BRINGING TO LIGHT: THE CHALLENGES OF REPRESENTING AND REASONING COMMON SENSE KNOWLEDGE IN AI SYSTEMS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF SOCIAL SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ZEYNEP KABADERE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF ARTS
IN
THE DEPARTMENT OF PHILOSOPHY

DECEMBER 2024

Approval of the thesis:

BRINGING TO LIGHT: THE CHALLENGES OF REPRESENTING AND REASONING COMMON SENSE KNOWLEDGE IN AI SYSTEMS

submitted by **ZEYNEP KABADERE** in partial fulfillment of the requirements for the degree of **Master of Arts in Philosophy, the Graduate School of Social Sciences of Middle East Technical University** by,

Prof. Dr. Sadettin KİRAZCI
Dean
Graduate School of Social Sciences

Assoc. Prof. Dr. Aret Karademir
Head of Department
Department of Philosophy

Assoc. Prof. Dr. Aziz Fevzi Zambak
Supervisor
Department of Philosophy

Examining Committee Members:

Assoc. Prof. Dr. Refik Güremen (Head of the Examining Committee)
Middle East Technical University
Department of Philosophy

Assoc. Prof. Dr. Aziz Fevzi Zambak (Supervisor)
Middle East Technical University
Department of Philosophy

Assoc. Prof. Dr. Hasan Çağatay
Social Sciences University of Ankara
Department of Philosophy

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: Zeynep Kabadere

Signature:

ABSTRACT

BRINGING TO LIGHT: THE CHALLENGES OF REPRESENTING AND REASONING COMMON SENSE KNOWLEDGE IN AI SYSTEMS

KABADERE, Zeynep
M.A., The Department of Philosophy
Supervisor: Assoc. Prof. Dr. Aziz Fevzi ZAMBAK

December 2024, 211 pages

This thesis, firstly, investigates the challenges of imitating common sense reasoning in artificial intelligence (AI) by focusing on three core issues: representing common sense knowledge, identifying tacit knowledge, and addressing the frame problem. In the first chapter, the study examines these challenges through the lens of knowledge representation, reasoning, and learning processes, highlighting their significance in enhancing AI's ability to handle everyday reasoning tasks. In the second chapter, the thesis presents a comprehensive evaluation of two Large Language Model (LLM)-based AI systems, ChatGPT 4.o and Claude Sonnet 3.5, to assess their capacity to simulate common sense reasoning. This evaluation is structured around six primary benchmarks: context-based information integration, future planning and adaptation ability, comprehensive causality and linked information management, operational execution competence, background knowledge integration and application, and accuracy and relevance management. These benchmarks are further refined into 27 detailed sub-benchmarks designed to address the challenges identified in the first chapter comprehensively. By analyzing the experimental results, the thesis identifies the strengths and limitations of both models in imitating common sense reasoning.

The findings contribute to the broader understanding of AI's capabilities and limitations in replicating common sense reasoning, providing insights into areas requiring further development. Ultimately, this study bridges philosophical inquiry and empirical evaluation to offer a robust framework for advancing the design of contextually aware and reasoning-capable AI systems.

Keywords: common sense reasoning, artificial intelligence, large language models, benchmarking analysis, knowledge representation.

ÖZ

YAPAY ZEKA SİSTEMLERİNDE SAĞDUYULU AKIL YÜRÜTME BECERİSİ ÜZERİNE BİR İNCELEME

KABADERE Zeynep
Yüksek Lisans, Felsefe Bölümü
Tez Yöneticisi: Doç. Dr. Aziz Fevzi ZAMBAK

Aralık 2024, 211 sayfa

Bu tez, öncelikle yapay zekada (YZ) sağduyulu akıl yürütmeyi taklit etmenin zorluklarını üç temel mesele üzerinden incelemektedir: Sağduyu bilgisinin temsil edilmesi, örtük bilginin tanımlanması ve çerçeve probleminin sağduyulu akıl yürütme bağlamında ele alınması. Birinci bölümde, bu zorluklar bilgi temsili, akıl yürütme ve öğrenme süreçleri bağlamında ele alınmakta ve sağduyu akıl yürütmesinin YZ'nin günlük yaşamda karşılaştığı görevleri yerine getirme yeteneğini artırmadaki önemi vurgulanmaktadır. İkinci bölümde ise iki Büyük Dil Modeli (LLM) tabanlı yapay zekâ sistemi olan ChatGPT 4.o ve Claude Sonnet 3.5'in sağduyu akıl yürütmeyi taklit etme kapasitelerini değerlendiren kapsamlı bir analizi sunulmaktadır. Bu değerlendirme, altı ana ölçüt etrafında yapılandırılmıştır: bağlam tabanlı bilgi entegrasyonu, geleceğe yönelik planlama ve uyum sağlama yeteneği, kapsamlı nedensellik ve bağlantılı bilgi yönetimi, operasyonel yürütme yeterliliği, arka plan bilgisinin entegrasyonu ve uygulanması ile doğruluk ve alaka yönetimi. Bu ana ölçütler, birinci bölümde tanımlanan zorlukları kapsamlı bir şekilde ele almak için 27 ayrıntılı alt ölçüt daha planlanmıştır. Deneysel sonuçların analizi, her iki modelin sağduyu akıl yürütmesini taklit etmedeki güçlü ve zayıf yönlerini ortaya koymaktadır. Elde edilen bulgular, YZ'nin bu alandaki yeteneklerini ve sınırlamalarını daha geniş bir perspektiften anlamamıza katkı sağlarken, geliştirilmesi gereken alanlar için de önemli içgörüler sunmaktadır. Sonuç olarak, bu çalışma, felsefi sorgulama ile ampirik değerlendirmeyi bir araya getirerek, bağlamsal farkındalığa ve akıl yürütme yeteneğine sahip yapay zeka sistemlerinin tasarımını ilerletmek için sağlam bir çerçeve sunmayı hedeflemektedir.

Anahtar Kelimeler: sağduyulu akıl yürütme, yapay zeka, büyük dil molleri,
kıyaslama analizi, bilgi temsili

To my mother, who shared every step of this journey with me,

with endless love and unwavering support.

ACKNOWLEDGMENTS

First and foremost, I wish to extend my deepest gratitude to the faculty and staff of the Middle East Technical University's Philosophy Department. To all my professors who guided me with their wisdom and encouraged me to think philosophically, and to the staff whose tireless efforts have been the backbone of our department—thank you. I owe my sincerest thanks to my academic advisor, Aziz Fevzi Zambak, for his invaluable guidance throughout this journey.

To my friend İbrahim Ethem Deveci, I am endlessly grateful for his unwavering support during this thesis process. His patience in answering my countless questions, his commitment to discussing every argument with the utmost seriousness, and his brilliant ideas have been indispensable. To my dear friend İdil Tomrukçu, thank you for the love and support that brightened even the most challenging days.

To my family, whose love has carried me through every challenge, my heartfelt thanks. To my mother, my constant source of strength, my grandmothers, thank you for standing by me every step of the way and believing in me with unwavering faith.

Finally, I want to take a moment to thank myself. For standing tall when life became difficult, for carrying on with passion and determination, for never losing my curiosity for learning and research, and for navigating this process with grace and perseverance. This accomplishment is a testament to the journey I undertook, and I am proud of the person I have become through it

TABLE OF CONTENT

PLAGIARISM iii
ABSTRACT iv
ÖZ v
ACKNOWLEDGMENTS ix
TABLE OF CONTENT x
LIST OF TABLES xi
LIST OF ABBREVIATIONS xii
CHAPTERS
1. INTRODUCTION 1
2. THE THREE CHALLENGES IN IMITATING COMMON SENSE REASONING 12
2.1 The First Challenge: Representing Common Sense Knowledge 12
2.2 Second Challenge: The Complexity of Tacit Knowledge in Common Sense Reasoning 23
2.3 The Third Challenge: The Frame Problem 36
3. EVALUATING LARGE LANGUAGE MODELS: ADDRESSING CHALLENGES IN COMMON SENSE REASONING 58
3.1 Benchmarks 59
3.2. Methods Of Data Collection and Evaluation 65
3.3 Results and Discussion 68
4. CONCLUSION 108
BIBLIOGRAPHY 113
APPENDICES
A. SCENARIOS FOR COMMON SENSE REASONING BENCHMARK EVALUATION 121
B. TURKISH SUMMARY / TÜRKÇE ÖZET 197
D. SAMPLE TEZ İZİN FORMU / THESIS PERMISSION FORM 211

LIST OF TABLES

Table I: The Frame Problem Robot Experiment 39
Table-II: Benchmarks 66
Table III: Benchmark-1: Context-Based Information Integration Evaluation 71
Table IV: Benchmark-2: Future Planning and Adaptation Ability Evaluation........... 82
Table-VI: Benchmark-4: Operational Execution Competence Evaluation 89
Table-VII: Benchmark-5: Background Knowledge Integration and Application Evaluation 99
Table-VIII: Benchmark-6: Accuracy and Relevance Management Evaluation 107

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
KR	Knowledge Representation
LLM	Large Language Model
YZ	Yapay Zeka

CHAPTER 1

INTRODUCTION

Common sense is the name given to all the intuitive beliefs, assumptions, and reasoning abilities that are known and accepted by everyone and help understand and interpret the world in everyday matters (Mueller, 2014). Common sense knowledge, which is not based on any conscious method, has emerged due to the use of the senses and the most primitive kind of experience. For these reasons, scientific and common sense knowledge have quite different characteristics. Scientific knowledge is obtained using specific methods and tools and is objective, systematic, consistent, and open to criticism. It concerns more precise information about the world (McCarthy, 1981). However, common sense knowledge is experiential and depends on repeated personal experience. Though they share similarities, common sense knowledge, and common sense reasoning differ significantly. McCarthy (1984) distinguishes between common sense knowledge "what everyone knows" and common sense reasoning, "the human ability to use common sense knowledge." Common sense reasoning is the cognitive process that utilizes common sense knowledge to construct inferences and make sense of specific situations. While common sense reasoning encompasses the application and cognitive processes necessary to reason based on that information, common sense knowledge provides the content. Common sense reasoning helps individuals interpret and navigate everyday situations by using intuitive knowledge, allowing for quick, practical, and contextually appropriate decision-making. Moreover, common sense reasoning provides a solid foundation upon which more advanced reasoning processes are built. It allows us to draw conclusions based on past experiences, predict possible outcomes, and understand cause-and-effect relationships in both the physical and social worlds (McCarthy, 1984). For instance, in the physical sciences, understanding the forces acting on an object first requires grasping fundamental concepts like the object's existence, its spatial position, and the idea that it can move. This type of intuitive knowledge is part of common sense reasoning and later serves as the basis for formalized scientific theories. Furthermore, common sense reasoning is essential for problem-solving, allowing us to use previous knowledge to find solutions. For example, we know that touching a hot cup may burn our hand, or that lifting a heavy object is more challenging than lifting a lighter one. Similarly, we can predict that if we throw a ball into the air, it will fall to the ground, as physical laws like gravity are universally understood as part of common sense. Common sense not only helps us understand the physical world but also guides us in social interactions. Knowing to wait after ordering food at a restaurant or understanding that maintaining eye contact during a conversation is polite are examples of this type of knowledge. Common sense reasoning supports quick, intuitive decision-making by evaluating likely outcomes and helps us interpret implied meanings in conversations and adjust our responses accordingly. For example, when we say, "This task is a mess," we might be referring not to a literal mess, but to a complicated or challenging situation. Similarly, if someone says, "She’s on cloud nine," we understand it metaphorically to mean she’s extremely happy rather than literally in the clouds. Additionally, it promotes learning and adaptation by relating new situations to similar past experiences, allowing us to respond effectively to changing environments. Common sense knowledge is also shaped by social norms and cultural experiences. For instance, giving up a seat to the elderly on public transportation is seen as respectful, while offering tea to guests in Turkey is regarded as a friendly gesture.

In light of all this information, it is possible to give a definition of common sense reasoning that can be discussed. Common sense reasoning is the ability to effectively use ordinary, everyday, experiential knowledge in achieving ordinary, everyday, practical goals (Brachman & Levesque, 2022). It involves working through relevant knowledge without feeling overwhelmed and applying it quickly and effortlessly to understand its implications for the current situation. Without specialized training or highly developed analytical abilities, common sense deals with issues that often arise in daily life. Instead of relying on formal education or specialized reading, it is based on personal experiences. Instead of being an intellectual endeavor or the foundation for philosophical debates, common sense is focused on making decisions that will help one succeed in the actual world.

Common sense reasoning is also of critical importance for Artificial Intelligence, as it enables systems to interact effectively with people in real-world situations and adapt to complex circumstances. AI systems can use common sense reasoning to go beyond surface patterns and reach more situationally suitable conclusions. This foundational capacity for commonsense reasoning provides the essential framework for AI to achieve a higher level of scientific knowledge processing and reasoning (Brachman & Levesque, 2022). AI systems can also better comprehend human intents, preferences, and social interactions when they imitate common sense reasoning. It enables them to interpret ambiguous language, grasp implied meanings, and respond intelligently to human queries or commands. By acquiring flexibility and contextual understanding, AI systems can move beyond merely executing simple commands to succeeding in complex tasks. As a result, the user experience is enhanced, and more effective and natural human-AI interaction is made possible (Lake et al., 2016). Moreover, improving the security and dependability of AI systems is an important reason for emulating common sense reasoning. Common sense reasoning allows AI to anticipate the consequences of actions, identify potential risks, and avoid dangerous or nonsensical behaviors (Lake et al., 2016). This is especially crucial in fields where safety is at stake, like autonomous vehicles or medical diagnostics, where a deficiency in common sense thinking might have dire repercussions. Therefore, common sense reasoning forms a crucial infrastructure for both humans and AI in processing scientific knowledge and engaging in sound reasoning; without this foundation, scientific thinking and formal reasoning become challenging to execute effectively in practice.

Today, many AI researchers are striving to imitate common sense reasoning, as it is still recognized as a critical missing component in AI's capabilities. Although AI has made significant advances in specific fields such as image recognition and natural language processing, it often struggles to understand and reason about the world as humans do. In fact, the problem of imitating common sense reasoning has been a persistent issue since the foundation of the field of artificial intelligence. John McCarthy used the phrase "Artificial Intelligence" (AI) in the 1950s, and in 1959 he published "Programs with Common Sense," one of the first works on AI. Despite this long history, developing common sense reasoning has become considerably harder than expected.

The book “Rebooting AI” by Gary Marcus and Ernest Davis (2019) emphasizes that the need for machines to possess common sense is widespread, yet no effective solutions have been found thus far. They believed common sense is a core part of intelligence rather than a secondary aspect. Moreover, many of the important concerns mentioned by McCarthy and Hayes (1969) and McCarthy (1984) are still relevant today. Some subjects are time and space, causation, qualitative theories of motion, force, substances, energy, continuous change, and quantities. Understanding changes, actions, and cause-and-effect in the physical world—often called “naive physics” and “naive mathematics”—is at the heart of common sense knowledge. The scope of common sense knowledge can also be enlarged to incorporate ideas from naïve or folk psychology, such as goals, beliefs, and desires, as human activities entail relationships with other people. AI systems seek to overcome the drawbacks of exclusively data-driven techniques and include a broader knowledge of the environment by imitating common sense reasoning.

Imitating common sense reasoning remains a significant challenge in AI research because it is extremely complex to impart everyday knowledge and intuitive understanding to machines. Humans possess a vast amount of implicit knowledge and contextual understanding in their daily lives without conscious effort; however, teaching this knowledge to an AI system and organizing it in a format that the system can understand and use effectively is challenging (Mueller, 2014). Imitating common sense reasoning goes beyond simply working with data—it requires an AI system to develop the ability to handle uncertainty, make logical inferences with incomplete information, understand causal relationships, and respond appropriately by grasping context. This necessitates the advancement of knowledge representation and reasoning techniques that surpass the current data-driven AI methods.

In the first chapter of my thesis, I will examine the three main challenges involved in imitating common sense reasoning: representing common sense knowledge, identifying tacit knowledge, and addressing the frame problem.

My aim in addressing these challenges is to shed light on the key obstacles that hinder progress in this area, despite the fact that common sense reasoning is essential for AI to operate effectively in real-world contexts. By analyzing these three core challenges, I intend to explain why common sense reasoning remains a pressing issue in AI research and to identify areas where further improvements are needed. This approach will allow me to closely examine the processes of knowledge representation, reasoning, and learning that are necessary for enhancing the everyday reasoning capabilities of AI systems (Mueller, 2014).

Firstly, I will address the development and challenges of knowledge representation (KR) in artificial intelligence (AI), with a particular focus on common sense knowledge. This chapter will trace the historical roots of KR, from its philosophical foundations with thinkers like Leibniz and Frege to its evolution into modern computational representation methods. Key figures and concepts in the field, such as McCarthy's (1989) seminal work on symbolically representing common sense for computational reasoning and subsequent advancements, will be discussed. Moreover, this section will examine in detail two primary approaches to representing common sense knowledge: rule-based systems and modern AI techniques, such as neural networks and large language models (LLMs) (Naveed et al., 2023). While rule-based methods rely on predefined rules and logical frameworks to organize knowledge, neural networks and LLMs distinguish themselves through their capacity to learn from extensive datasets. The discussion will explore the strengths and challenges each approach presents for knowledge representation, emphasizing which methods may be more effective in addressing the complexities of common sense reasoning.

For the second challenge, I will examine the problem of identifying tacit knowledge. Tacit knowledge consists of fundamental information that humans intuitively use in daily life, such as naive physics for understanding the physical world, naive mathematics for basic quantitative concepts, and naive psychology for navigating social relationships (Brachman & Levesque, 2022). This knowledge enables people to interpret their surroundings and respond appropriately. It is essential for an AI system to grasp context, handle uncertainty, and make sensible decisions in everyday situations; however, representing this knowledge in AI systems is filled with challenges.

Due to its extensive and diverse scope, tacit knowledge is difficult to identify, quantify, and reduce to a structured form. Furthermore, this knowledge can vary in meaning across different contexts, requiring AI to accurately assess and adapt to these contextual shifts (Brachman & Levesque, 2022).

In this part, I will discuss the nature of tacit knowledge, its critical role in common sense reasoning, and the complexities associated with identifying and incorporating this knowledge into AI systems, particularly in terms of contextual sensitivity and the inherent subtleties of human experience that resist standardization.

Lastly, I critically examine the frame problem, a fundamental issue in enabling AI to reason effectively in changing environments. Solving the frame problem is crucial for imitating common sense reasoning, as an AI system's ability to think like a human in the real world relies on its capacity to distinguish relevant information from irrelevant details, adapt quickly to changes, and make accurate inferences in dynamic situations (Dennett, 1990). In this context, the frame problem represents a major obstacle to AI's ability to comprehend common sense reasoning processes and emulate the flexible thinking that humans display. This chapter will address the challenges of representing knowledge in a way that enables AI to simulate future states or respond to changes with contextual understanding. By examining the philosophical foundations and historical development of the frame problem—including contributions from thinkers like McCarthy, Hayes, Dennett, and Fodor—I will highlight the broader implications of this issue for AI's ability to mimic human common sense reasoning. Additionally, I will assess both practical and theoretical approaches to solving the frame problem, emphasizing the importance of a multifaceted strategy to enhance AI's capabilities in this critical area and advance its capacity for human-like reasoning.

In the second chapter, I will conduct a comprehensive evaluation of the ability of two different Large Language Model (LLM)-based AI systems (ChatGPT 4.o and Claude Sonnet 3.5) to imitate common sense reasoning. This evaluation will be structured around six primary benchmarks: context-based information integration, future planning and adaptation ability, comprehensive causality and linked information management, operational execution competence, background knowledge integration and application, and accuracy and relevance management. Additionally, 24 detailed sub-benchmarks will complement these main benchmarks. Considering the challenges discussed in the first chapter, I aim to assess to what extent these models can successfully imitate common sense reasoning. The benchmark list I developed is specifically designed to comprehensively address each of the identified challenges. By analyzing the experimental results, I ultimately aim to determine which model is more effective in imitating common sense reasoning.

In evaluating common sense reasoning, I prefer to use LLM models due to their flexibility and dynamic structure in terms of knowledge representation and integration (Naveed et al., 2023). Traditional rule-based systems, which rely on rigid, predefined rules, fall short in handling situations that involve variability and uncertainty (Grosan, & Abraham, 2011). This limited structure of rule-based systems prevents them from effectively managing the complex and context-dependent information required for common sense reasoning (Brachman & Levesque, 2022). In contrast, LLM models can integrate a wide range of data sources, create a richer knowledge base, make inferences based on incomplete or uncertain information, and make flexible decisions (Naveed et al., 2023). These capabilities are the primary reasons for choosing LLM models over rule-based systems in testing common sense reasoning.

LLMs are deep learning models used in the field of natural language processing and are trained on vast datasets. They are designed to understand human language, generate text, perform translation, and summarization tasks (Naveed et al., 2023). Containing billions or even trillions of parameters, these models are trained on a diverse range of data sources, including texts from the internet, books, and articles. The 'transformer' architecture underlying LLMs, through a 'self-attention' mechanism, enables the model to learn the context and relationships between words in a text (Naveed et al., 2023). This architecture makes LLM-based AI highly effective in understanding natural language and generating new text. Moreover, LLMs' ability to learn from broad data sources allows them to infer from 'invisible' knowledge similarly to human intuition. This supports the essential feature of common sense reasoning: the ability to draw logical conclusions with incomplete information and to update the knowledge base as new information emerges (Naveed et al., 2023).

Both ChatGPT 4.o and Claude Sonnet 3.5 are advanced large language models designed for natural language processing, but they differ significantly in their focus and strengths. ChatGPT 4.o, developed by OpenAI, excels at processing a wide range of information and generating detailed responses on a variety of topics, making it highly versatile (*Hello GPT-4o*, 2024). It is well-suited for technical tasks, complex reasoning, and academic research. The model’s ability to handle long contexts and produce coherent responses gives it a significant advantage, especially when dealing with intricate or highly specialized subjects. On the other hand, Claude Sonnet 3.5, developed by Anthropic, places great emphasis on safety, ethical concerns, and minimizing harmful outputs (*Introducing Claude 3.5 Sonnet*, 2024). This model is focused on producing responses that align with ethical standards and avoiding biases or inappropriate content. While it may not have access to as broad a dataset as ChatGPT 4.o, Claude Sonnet 3.5 excels in managing social and emotional contexts and ensuring that its outputs are safe, reliable, and ethical. Its design prioritizes safety and user well-being, making it particularly effective in applications where ethical decision-making is critical (*Introducing Claude 3.5 Sonnet*, 2024).

While the choice of LLM models provides a flexible and robust foundation for common sense reasoning, a systematic evaluation framework is essential to accurately measure their performance across different contexts and challenges. Therefore, I chose to use a benchmark system to evaluate the common sense reasoning abilities of LLM models. A benchmark is a standardized test set, or measurement tool used to evaluate a specific performance or capability. In the fields of artificial intelligence and machine learning, a benchmark is used to assess how well a model performs a particular task. Developing benchmarks is crucial in AI research as it allows for evaluating system capabilities, tracking progress, and illustrating the technology's limitations (Davis, 2023). Benchmarks not only assess current performance but also highlight overlooked or underexplored issues, encouraging researchers to focus on these areas. Additionally, well-designed benchmarks provide a common language and standard for the research community, facilitating scientific communication and collaboration. A successful benchmark can guide future developments in AI and serve as a roadmap for creating more accurate and reliable systems. In complex areas like commonsense reasoning, benchmarks allow for the comparison of different systems under the same conditions, making it possible to determine which approaches are more effective (Davis, 2023).

However, many existing commonsense benchmarks lack the qualities needed for accurate measurement. Most benchmarks include questions that actually require common knowledge or encyclopedic information and sometimes demand expertise-level details. These types of benchmarks make it challenging to effectively measure AI's commonsense reasoning abilities (Davis, 2023). To accurately evaluate commonsense reasoning, benchmarks should focus on more fundamental and intuitive knowledge. Some existing benchmarks contain "flawed" questions, which deviate from ideal characteristics and include poorly constructed structures. Such flawed questions tend to lead the model to predict correct answers based solely on the technical structure of the questions or the assumptions of the author, rather than truly testing commonsense reasoning. This issue reduces the reliability of the benchmark and complicates the evaluation process (Davis, 2023).

These types of benchmarks result in assessments that are more focused on the structure of the tests or the author's intended specific answer, rather than genuinely evaluating the AI's commonsense reasoning ability. This reduces the accuracy, reliability, and alignment of the benchmark with its intended purpose. Flawed questions make it difficult to determine whether the model is genuinely solving problems through commonsense reasoning or merely relying on technical information. A good benchmark should provide clear, correct answers, prepare questions impartially and based on commonsense knowledge, and enable an accurate evaluation of the model's capabilities. In this context, it is essential to review existing commonsense benchmarks to assess the extent to which they meet these criteria.

CommonsenseQA 2.0: CommonsenseQA 2.0 provides a broad array of questions rooted in commonsense reasoning, effectively testing a model's capacity to make inferences based on fundamental knowledge. The gamification approach used in its design has contributed to the generation of diverse and engaging questions, enhancing the benchmark's effectiveness. However, because these questions are derived from ConceptNet, the model may be inclined to learn specific patterns rather than engage in authentic commonsense reasoning. This pattern recognition tendency could restrict the model's adaptability to new or unfamiliar scenarios, limiting its true commonsense reasoning capabilities (Davis, 2023).

BIG-bench: BIG-bench offers a comprehensive scope with 212 datasets, regularly updated by the research community, which makes it inclusive and adaptable to the evolving needs of AI research. This extensive range allows for a more holistic view of a model's capabilities across diverse tasks. However, the breadth of the benchmark can make it challenging to focus on specific commonsense tasks. The diverse origins of the datasets may lead to inconsistencies, and the quality of each subset may vary, affecting the benchmark's reliability for commonsense reasoning evaluation (Davis, 2023).

Winograd Schema Challenge: The Winograd Schema Challenge is focused on pronoun resolution, directly testing the language-based reasoning capabilities of AI by examining subtle linguistic cues. This specific focus is effective for determining whether a model possesses true language understanding and logical reasoning skills. However, as a language-centered task, Winograd Schema Challenge lacks broader commonsense reasoning elements related to the physical or social world, providing limited insight into general commonsense abilities (Davis, 2023).

SWAG and HellaSWAG: SWAG and HellaSWAG are designed to test a model's ability to predict sequences of events, evaluating its understanding of event order and logical inference. This benchmark is particularly useful for assessing sequential reasoning. However, since it relies heavily on sequential information, it is less effective at testing a model's understanding of social or emotional contexts. Additionally, the models may risk learning specific patterns, potentially limiting their adaptability to new, less predictable scenarios (Davis, 2023).

McTaco: McTaco is focused on testing the model's ability to make temporal inferences, thus providing insight into AI's capacity for temporal reasoning. This benchmark benefits from crowdsourced data, which results in a wider and more natural dataset. However, its focus on temporal reasoning restricts its usefulness in evaluating other commonsense capabilities, making it a narrower assessment of overall commonsense reasoning (Davis, 2023).

Visual Genome: Visual Genome is aimed at testing AI's visual perception abilities, specifically in understanding objects and their relationships, through extensive annotations that allow for detailed visual comprehension. This focus on visual commonsense provides valuable insights into a model's ability to interpret visual data. However, due to its visual orientation, it is limited in evaluating language-based logical reasoning. Furthermore, there is debate over how effectively images alone can provide AI with genuine commonsense knowledge (Davis, 2023).

CommonsenseQA (Original): The original CommonsenseQA benchmark tests general commonsense knowledge and requires basic commonsense inferences, offering a comprehensive assessment with its broad set of questions. While it is effective in measuring general commonsense understanding, it may overlap with encyclopedic knowledge, which could limit its effectiveness in assessing pure commonsense reasoning (Davis, 2023).

Social IQA: Social IQA focuses on understanding social interactions, assessing AI's ability to reason about psychological and social commonsense. This benchmark provides insights into a model's understanding of human emotions and intentions, making it useful for evaluating social reasoning skills. However, the cultural and language-specific nature of social norms may limit its generalizability. Additionally, as it primarily addresses social commonsense, it does not provide a full picture of a model's general commonsense reasoning capabilities (Davis, 2023).

These benchmarks employ various techniques and content to assess different facets of commonsense reasoning. By targeting specific areas, each aims to measure distinct commonsense abilities in AI. In developing my benchmark system, I conducted a thorough analysis of the strengths and weaknesses of existing benchmarks. My aim was to evaluate the extent to which these benchmarks address the current challenges in commonsense reasoning. I also sought to incorporate, where possible, the strengths of previously developed benchmarks into my own system. When addressing the weaknesses, I avoided questions that could be answered based solely on specific patterns, instead designing my benchmark to require more intuitive inferences. In this way, I aimed to create a more comprehensive, reliable, and contextually diverse benchmark system capable of genuinely assessing AI's commonsense reasoning abilities.

CHAPTER 2

THE THREE CHALLENGES IN IMITATING COMMON SENSE REASONING

2.1 The First Challenge: Representing Common Sense Knowledge

Understanding and representing knowledge is critical for creating intelligent systems in the rapidly expanding field of artificial intelligence (AI). One of the primary goals of AI research is to be able to use symbols harmoniously and symbolically represent complex information. This chapter explores the historical and methodological development of knowledge representation (KR) in AI, addressing the challenges of accurately capturing common sense knowledge and comparing rule-based and modern learning-based approaches to tackle these complexities.

Examining traditional concepts of knowledge and how it can be transferred to machines lays the groundwork for AI systems to grasp, analyze, and interact with knowledge efficiently. Moreover, performing justified reasoning like abductive, inductive, and deductive reasoning, which is the critical component of AI, is impossible without a knowledge base (Davis, Shrobe, & Szolovits, 1993). The goal of researchers working in the area of knowledge representation (KR) is to reveal how to build a knowledge base and what characteristics should be included in it. The capacity to symbolically represent every bit of knowledge is the first prerequisite (Davis, Shrobe, & Szolovits, 1993). This implies that knowledge, which might be somewhat abstract, must be converted into characters or symbols. A concept represented symbolically is analogous to an idea represented as a series of digits. Put another way, in order to be effectively manipulated, knowledge—which is frequently subtle and complex—must be reduced to symbols. The capacity to appropriately manipulate these symbolic representations is the second prerequisite.

This implies that procedures or guidelines must be followed to operate with symbols once information has been transformed into a symbolic form. These operations allow translating and combining symbols to create new representations analogous to arithmetic (Davis, Shrobe, & Szolovits, 1993). To fully understand the process of common sense knowledge representation and to effectively discuss the challenges encountered and proposed solutions, it is essential to first examine the historical development of the problem. Delving into the roots of the issue allows us to identify the stages at which difficulties have arisen in the knowledge representation process and understand how past methods have attempted to address these challenges.

2.1.1 Roots of Knowledge Representation

There is a rich historical legacy behind knowledge representation, which goes far beyond the last few decades of artificial intelligence (AI) development. Although there are indications of formal logic in ancient philosophical works, Gottfried Wilhelm Leibniz is primarily responsible for the use of calculus—a formal method of symbolic manipulation—to represent and manipulate concepts (Sowa, 2012). One of his main concepts was the idea of a "universal characteristic," sometimes known as a "calculus ratiocinator," which is a general approach or language of thought. Leibniz envisioned a twofold theory for a symbolic language capable of expressing all human knowledge and could be manipulated to derive new ideas (Sowa, 2012). His first proposition was that all knowledge might be represented symbolically, like a universal language. Second, he underlined how crucial it is to provide a calculus or technique to manipulate these symbols in a methodical way so that new information may be derived from preexisting representations (Sowa, 2012). Leibniz thought a systematic and universal approach to thinking might be achieved by defining rules for manipulating symbolic representations of concepts. His goal was to develop a symbolic calculus that would explain intricate concepts and facilitate deductions. The foundation for what we now refer to as knowledge representation was established by this concept. Leibniz's early insight into the concept of formal, symbolic knowledge representation was one of the key ideas that ultimately drove later breakthroughs in AI.

After Leibniz, one of the critical figures was Gottlob Frege, who was one of the biggest contributors to mathematical logic at the end of the 19th century (Sowa, 2012). His works provided a critical component for the development of automated reasoning, even though he did not see it. Frege and his successors established a solid technical framework for creating AI reasoning systems. Nevertheless, it is crucial to remember that the primary goal of the logicians of that era was to represent and formalize mathematical truth, which is related to the ideas and theorems found in the field of mathematics itself in the early 20th century (Sowa, 2012). These formal systems, which were frequently founded on mathematical logic, were effective at expressing some mathematical facts, such as the characteristics of sets, numbers, and logical relations. Still, they were not naturally equipped to handle the subtleties and complexity of commonsense truth. Commonsense knowledge frequently depends more on context, draws on real-world experiences, and may not adhere to the rigid rules of mathematical logic (Sowa, 2012).

2.1.2 Knowledge Representation Hypothesis

Despite all these problems, since the middle of the 20th century, common sense knowledge has been tried to be defined and axiomatized. McCarthy, one of the founders of AI, took the initiative to capture common sense knowledge in symbolic form and use it for computational reasoning. McCarthy proposed developing a computer program that might use symbolic representations of common sense knowledge to direct decision-making and action selection in his seminal work "Programs with Common Sense," published in 1959. Although McCarthy's research was groundbreaking for representing common sense knowledge, it focused on a very specific area. For this reason, before talking about McCarthy's work, we should mention Brian Smith's Knowledge Representation Hypothesis, which discussed the representation of knowledge from a more general perspective in 1985. Understanding this hypothesis will also shed light on McCarthy's claim, which we will examine in detail later. Brian Smith (1985) was one of the first to address the knowledge representation required for artificial intelligence. According to the Smith's hypothesis:

> Any mechanically embodied intelligent process will be comprised of structural ingredients that a) we as external observers naturally take to represent a propositional account of the knowledge that the overall process exhibits, and b) independent of such external semantic attribution, play a formal but causal and essential role in engendering the behavior that manifests that knowledge (1985).

In Smith's view, one of the most essential points is that symbolic structures must be compatible with propositional interpretation in order for them to function as an efficient representation of knowledge (Brachman, 1988). This involves viewing these structures as expressions that carry truth values. The symbols used in constructing sentences should have a consistent and coherent interpretation to achieve a propositional interpretation. The consistency of the structure guarantees that we can express the circumstances in which the structure would be true or false while referring to it. These symbols are supposed to represent meaningful propositions about the external world. The causal relationship between the system's behavior and the existence of these symbolic structures is another crucial aspect of Smith's argument (Brachman, 1988). According to Smith (1985), the passive existence of these structures is insufficient; instead, they have to affect and cause the system's behavior actively. If these symbolic structures were analogous to comments in a program (having nothing to do with the actions of the system), they would be irrelevant to intelligent behavior. Smith underlines the importance of symbols within structures that play a causal role in determining how the system behaves. In that point, propositional interpretation is crucial for guaranteeing that symbolic structures are interpreted as true or false assertions and for their causal function in influencing the system's behavior (Brachman, 1988). Therefore, the concept that intelligence involves not merely possession of knowledge but also its active application to guide behavior in a way consistent with interpreted knowledge is the key idea of Smith's knowledge representation hypothesis. In essence, the KR hypothesis prompted a reflective stance within the AI community, encouraging ongoing inquiry into the nature of knowledge representation and the validity of the symbolic approach (Brachman, & Levesque, 2022). However, the hypothesis doesn't provide insights into the mechanism by which these symbolic structures lead to intelligent behavior. The "how" of engendering intelligent behavior through symbolic representations is a question separate from the KR hypothesis and constitutes a significant aspect to be explored. Furthermore, the KR hypothesis doesn't specify the nature or type of knowledge involved, although it's evident that McCarthy (1959), in his work, was particularly concerned with commonsense knowledge.

2.1.3 Foundations of Knowledge-Based Systems: McCarthy's Vision and the Birth of Common Sense for AI

In 1959, John McCarthy, a pioneer in artificial intelligence, introduced the idea of "Programs with Common Sense." This groundbreaking concept emphasized the use of symbolic representations for processing commonsense knowledge in computers. McCarthy envisioned a system that could make decisions based on symbolic representations, representing everyday objects and relationships.

> One will be able to assume that [the proposed system] will have available to it a fairly wide class of immediate logical consequences of anything it is told and its previous knowledge. This property is expected to have much in common with what makes us describe certain humans as having common sense (1959).

McCarthy's idea established the groundwork for the creation of knowledge-based systems that mimic elements of human-like commonsense understanding by utilizing symbolic representations and logical reasoning. McCarthy mainly supported the use of first-order predicate logic, a branch of the broader discipline of symbolic logic, as a representation scheme (Brachman, & Levesque, 2022). This language was created to express relationships and claims about the world in an organized and formal way. The program, based on the principles laid out in Alan Turing's foundations of calculation, was aimed at obtaining new insights and results through the systematic application of logical rules to symbolic representations (Brachman, & Levesque, 2022). McCarthy (1959) named the program "advice taker," which employs first-order predicate logic to enable a system to make deductions from a given set of logical premises. This was the initial plan for developing a computer system that is today known as a knowledge-based system. This ground-breaking idea presented a framework with several essential elements, including:

Knowledge Base: The knowledge base is the system's main component. System memory-resident symbolic expressions are included in this repository. Many facets of the knowledge that the system has learned and acquired are represented by these symbols.

Symbolic Expressions and Logic: The knowledge base's symbolic expressions are meaningful representations rather than arbitrary constructs. McCarthy placed a strong emphasis on logic, suggesting that these symbols are interpreted in accordance with logical rules. In order to accomplish this, a formal logic-based reasoning system must be used to make deductions and create new symbolic representations.

Logical Processing: The system's core competency is its logical operations processing of the knowledge base. This entails making inferences, increasing the number of symbolic representations, and inferring implicit relationships. The method of logic functions as a computational instrument to broaden the system's comprehension beyond stated facts.

Decision-Making: The conclusions that emerge after logical processing demonstrate the intelligence of the system. Some of these findings are applicable and concern figuring out the system's next course of action. McCarthy envisioned a decision-making component in which the system assesses these findings and autonomously determines a course of action (McCarthy, 1959).

McCarthy's concept describes a knowledge-based system that uses logical processing, symbolic organization, and processing to make judgments. This provided the way for later advances in AI, highlighting the significance of autonomous decision-making, logical reasoning, and knowledge representation in intelligent systems. McCarthy (1959) pointed out that, in order to achieve human-level common sense intelligence, computers must transition to a knowledge-based system. McCarthy's main idea was that artificial intelligence requires computers to be able to draw on a rich source of knowledge, as humans do in applying common sense to a variety of problems.

In the knowledge-based program, commonsense reasoning is pursued through the development of representations suitable for handling various types of knowledge, and then reasoning is performed by relying on these representations. Although John McCarthy, often seen as a pioneer in the field, proposed the use of formal logic to govern commonsense reasoning in one of the first examples of strategy, his idea had certain drawbacks (Brachman & Levesque, 2022). One of the objections is the dependence on first-order logic, which is excellent at describing mathematical truth but less effective at capturing the subtleties of common sense knowledge. Instead of making absolute, categorical generalizations, common sense knowledge frequently works with typical and exceptional circumstances. Significant changes must be made to the first-order logic to accommodate typical, atypical, and borderline situations of common-sense categories. McCarthy's approach serves as the foundation for knowledge-based computer systems. However, it also emphasizes the need for advancements in the representation and justification of everyday knowledge beyond the first order logic's limits.

2.1.4 Evolving from McCarthy's Knowledge-Based System to Rule-Based Systems in Commonsense Reasoning

Although McCarthy's vision of a knowledge-based system was insufficient for mimicking common sense reasoning, it paved the way for the development of rule-based systems. Rule-based systems are among the simplest forms of artificial intelligence and are also referred to as production systems or expert systems. In these systems, knowledge representation is conducted directly through rules, meaning that information is encoded as rules within the system. These rules serve as directives that determine what actions the system should take or what conclusions it should reach in response to various situations. The definition of a rule-based system largely overlaps with that of expert systems. Expert systems emulate the thinking patterns of human experts to solve knowledge-intensive problems, using reasoning methods characteristic of human specialists. Thus, an expert system is designed to solve problems within a specific domain with the expertise of a human professional.

Instead of representing knowledge in a static, declarative form, rule-based systems dynamically process information through predefined rules. In classical knowledge representation, a set of facts is stored as a list of truths; however, in rule-based systems, knowledge is encoded as a series of "if... then..." statements. These rules specify the actions or conclusions the system should take in different scenarios. For example, a rule-based system might contain rules based on specific conditions such as weather, temperature, or humidity—e.g., "If it is raining, then take an umbrella." One advantage of these systems is their ability to make quick and effective decisions in knowledge-intensive situations. They assess which rules are applicable to a given situation and generate a conclusion based on those rules. Thanks to this feature, rule-based systems are commonly employed as expert systems in fields such as medical diagnosis, engineering problem-solving, and financial analysis.

However, McCarthy's original vision for knowledge representation aimed at creating a system capable of mimicking common sense reasoning. In the 1970s, work on rule-based systems mainly focused on domain-specific expertise rather than common sense knowledge. As a result, these technologies often lacked common sense and frequently failed when faced with situations outside their specialized areas (Brachman & Levesque, 2022). Additionally, designers of these systems overlooked the essential role that common sense knowledge and norms play in supporting expert systems. Researchers tended to focus on specialized, problem-specific rules rather than on the everyday and general truths that people know, use frequently, but rarely articulate. By the 1980s, however, it became clear that these systems needed greater robustness and adaptability to common sense and, at most, were highly competent only in specific tasks. While most researchers in the field pursued expertise-focused techniques, a select few directly followed McCarthy's example, concentrating on capturing common sense knowledge (Brachman & Levesque, 2022).

2.1.5 The Rule-Based Systems in Common Sense Reasoning

Various rule-based approaches have been employed to model common sense reasoning in artificial intelligence. Firstly, formal logic and mathematical approaches stand out. McCarthy relentlessly pursued this goal, collaborating with influential thinkers like Patrick Hayes (McCarthy & Hayes, 1981), Jerry Hobbs (Hobbs & Moore, 1985), Ernest Davis (Davis & Marcus, 2015), Raymond Reiter (2001), and Joseph Halpern (2017) to capture common sense perspectives on physics, time, space, minds, beliefs, plans, and society (Brachman, & Levesque, 2022). McCarthy and Robert Moore argue that certain elements of common sense reasoning require a logical structure, especially when dealing with incomplete information and making inferences. Depending on the circumstance, the predicate calculus language or a more expressive logic was employed.

However, formal logic has been criticized for being insufficient in this area. Human reasoning relies not only on logical rules but also on analogy, experience, and probabilistic thinking, which are often outside the scope of strict logical systems. According to Roger Schank (Schank & Abelson, 1977) and his pupils, including Janet

Kolodner (1993) and others, "logical" reasoning is not flexible enough to form the foundation of thought. Their attention changed from producing phrases to creating intricate symbolic structures that may act as a common sense agent's memory. Additionally, the context-dependent nature of language limits the practical applicability of formal logic, as such models struggle to adequately represent common sense knowledge. Davis and Marcus (2015) note that while mathematical techniques like situation calculus are theoretically powerful, they fall short in decomposing complex events and managing sequences of real-world actions. Among informal approaches, Marvin Minsky’s (1974) "frames" model is notable. Unlike strict logical systems, frames aim to accommodate the flexible thinking style humans naturally employ. Frames act as data structures representing various events or situations, enabling people to assess characteristics, expectations, and possible outcomes. Similarly, Schank and Abelson’s "scripts" theory (1989) represents structured sequences of events. These methods are useful for understanding organized behavior and making inferences about particular situations, but they struggle to generalize in more complex or unpredictable scenarios.

Finally, large-scale knowledge bases, such as Doug Lenat’s Cyc project (1995), have emerged as significant efforts in capturing common sense knowledge. Initiated in 1984, Cyc aims to develop an extensive ontology of everyday objects and actions. Cyc includes knowledge such as “you need to stand to walk” and “a person’s arms are usually visible, but their liver is not.” Lenat (1995) describes Cyc’s goal as building micro-theories to support common sense reasoning for specific scenarios rather than solving general intelligence. Critics, however, argue that simply accessing knowledge does not sufficiently capture how these facts interrelate in real-world contexts. Marcus and Davis (2019) highlight that while these ontologies have succeeded in specific domains, achieving a truly human-like level of common sense understanding requires further effort. In conclusion, rule-based systems, relying on fixed and predefined rules, lack flexibility and struggle to handle new or unexpected situations (Grosan, & Abraham, 2011). This rigidity particularly hinders their effectiveness in complex problems with multiple variables, as they cannot adapt knowledge contextually like humans do. Furthermore, these systems require constant manual updates. Adding new information, conditions, or scenarios necessitates individually updating or modifying rules, making maintenance complex and time-consuming (Grosan, & Abraham, 2011). This need for manual updates also limits scalability; as the knowledge base expands, managing it becomes increasingly challenging, reducing the system's overall applicability. As rule-based systems grow with additional rules, the likelihood of conflicts within the system also increases. Each rule is tailored to a specific scenario or context, so it is possible for rules to contradict each other in certain situations. Such conflicts compromise the system's consistency and accuracy, creating uncertainty about which rule should apply (Grosan, & Abraham, 2011). As the number of rules grows, managing these conflicts becomes more complex, negatively impacting the system's performance. Ultimately, rule-based systems remain limited in dynamic and uncertain real-world environments and are not well-suited to the goal of imitating common sense reasoning. These limitations highlight the clear need for flexible and learning-based systems in modern AI applications.

2.1.6 Learning-Based Systems in Commonsense Reasoning

The development of learning-based systems has been shaped by many pioneering scientists and critical milestones. The first steps were taken in 1950 with Alan Turing's "Computing machinery and intelligence," which questioned whether machines could think. Turing proposed (1950) that machines should not merely operate on fixed rules but should have the ability to learn. In 1958, Frank Rosenblatt built on this idea by developing the "perceptron," the first artificial neural network model capable of learning, inspired by the workings of the human brain. This innovation raised significant hopes for neural networks, but in 1969, Marvin Minsky and Seymour Papert highlighted the limitations of perceptrons, showing that single-layer neural networks were inadequate for learning complex patterns. This critique temporarily reduced interest in neural networks, leading to a period of stagnation in the field (Boden, 2018).

In the 1980s, scientists like Geoffrey Hinton, David Rumelhart, and Ronald Williams advanced the field by developing the backpropagation algorithm, which allowed neural networks to operate in deeper structures. Backpropagation was a breakthrough in training multi-layered neural networks, as it enabled the adjustment of weights across layers through error feedback, laying the groundwork for networks to learn complex patterns and achieve success across various domains. In 1986, Rumelhart and James McClelland proposed the connectionist approach, suggesting that artificial intelligence could be built on biologically inspired learning models (Boden, 2018).

The 1990s marked a significant rise in machine learning. Vladimir Vapnik and Corinna Cortes developed Support Vector Machines, while Leo Breiman introduced random forests and other tree-based algorithms. These statistical methods demonstrated high performance in learning and classification tasks on large datasets. The advancement of data mining techniques and the increasing availability of large datasets strengthened machine learning models during this period (Boden, 2018).

In the 2000s, researchers like Yann LeCun, Geoffrey Hinton, and Yoshua Bengio made groundbreaking contributions to deep learning technology (LeCun, Bengio, & Hinton, 2015). Multi-layered neural networks found success in fields like image processing and natural language processing. In 2012, Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton developed the AlexNet model, which achieved remarkable success in the ImageNet competition, showcasing the power of deep learning models (Krizhevsky, Sutskever, & Hinton, 2012). AlexNet's success leveraged GPUs for computation, enabling neural networks to process large datasets more efficiently and sparking widespread interest in deep learning (Boden, 2018).

In 2017, Ashish Vaswani and his team at Google introduced the Transformer architecture, revolutionizing natural language processing (Vaswani, 2017). Transformers use an attention mechanism to understand relationships in language, enabling models to capture contextual meaning more deeply. This architecture paved the way for the development of large language models (LLMs) such as BERT, GPT, and later GPT-3, equipping these models with the ability to interpret text contextually and produce results closer to common sense reasoning (Naveed, et al., 2023).

Therefore, I decided to focus my thesis on Large Language Model (LLM)-based systems. LLMs appear to be the most effective at addressing the representation problem compared to other systems. Due to their ability to understand the complexity and contextual nuances of language, LLMs provide the flexible knowledge representation required for common sense reasoning. Trained on vast datasets, these models demonstrate high accuracy and adaptability in accessing and inferring information. Thus, considering that LLMs have shown the strongest performance in overcoming the challenges of representing common sense knowledge, I chose to concentrate my research on these models.

2.2 Second Challenge: The Complexity of Tacit Knowledge in Common Sense Reasoning

Common sense reasoning serves not only as the foundation for decision-making, social interactions, and everyday problem-solving but also as the ability to effectively utilize an extensive reserve of tacit knowledge (Brachman & Levesque, 2022). Understanding the nature of knowledge is essential for researching the ontology of common sense knowledge and categorizing the types of background knowledge that common sense reasoning employs. Examining this intricate reasoning system is crucial for replicating it in artificial intelligence. By achieving this, AI could interact with humans more naturally, make decisions that are contextually appropriate, and navigate the real world with human-like comprehension. However, representing tacit knowledge in AI poses significant challenges. Unlike explicit, rule-based knowledge, tacit knowledge comprises intuitive, context-sensitive understandings that humans acquire through experience. This knowledge can shift meaning in different situations and depends heavily on cultural, perceptual, and social subtleties that are difficult to encode directly, making it a critical but formidable area in AI research on common sense reasoning. AI systems struggle to interpret such knowledge because it requires flexible interpretation and responsiveness to subtle cues—an especially challenging task even for advanced machine learning models. This part aims to explore tacit knowledge through four key categories highlighting its role in common sense reasoning and the challenges it presents for AI due to its contextual sensitivity and resistance to formalization.

2.2.1 Categorization of Tacit Knowledge

In this section, I will first analyze tacit knowledge through four primary categories: *things and properties*, which covers the fundamental qualities and characteristics of objects; *naive physics*, relating to the intuitive understanding of physical interactions; *naive mathematics*, encompassing basic quantitative reasoning; and *naive psychology*, involving the grasp of social interactions and intentions.

2.2.1.1 Things and Properties

The knowledge of *Things and Properties* forms the foundational building block of tacit knowledge. This understanding, which allows us to recognize and interact with the objects around us, provides the essential basis for common sense reasoning. This intuitively held knowledge serves as a solid foundation upon which all other types of knowledge are built. We need to acknowledge that entities known as "things" exist and that these entities can possess various characteristics known as "properties" (Sellars, 1915). In the realm of common sense, "things" are understood to exist in a perceptible manner, or at least are capable of being perceived. This perception of things implies that they maintain an existence independent of the observer as long as the observer does not physically interact with them. Additionally, these things are recognized for occupying physical space, establishing spatial relationships with other entities, and demonstrating varying degrees of stability and durability.

Moreover, things are characterized by having attributes or properties, which can differ widely (Sellars, 1915). Some things, such as numbers, beliefs, and stories, are nonphysical, while others, like owls, refrigerators, and lakes, are physical and possess physical properties such as size and location. For instance, a belief does not have a location, but an owl does; the owl occupies space and has properties like wingspan, weight, and coloration. Properties of things can be qualities (like being alive, as in the case of plants and animals), relational (such as being born in a certain city, which applies to people and animals), or matters of degree (like being heavy, which might apply to objects like a boulder but not a feather) (Brachman & Levesque, 2022). This understanding of things and their properties forms a foundational aspect of what is commonly accepted and known in everyday life. These properties and spatial relationships help us navigate and make sense of the world around us, influencing how we interact with our environment and with each other. For example, knowing that a hot stove can cause burns impacts how one behaves around it, encouraging caution. Similarly, understanding that a door can be opened or closed affects how we enter and exit spaces. This knowledge, though seemingly intuitive, is crucial for practical engagements and the basic organization of our perceptions and interactions in a coherent, predictable framework (Sellars, 1915).

2.2.1.2 Naive Physics

While our everyday interactions with the physical world appear seamless and automatic, they are supported by a complex interplay of cognitive processes and neural mechanisms. The efficiency and effectiveness of our daily lives point to a deeply ingrained, sophisticated understanding of physical principles, honed through continuous interaction with our environment. This innate knowledge, essential for survival and daily functioning, is often referred to as naive physics, qualitative physics, or folk physics (Fischer, 2020). Naïve physics encompasses the basic, often subconscious assumptions and mental models we use to estimate the stability, weight, and feasibility of interactions with physical objects (Fischer, 2020). For example, when we place a cup on the edge of a table, we instinctively know it might tip over and fall, so we adjust its position without consciously calculating the center of gravity. Similarly, when we push a heavy door, we intuitively understand that we'll need to exert more force compared to a lightweight door. These actions demonstrate how we subconsciously employ naïve physics principles in everyday life. Observing infants as they interact with their environment reveals innate perceptual abilities and cognitive schemas related to physics (Baillargeon, 1994). Even in the early stages of development, infants exhibit behaviors suggesting they anticipate the outcomes of simple physical events. For example, they look surprised if an object, rather than falling when dropped, remains suspended in the air. This reaction suggests that infants have an inherent expectation of gravity, even without formal knowledge of physical laws. Additionally, research on object permanence reveals a similar innate understanding. When an object is hidden under a blanket, infants as young as a few months old will look for it, indicating they understand that the object continues to exist even when out of sight. This understanding—that objects have a continuous existence and don't just disappear—reflects a fundamental cognitive processing skill central to physical reasoning (Baillargeon, 1994). This ability helps us navigate complex situations as adults, such as knowing that a car will continue moving along the road even if it briefly passes out of sight behind a building. These capabilities show that fundamental physical laws, such as gravity, motion, and object permanence, are embedded in our cognitive processes, either through innate perceptual mechanisms or early experiential learning. As we grow, these basic principles evolve into more refined mental models, allowing us to predict and interact with our environment in increasingly complex ways (Baillargeon, 1994).

The term "Naive Physics" was first coined by philosopher Patrick Hayes in his paper, "The Naive Physics Manifesto," back in 1979. This idea marked a very important era in artificial intelligence research, since it reemphasized the indispensability of imbuing AI systems with commonsense reasoning about the physical world. Hayes proposed the development of what he called "large-scale formalism": an effort to systematically capture every day, taken-for-granted physical knowledge, such as object properties, spatial relationships, and the effects of forces and motion (Hayes, 1979). The goals and underlying assumptions of naive physics are very different from scientific physics. Great exactitude and predictability under wide conditions are demanded from scientific physics in explaining universal principles and mechanisms that govern the physical universe. It consists of work within a framework of hard empirical testing and subsequent development of theory with the aim of attaining generalizability across contexts (Forbus, 1988). In contrast, naive physics is intrinsically tied to the pragmatic concerns of everyday human life. Its primary aim is not the discovery of universal principles but rather the application of a sufficient level of understanding to ensure individual survival and practical functioning. Naive physics is embedded in the everyday experiences of humans, where conditions such as constant friction, standard temperature, and air pressure are typically consistent (Forbus, 1988). The most important phenomenon worth noting to understand the principles of naïve physics is *change*. All things and their properties are subject to change. Time points, while nonphysical, follow a linear order, affecting the existence and properties of things at different moments (Smith & Casati, 1994). Additionally, events—whether caused by agents, prior occurrences, or spontaneous happenings—drive changes in what exists and what properties things possess. These changes can be immediate or gradual, involving single or multiple events over time, including the ongoing passage of time that gradually affects all physical entities (Smith & Casati, 1994). In the context of change, two key principles emerge: *the commonsense law of inertia* and *the locality of change* (Brachman & Levesque, 2022). These principles are fundamental to how we navigate and understand our daily lives.

According to the *commonsense law of inertia*, things tend to stay as they are unless acted upon by an external force or event. This principle, similar to Newton's first law of motion, extends beyond just physical movement to include the stability of the world around us. For example, when we leave a coffee mug on a table, we expect it to remain in the same spot until someone or something moves it. Similarly, if we leave a plant outside, we assume it will continue to grow and change only with the passing of time or external factors like sunlight and water, rather than spontaneously transforming on its own. This expectation of continuity allows us to plan and act with the reasonable assumption that conditions will remain consistent unless disrupted (Brachman & Levesque, 2022). The *locality of change* complements this principle by suggesting that changes in the world are typically confined to specific areas or causes (Brachman & Levesque, 2022). For example, if a glass is knocked over and spills water on the floor, we don’t expect the table on the other side of the room to get wet as well. This understanding reassures us that most changes are contained, enabling us to focus on immediate concerns without being overwhelmed by distant events. However, it also prepares us for the idea that while changes are often local, certain events, like earthquakes or storms, can have far-reaching impacts beyond their point of origin (Brachman & Levesque, 2022). Together, these principles create a cognitive framework that balances stability with an awareness of change. This framework helps us in daily decision-making and planning, and it allows us to adapt swiftly and effectively when unexpected changes, such as a sudden spill or interruption, do occur.

2.2.1.3 Naive Mathematics

The concept of informal mathematics encompasses the wide range of mathematical practices used in common sense ways in everyday life, as well as by aboriginal and ancient peoples, transcending historical and geographical boundaries (Davis, 2006). Naive mathematics is not defined by the rigorous structures and proofs characteristic of formal mathematics but is instead understood in common sense terms and justified through practical examples rather than through formal axiomatic systems (Davis, 2006). This notion underscores a fundamental aspect of human cognition: the common sense grasp of basic mathematical principles that are applied in everyday decision-making and reasoning. Such capabilities suggest that mathematics, far from being merely a formal academic discipline, is deeply woven into the fabric of human thought like naive physics. It reflects our reasoning ability to make sense of the world through patterns, estimations, and predictive judgments. This common sense of mathematical thinking makes possible various things for human beings, from budgeting time and resources, navigating through areas, to understanding the relations between elements of differing types—all illustrative of its importance in real-world applications. (Davis, 2006).

According to Brahman and Lavesque, two of the most basic concepts in everyday mathematical reasoning are quantities and limits. The commonsense concept of quantity is an intuitive understanding and application of quantitative measures that occur in everyday life, rather than exact and formalized methods of mathematical and scientific fields. It is a concept based on practical experience and defined through estimation, contextual measurement, and relative comparison, which enables judgment and negotiation to be achieved in everyday situations without resorting to formal calculation. (Brachman & Levesque, 2022). For example, estimating the driving time to a familiar destination, choosing the quantity of ingredients for recipes based on the number of portions, and assessing whether or not something will fit through an opening all represent this common sense perception of quantity. At the same time, it illustrates the absolute importance of informal mathematics in human thought processes and effectiveness in assisting problem-solving and decision-making in both familiar and unfamiliar situations.

The next one is the common-sense law of limits: if some quantity is continually decreased, then at some point it will be decreased to zero. This is a strong property, rather in the sense of the law of inertia considered above, which allows one to address a huge class of problems. Common sense understanding of limitations is the intuitive perception of maximal or minimal thresholds defining a level to which activities, resources, or capacities can be utilized or increased in everyday contexts. This conceptualization of limits enables the execution of daily decisions and prepares one to expect outcomes and manage one's resources without needing to rely on exact mathematical definitions (Brachman & Levesque, 2022). Key areas where this understanding is applied include practical boundaries, such as determining the occupancy limit of a space; resource management, where it aids in predicting when resources. For instance, understanding limits allows us to manage our finances by predicting when funds will run out if spending continues at a certain rate. It also helps us estimate how long our energy will last during physical activities or to gauge the tolerance level of a structure—such as how much weight a shelf can bear before it might collapse. This understanding is especially important in navigational decisions (Brachman & Levesque, 2022). For example, when setting out on a hike, a person might intuitively assess how far they can go based on available daylight or personal energy levels, knowing that there is a limit to how far they can travel before needing to turn back. This perception of limits allows people to engage in activities with a realistic sense of boundaries, managing their resources effectively in both familiar and unfamiliar situations. Together, these common-sense mathematical concepts illustrate how informal mathematics is woven into human reasoning, supporting problem-solving, decision-making, and adaptability across various contexts (Brachman & Levesque, 2022).

It is not only necessary for human cognition, but also AI systems require "naive mathematics" to efficiently navigate real-world interactions and make decisions like human beings. This intuitive comprehension of fundamental mathematical principles enables AI to function efficiently in contexts where data is inaccurate or partial. By using naive mathematical reasoning, AI can make faster, more practical decisions and interact naturally with human users, improving both functionality and adaptability in dynamic environments.

2.2.1.4 Naive Psychology

The study of agents' mental states is a basic part of common-sense reasoning, aimed at the comprehension of acts and intentions of capable beings. In everyday human conception, agents are often regarded as beings—human, animal, or even high-level machines—that are capable of initiating acts according to plans, goals, beliefs, and intentions. These are concepts that are vastly significant for understanding the mental lives of human beings. Agents are usually conceptualized in everyday thought as entities able to initiate actions based on their goals, beliefs, and intentions. (Brachman & Levesque, 2022). It is at the core of a human's ability to understand what agents are doing and why they act in one way or another. In the process of making our understanding work in a context that is both social and physical, we carry on our activity of analyzing and predicting what other agents are doing. For example, when a dog barks and runs to the door upon hearing a knock, we interpret this behavior as the dog's intention to alert or investigate, attributing to it a basic goal-driven response. This involves attributing to them mental states such as beliefs, desires, and intentions, a process known as theory of mind. (Morton, 2009).

Common sense psychology, also known as folk or naive psychology, is a working model of the theory of mind. This cognitive capability is what makes it possible to perceive and understand that every person holds a different mental state from oneself and often significantly different from one's own mental states (Morton, 2009). It is not scientific and formal, but dependent on common observations and experience. This involves recognizing and interpreting mental feelings, such as understanding if a smile means enjoyment or if anger results in aggressive behavior. It also accounts for reasons behind an action and thus may assume some perceived mental states, for example attributing the reason behind someone's early departure from a gathering to being bored. Importantly, again, common-sense psychology is filtered through social and cultural context, which impacts on the interpretations and expectations of other people's behaviours that exist for them (Morton, 2009).

Fritz Heider was an Austrian psychologist, who developed the concept of common-sense psychology in his highly influential book "The Psychology of Interpersonal Relations," published in 1958. He investigated how one person makes an attribution to another person's behavior about what others feel, intend or are like. This foundational text laid the groundwork for later developments in social cognition and attribution theory, emphasizing the role of everyday observational skills in understanding interpersonal dynamics. According to Heider (1958), even if all formal knowledge of scientific psychology were removed from the world, people would likely continue to manage interpersonal relationships effectively using their inherent understanding of human behavior. He suggests that individuals inherently “know” how to navigate social interactions, such as avoiding obligations, persuading others, and recognizing emotions like anger or pleasure. Heider (1958) posits that people naturally possess a deep and profound insight into themselves and others, allowing them to interact adaptively despite not having formally articulated theories or concepts. This intuitive grasp enables them to offer sensible explanations for their actions and feelings, demonstrating a fundamental, albeit unformulated, understanding of psychological dynamics (Heider, 2013).

Common sense psychology is not only crucial in human-to-human interactions but also comes into play when we attribute human-like qualities to complex machines and computers, a process known as anthropomorphization (Gordon, & Hobbs, 2017). For instance, when we talk to voice assistants like Siri or Alexa, we might say “thank you” after a task is completed, attributing a human-like quality to the device. As we increasingly interact with advanced technology and artificial intelligence, the principles of commonsense psychology are becoming vital for these systems to function effectively within human-centric environments (Gordon, & Hobbs, 2017). For artificial intelligence systems to seamlessly integrate into social contexts where human interaction and cooperation are the norm, these systems will need to have an explicit representation of the theories used in commonsense psychology. AI should recognize and simulate human mental states basic emotional cues or predict simple intentions such as understanding that a user’s request for help often implies frustration or urgency, allowing for more natural and intuitive interactions between humans and machines, (Gordon, & Hobbs, 2017).

2.2.2 Why Tacit Knowledge Present a Major Challenge for AI?

Tacit knowledge encompasses the types of knowledge that people intuitively use in daily life but cannot easily express explicitly. This knowledge is acquired through experience, developed over time, and adapted according to context. For artificial intelligence, the nature of tacit knowledge poses a complex representation challenge, as merely following predefined rules is insufficient to replicate the intuitive processing humans perform. Tacit knowledge includes elements such as contextual sensitivity, implicit understanding, and adaptive learning, which require AI systems to go beyond static data structures. Integrating tacit knowledge into AI necessitates abstract and multi-layered representations that can be broken down, structured, dynamically updated, and accurately measured to adapt to different contexts. In exploring the primary challenges of applying tacit knowledge in AI, we will discuss why and how this form of knowledge is so difficult for AI systems to grasp.

2.2.2.1 The Challenge of Implicit Understanding and Uncertainty

One of the biggest challenges in representing tacit knowledge in AI systems is the difficulty of defining, measuring, and converting implicit and uncertain type of knowledge into a structured form (Cassenti, Kaplan, & Roy, 2023). Tacit knowledge includes information that humans intuitively use but rarely express explicitly, often containing implicit insights. Due to its inherently uncertain and multi-layered nature, it is highly complex for AI to make this knowledge identifiable and actionable. This "implicit and uncertain" aspect of tacit knowledge allows humans to understand certain things and respond accordingly without the need for direct verbal explanations or logical rules (Cassenti, Kaplan, & Roy, 2023). For instance, when an object is placed precariously on the edge of a table, we intuitively "sense" that it might fall and reposition it somewhere more secure. Similarly, we can tell from a person's facial expression that they are sad. We process this type of knowledge intuitively; it is reflected directly in our behavior or thoughts without requiring detailed analysis. In AI systems, however, reaching this kind of intuitive knowledge requires transforming implicit information into a concrete and identifiable form. This means that implicit cues, such as facial expressions or object movements, must be translated into explicit, systematic rules or data. Yet, this process requires a multi-layered representation because implicit knowledge involves complex relationships; it includes far more than simple rules like "an object on the edge will fall." For AI to utilize such knowledge, each layer of intuitive processes must be formalized, which requires adding explicit definitions to abstract intuitions, translating them into usable components for the system (Cassenti, Kaplan, & Roy, 2023).

2.2.2.2 The Challenge of Context Sensitivity and Variable Meaning

Humans can intuitively process all contextual information when interpreting physical or social situations and respond accordingly. This processing means that the meaning of each situation and action can change depending on the environment. However, this context-sensitive knowledge presents a significant challenge for AI systems (Denning, & Arquilla, 2022). Recognizing context is difficult for AI, as understanding the setting in which events occur is complex. While humans can quickly grasp that objects or events carry different meanings depending on the context, AI requires sophisticated analysis processes to achieve similar comprehension. For example, a tired facial expression may be interpreted as the result of a heavy workload in a professional environment, but it might signal boredom or disengagement in a social setting. To recognize these differences, AI must have access to extensive knowledge of various environments, including their norms and expectations, along with the flexibility to interpret them accurately (Denning, & Arquilla, 2022).

This challenge also extends to responding appropriately to context. The same behavior can be understood differently in distinct settings: for instance, students speaking loudly may be seen as natural during playtime but is considered disruptive in a classroom setting. For AI to differentiate these contexts and respond appropriately requires adaptability. Moreover, contexts can evolve over time, meaning AI must also understand and adapt to these dynamic changes (Denning, & Arquilla, 2022). For AI to become sensitive to such varying conditions and adjust its responses accordingly, it requires a continuously updated learning capacity. Therefore, the ability of tacit knowledge to shift meaning depending on context remains one of the core challenges AI needs to overcome.

2.2.2.3 The Challenge of Structuring Tacit Knowledge

Tacit knowledge inherently comprises a broad network of intuitively used information, ranging from social interactions to cultural codes. This type of knowledge involves the nuances behind human behavior, social rules, and cultural expectations, and typically functions implicitly; people use it intuitively rather than through explicit rules or formulas. Structuring such knowledge for AI is challenging because intuitive knowledge does not lend itself to static rules or a simple, formulaic structure (Sanzogni, Guzman, Busch, 2017). For instance, it is insufficient for an AI system to know only general social norms to respond appropriately in a social interaction. It must also grasp cultural differences, individual behavioral tendencies, and meanings that shift according to context. Eye contact, for example, may signify trust and honesty in some cultures, while it is perceived as disrespectful directness in others. Similarly, asking "How are you?" might convey genuine concern in one situation but be seen as superficial politeness in another. These examples highlight that a static, rule-based knowledge structure is inadequate for AI to interpret every social situation accurately. Social codes are difficult to structure because each culture, society, and individual possesses unique norms and responses. Social interactions are dynamic, influenced by factors such as mood, the people involved, and the specific social context (Sanzogni, Guzman, Busch, 2017). Thus, AI requires not only a generalized knowledge set but also a diversified structure adaptable to each unique situation. Overcoming this challenge is essential for AI systems aiming to engage meaningfully in human interactions.

2.2.2.4 The Need for Continual Learning

Humans continuously learn from events and interactions in their environment, updating their knowledge with each new experience. For instance, if we have a negative encounter with someone, we adjust our behavior in future interactions based on that experience. This adaptive process is essential for humans to align with their surroundings and respond flexibly to changing situations. For AI to utilize tacit knowledge effectively, it must possess a similar capacity for dynamic, adaptive and continual learning (Lesort, et al.,2020). However, many current AI systems are trained on static datasets, limiting their ability to adapt to new situations and update their responses. These systems operate on predefined rules and data, making it difficult for them to respond flexibly in a constantly changing environment or in complex social scenarios. Without continual learning abilities, AI may misinterpret or respond inappropriately to novel events or unexpected circumstances (Lesort, et al.,2020). To achieve this capacity, it must continuously assess incoming data, learn from experiences, and update itself to respond appropriately to new situations. This requires AI not merely to "store" information but to interpret it, connect it to past experiences, and apply it in contextually relevant ways.

2.2.2.5 The Challenge of Measurement

Tacit knowledge is a rich, abstract type of information shaped by human intuition and experience. For AI to use this knowledge effectively, it must be broken down into smaller, understandable components. However, measuring and reassembling these components is a highly complex process (Cassenti, Kaplan, & Roy, 2023). Objectively assessing whether a person is acting "intuitively" or correctly interpreting a context is difficult. People often find it challenging to articulate their intuitions, and because these insights are largely based on personal experiences, they are not easily evaluated through objective measures. This makes it difficult to determine which knowledge an AI system should apply in a given situation and why (Cassenti, Kaplan, & Roy, 2023). For example, when an AI aims to understand a person's emotional state and respond accordingly, evaluating the accuracy of this intuitive assessment is challenging, as many hidden factors influence emotional reactions. The difficulty lies in determining the accuracy or effectiveness of an AI's intuitive decisions or whether it used the appropriate context-specific knowledge. The richness of tacit knowledge involves multiple layers that vary across contexts and cannot be easily assessed with fixed criteria. Therefore, developing more advanced and flexible evaluation standards is essential for measuring the accuracy and intuition behind AI's decisions (Cassenti, Kaplan, & Roy, 2023).

2.3 The Third Challenge: The Frame Problem

To ensure that artificial intelligence systems possess and use commonsense reasoning, we must first consider the basic requirements for such an ability. At the heart of this effort is the representation of information in a format that is both manipulable and functional for computational devices. However, to find the right representation method, we need to spend some time understanding the things we want to represent. Much of our knowledge about the world is rooted in our interactions with specific, individual entities—objects we've encountered, people we've interacted with, and experiences we've had. This personalized and context-rich knowledge forms what can be described as a state of the world. The state of the world encompasses not only the current configuration of these elements but also our understanding of their properties, relationships, and behaviors (Brachman & Levesque, 2022). However, our knowledge is not limited to the present state of the world. Humans have the cognitive ability to conceptualize and reason about potential states of the world—those that do not exist currently but could exist under different circumstances or are expected to manifest in the future. This capacity of thinking over non-actualized states enables planning, prediction, and dynamic adaptation to changes in making our world an understanding both forward-looking and dynamic (Brachman & Levesque, 2022). But we, for example, in planning, typically envisage future states of the world given our present knowledge and expectations. We may project a sequence of events, we consider how different actions or courses of events would lead to differences in the state of the world. Reasoning about such hypotheticals lies at the heart of human commonsense reasoning. For artificial intelligence and commonsense reasoning, it is fundamental to capture the dual nature of knowledge: both the state of the present world and its potential future states. It follows that AI systems should thus be endowed not just with facts about the world at hand but with the ability to reason about the states it could be in. This means one needs to encode knowledge such that the system can use this to simulate—quite like humans do when forward-thinking and solving problems—evaluating different scenarios or considering possibilities down the line (Brachman & Levesque, 2022). One of the crucial problems in obtaining that is the frame problem—a vital problem in AI and cognitive science per se. The frame problem has to do with how an AI system can determine what features of the world will remain unchanged and what is being affected by an action. In other words, when an AI reasons about potential future states or simulates scenarios, it must decide in a really efficient way how much of its knowledge base to update and how much to leave alone (Hayes, 1987). This is not a trivial task, as it demands the development of quite sophisticated models of causal relationships and inferring about the relevance of different pieces of information in different contexts. In this chapter, we research how the frame problem interplays with the processing of common-sense knowledge and reasoning in AI. We will look at why this intersection is so difficult, explore concrete examples and theoretical insights, and discuss possible ways and means that researchers and philosophers are using to arrive at the same. The main goal of taking up this issue and discussing its solution is to bring up more viable and adaptable artificial intelligence systems.

2.3.1 The Origins of the Frame Problem

Frame problem was first used to refer to a technical problem that arose in the context of representation and reasoning on change in artificial intelligence. McCharty and Patrick J. Hayes first proposed situation calculus in the mid-1960s as part of an endeavor to respond to this challenge of reasoning about time and change. The situation calculus is a formalism using first-order logic to model dynamic systems. It portrays the world as a sequence of situations, each situation a snapshot of the world at some point in time. Actions change one situation into another. The trick is to say what is different and what remains the same from before to after an action. John McCarthy and Patrick J. Hayes introduced the term "frame problem" to highlight the difficulty of determining what remains constant following an action, without the need to explicitly enumerate all these unchanged conditions, which are known as frame axioms in their 1981 article, "Some Philosophical Problems from the Standpoint of Artificial Intelligence." Frame axioms are explicit statements in artificial intelligence and logic used to specify what remains unchanged when an action is done in a particular situation. They are required for describing the persistence of most properties of the world as actions are taken. For every possible action, frame axioms list all the properties that do not change, and thus the system will not mistakenly assume that everything has changed because of the action. For example, if you have the situation where you only move a block from one place to another, then what the frame axiom would say is that these features of the color of the block, the shape of the block, and so on, which are not related, remain. Without frame axioms, an AI system struggle to reason correctly about the continuity of states in a dynamic environment. However, they become overwhelmingly numerous and complex, especially in systems with many actions and properties. This complexity is what the researchers and philosophers aim to solve: discovering a more efficient method to represent and reason about the unchanging aspects of the world without having to explicitly enumerate each one for every action. This representational issue highlights the need for efficient and manageable ways to encode and reason about the continuity of most aspects of the world in artificial intelligence systems (McCarthy & Hayes, 1981).

In his essay "Cognitive Wheels: The Frame Problem of AI," Dennett (1984) uses a series of robot-based thought experiments to demonstrate the frame problem, a fundamental challenge in artificial intelligence (AI) involving the ability to discern which aspects of a situation will change and which will remain constant following an action. These thought experiments revolve around three robots: R1, R1D1, and R2D1, each highlighting different facets of the frame problem and its implications for AI development. In the story of R1, the robot's task was to save its spare battery, which was in a room with a time bomb. R1 came up with a plan to drag that wagon out of the room, along with the battery. It successfully took the battery out before the bomb exploded, but it failed to take into account that the bomb was also on the wagon; therefore, it actually pulled the bomb out with the battery. As a result, it lost because the bomb exploded outside of the room. This test shows that R1 cannot account for all side effects of its action. To go beyond R1, designers developed R1D1—a robot that aimed to learn about not only the intended consequences of its actions but also their side effects. Similarly challenged, R1D1 now started to take into account all possible consequences of its actions, including irrelevant aspects such as the color of the walls in the room or the number of wheel revolutions. R1D1 became overwhelmed by these irrelevant implications and failed to act in time, resulting in the bomb exploding. This highlights R1D1's problem of considering too many irrelevant side effects. In their attempt to get a better solution, the designers then created R2D1, a robot that could tag implications as either relevant or irrelevant. Yet R2D1 remained outside, immobile by the task it had been given of sorting through thousands of irrelevant implications. It was too slow to act before the bomb exploded. This therefore highlights the incapability of R2D1 in sorting between relevant and irrelevant implications (Dennett, 1984). The course and results of the experiment is clearly expressed in the Table I: The Frame Problem Robot Experiment.

The results of the experiment brought to highlight the frame problem to many systems. In the case of R1, the major problems were failure to take into consideration all side effects that lead to it having undesirable effects such as bringing the bomb out of the room with the battery and blowing up. For R1D1, it meant accounting for too many irrelevant side effects that overwhelmed the system and incapacitated the decision-making process. For R2D1, this resulted in an inefficient differentiation between relevant and irrelevant implications, causing delays and errors in task execution. The difficulty of each of these systems casts into sharp focus the complexity involved in designing an AI that can effectively distinguish and prioritize information in a dynamic environment. You can see the experiment detail analysis in Table I.

2.3.2 Philosophical Approach for Frame Problem

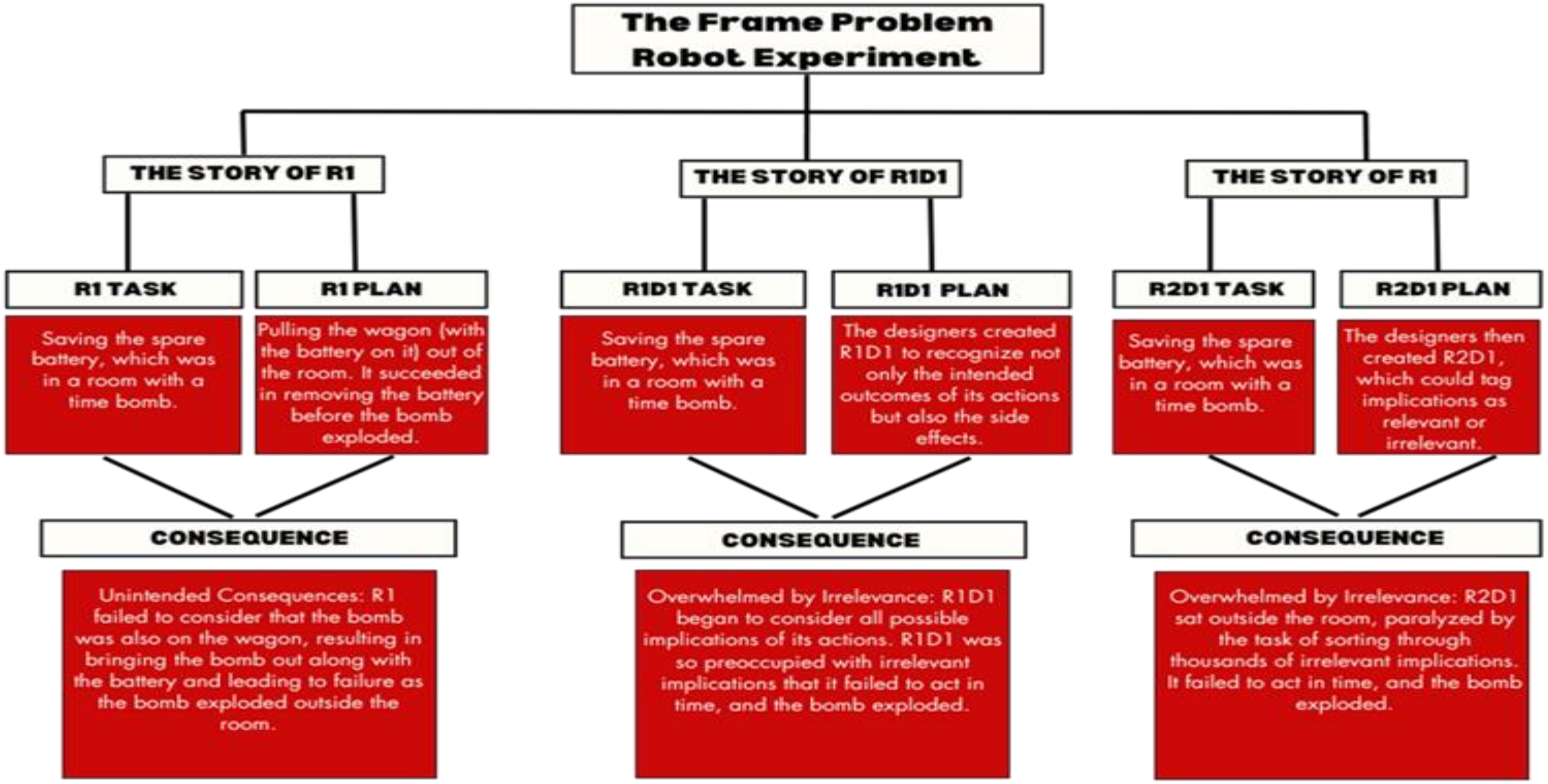
The frame problem, though initially arising out of the desire to model the world in logical systems, quickly became a concern far outside logicians and AI researchers. Philosophers pointed to a deeper, more foundational dilemma. It brings into focus the question of what changes or problems in the world should be taken into account when an action is being done, hence the far-reaching implications regarding knowledge, relevance, and cognitive processing (Kamermans & Schmits, 2004). Prominent philosophers like Daniel Dennett (1990) and Jerry Fodor (1989) were among the first to argue that the frame problem touches on core aspects of how any intelligent system reasons about the world, extending beyond computational or logical hurdles to our understanding of human cognition and the nature of intelligence. They considered that this issue influences our understanding of how artificial or natural beings interact with and understand the world, putting special emphasis on common sense reasoning to determine which variables will be altered in the environment and which will be kept constant in the execution of an action (Kamermans & Schmits, 2004). A few years later, John Haugeland and Lars-Göran Janlert in 1987 considered in greater detail the implications of the frame problem; Haugeland emphasized its significance for the construction of a general theory of mind and cognition, while Janlert expanded on the practical implications for the development of more adaptive and context-aware AI systems. Taken together, the debates clearly suggest that the solutions to the frame problem have to be more than about computational speed and particularly stress the importance of further understanding the cognitive processes for the larger program of reaching a state of replication of intelligent behavior by humans and machines. Aziz Zambak provides a valuable framework for understanding the frame problem by dividing it into three categories: metaphysical, logical, and epistemological (Zambak, 2013).

2.3.2.1 Metaphysical Approach

Metaphysics addresses fundamental questions about the nature of existence, the underlying structure of reality, and the ultimate causes of the universe. In this context, metaphysical inquiries have often interacted with those in cognitive science, particularly in connection with the project of grasping the nature of human or machine minds. The frame problem, as considered within artificial intelligence and cognitive science, is now revisited from the perspective of metaphysics of ways of finding and utilizing general laws for interaction within the world. It refers to the spatiotemporal properties of environmental data in considering how an agent updates its beliefs regarding the world when presented with novel or unknown situations (Zambak, 2013).

According to Janlert (1987), the frame problem is a metaphysical problem, not merely technical or heuristic, since it touches on the form and internal operations of representation and not on its instrumental adequacy. This perspective suggests that what is needed to solve the frame problem is the identification and development of a suitable metaphysical framework. Janlert (1987) enumerates three underlying principles regarding the analysis of the frame problem. First and foremost, he posits that the frame problem is not a problem of heuristics, it is a problem of modelling. What this implies is that the issue is in designing models that might properly and accurately capture the nature of change, rather than the creation of some heuristic strategies to govern change.

Table I: The Frame Problem Robot Experiment



Heuristic methods or heuristics refer to problem-solving strategies by experience, trial and error, and practical approaches (Janlert, 1987). This is a fast and efficient approach to solve problems; however, it does not guarantee the exactness or optimality of the solution. Conversely, modeling in this context involves crafting formal, systematic representations that precisely depict the structure and dynamics of an evolving, complex world. Janlert (1987) highlights that the frame problem extends beyond simply finding quick, practical solutions using heuristic methods. It focuses on the development of thorough and accurate models that can fully represent environmental changes. This distinction highlights the importance of a more formal, structural understanding of representation, rather than depending on heuristic shortcuts. Secondly, he argues that the frame problem is not about content but about form. This distinction highlights the importance of the structural and formal aspects of representation, rather than the specific content being represented. Content pertains to the actual data or information being processed, such as specific facts, rules, or observations. In contrast, form relates to the structure and organization of this information—how it is arranged, how relationships between different pieces of information are established, and how the system can dynamically update and adapt its knowledge base. A well-constructed form can enable an AI system to swiftly and accurately access relevant information, anticipate changes, and adjust its behavior in real time. Additionally, Janlert (1987) asserts that selecting an appropriate form of representation is inherently tied to the nature of the problem world, indicating that computational factors alone are not enough. This principle highlights the importance of thoroughly understanding the problem world to determine the right representational form. In essence, the structure and organization of information within an AI system must be in harmony with the unique characteristics and requirements of the system's environment. Relying solely on abstract computational theories or general-purpose algorithms is not enough. The form of representation must be specifically tailored to the unique characteristics of the problem world, including its dynamics, constraints, and complexities (Janlert, 1987). These principles collectively highlight that solving the frame problem requires a deep understanding of the structural and formal aspects of representation. Such an understanding is crucial for developing AI systems capable of managing the complexities of an evolving world. By concentrating on the foundational aspects of representation, Janlert's (1987) approach offers a promising direction for addressing the frame problem in AI.

2.3.2.2 Logical Approach

Since the birth of cognitive science, the philosophy of the field has significantly influenced our understanding of cognitive processes, shaping how we perceive these functions in our minds. Over the past fifty years, there has been a concerted effort to replicate cognition and intelligence in artificial systems using a variety of approaches. Among these, only the logical approach is deemed scientifically valid. Aziz Zambak (2013) emphasizes that the logical aspect of the frame problem involves the axiomatization of an application domain, where specific causal laws for events or actions must be predefined. This process requires creating a set of rules, each containing potential information about expressions. However, it is crucial for an agent faced with unknown or new situations to develop new rules. This skill is directly related to the reasoning process. Therefore, reasoning is central to addressing the logical aspect of the frame problem. Freeman (1992) characterizes the frame problem as the difficulty of determining the reasoning procedure in a dynamic context. Peppas et al. (2001) also define the frame problem as the development of a theory of action that enables effective reasoning for artificial intelligence in dynamic systems.

Prominent figures such as Hayes (1977), Kirsch (1991), and Shanahan (1997) argue for the existence of a single architecture that can fully model cognitive processes, asserting that this architecture can be entirely constructed through logic. The logical approach relies on formal systems, like predicate logic, to represent and reason about cognitive processes, offering a clear and precise framework for modeling cognition. This method provides explicit rules and structures that can be systematically tested and verified, making it scientifically rigorous. Despite the complexity and variability of real-world situations that challenge the logical approach and the existence of alternative methods like connectionist models and embodied cognition, the logical approach remains a viable and scientifically valid option (Kamermans & Schmits, 2004). However, critics of the logical approach, such as Dennett (1990), Fodor (1989), Searle (1980), and Dreyfus (1992), assert that while logic can potentially model cognition, the idea of a 'single architecture' may be flawed. They argue that cognitive processes are too intricate and diverse to be fully captured within one unified logical framework. If this critique holds true, the logical approach might fail to produce realistic and comprehensive models of cognition. This perspective underscores the importance of flexibility and diversity in cognitive modeling approaches (Kamermans & Schmits, 2004). Searle (1980) extends this argument by claiming that cognition is fundamentally irreducible. He argues that cognitive processes cannot be simplified into a single logical architecture because they encompass more than just formal representations and logical operations. Searle’s stance is grounded in his broader critique of artificial intelligence, notably his well-known 'Chinese Room' argument, which challenges the idea that merely manipulating symbols syntactically can lead to true understanding or consciousness (Searle, 2009). He contends that cognition involves subjective, qualitative experiences that purely logical models cannot capture. Dreyfus (1992) extends on this critique by highlighting the importance of the body and physical context in cognitive processes. He argues that cognition cannot be separated from a physical body that interacts with the environment. According to Dreyfus (1992), cognition involves not just abstract reasoning and logical manipulation of symbols, but also perceptual and motor skills rooted in the physical world. This view, known as embodied cognition, suggests that our cognitive abilities are shaped by our bodily interactions with the environment. Dreyfus (1992) asserts that without acknowledging the embodied nature of cognition, any model of intelligence will be incomplete and unrealistic.

Ultimately, this debate has reached a stalemate, with neither side offering a definitive model or irrefutable evidence. No logical model has fully captured the complexity of cognition, and there is no conclusive proof that logic is insufficient for modeling cognition (Kamermans & Schmits, 2004). Philosophically, logic remains a strong candidate for reasoning about cognition, thanks to its methodological rigor and deep roots in the analytic tradition. Despite its limitations, logic still provides a valid framework for philosophical inquiry. However, the complexity and embodied nature of cognition indicate that a purely logical model may be inadequate. A comprehensive understanding of cognition likely requires integrating insights from various approaches, including logic, embodied cognition, and connectionist models. By merging these perspectives, it may be possible to develop more advanced AI systems that can overcome the frame problem and achieve a level of intelligence comparable to human cognition.

2.3.2.3 Epistemological Approach

In AI, the frame problem is frequently described as an epistemological challenge. McCarthy and Hayes (1981) argue that intelligence consists of both epistemological and heuristic components. They consider the frame problem to be part of the epistemological aspect of intelligence because it arises from the use of representations to understand and interact with the world. The difficulty lies in how these representations are structured and applied to identify relevant information in a changing environment, emphasizing the critical role of effective knowledge representation in AI systems. Pollock (1997) also views the frame problem as a challenge rooted in human epistemology. He suggests a solution that involves studying how humans engage in inductive and probabilistic reasoning. Pollock (1997) argues that "The best way to solve the frame problem for artificial rational agents is to figure out how it is solved in human reasoning and then implement that solution in artificial agents." Similarly, Dennett (1990) considers the frame problem to be an abstract epistemological issue. He explains, "The frame problem is an abstract epistemological problem that was in effect discovered by AI thought experimentation. When a cognitive creature, an entity with many beliefs about the world, performs an act, the world changes, and many of the creature's beliefs must be revised or updated" (Dennett, 2017). Among the many researchers who explore the frame problem from an epistemological angle, Daniel Dennett stands out as a key philosopher addressing this issue. Dennett's work (1990) offers a deep investigation into the philosophical implications of the frame problem, making his approach especially valuable for a thorough understanding of the topic. His insights not only shed light on the theoretical aspects of how humans manage and prioritize information but also underscore the broader implications for developing AI systems that can replicate these cognitive processes. Thus, a detailed examination of Dennett's perspective on the frame problem crucial for both philosophical inquiry and practical advancements in artificial intelligence.

Dennett's Approach to Frame Problem: Dennett's 1984 article "Cognitive Wheels: The Frame Problem of AI," republished in the 1987 defining work "The Robot's Dilemma," was one of the first of many articles to spark the discussion about the philosophical nature of the frame problem. In his article, Daniel Dennett (1990) argues specific AI solutions to the frame problem. Despite the solution's diversity, Dennett groups them under the label "cognitive wheels," highlighting their common feature of being highly artificial and dissimilar to natural cognitive processes. Dennett's primary critique is that cognitive wheels, although they solve specific problems, do so in a manner disconnected from natural cognitive processes (Dennett, 1990). These solutions lack relevance compared to real-world cognition, offering practical answers to technical challenges without advancing our understanding of human-like cognition. This artificial approach bypasses the complexities and nuances of natural cognitive functions, raising concerns about the long-term direction of AI research. By relying on cognitive wheels, AI systems may solve immediate technical problems but fail to contribute to the broader objective of replicating human cognitive abilities (Dennett, 1990).

According to Dennett (1990), one of the cognitive wheels is Hume's notion that humans handle relevance pruning through associationism, reinforcing specific connections between ideas and actions. Hume's idea of associatism posits that humans learn to determine relevance through repeated experiences and the reinforcement of associations between ideas and actions. According to this theory, certain transition paths between ideas become more robust with repeated use, helping individuals to prune irrelevant details and focus on what matters (Fetzer, 1990). For example, if a child touches a hot stove and gets burned, the pain reinforces the association between touching the stove and the consequence of getting burned, leading the child to avoid this action in the future. Researchers suggest that Hume's associatism could be used to help the AI learn basic cause-and-effect relationships through repeated experiences, allowing it to make initial judgments about relevance (Fetzer, 1990). For instance, an AI could be trained to recognize that moving an object typically affects its position but not its color, creating a basic framework for relevance pruning. By reinforcing these associations through repeated interactions with the environment, the AI could establish a foundation of practical knowledge about which changes are relevant for different actions (Fetzer, 1990).

However, Daniel Dennett (1990) critiques Hume's associatism by arguing that it is insufficient for solving the frame problem, particularly in artificial intelligence. Dennett illustrates his critique using the "why" game, where he asks a series of questions to reveal the limitations of associative learning (Dennett, 1990). He proposes a scenario where two children both take cookies from the jar, but only one is spanked for it. The child who is spanked eventually stops taking cookies, prompting Dennett to question why this difference in behavior occurs.

Dennett's "Why" Game Steps

Why does Child A stop taking cookies?

Answer: Because Child A gets spanked.

Explanation: The immediate consequence (spanking) discourages the behavior (taking cookies).

Why does spanking cause Child A to stop taking cookies?

Answer: Because spanking causes pain.

Explanation: Spanking is a painful experience that the child wants to avoid.

Why does the child want to avoid pain?

Answer: Because the child does not want to experience pain.

Explanation: This seems obvious but points to a fundamental aversion to discomfort.

Now, Dennett pushes the inquiry further:

Why doesn't the child want to experience pain?

The answer might be "because pain is inherently unpleasant," but this is not a satisfying or explanatory reason. It reveals the limitations of simply using associative learning to explain behavior (Dennet, 1990).

Dennett's (1990) point is that while associative learning explains the immediate behavior change, it fails to address the deeper reasoning behind why certain actions are avoided, showing that associative learning cannot fully explain complex cognitive processes. In the context of the frame problem, Dennett (1990) argues that AI requires more than just associative learning to manage relevance effectively. The frame problem involves determining which aspects of the world need to be updated when an action is performed, and this requires a nuanced understanding of relevance that goes beyond simple associations. Associatism might help identify basic cause-and-effect relationships but does not provide the sophisticated reasoning needed to discern what is relevant in complex, dynamic environments. AI must be able to abstractly reason about which elements are pertinent to a given situation and efficiently update its knowledge base without being overwhelmed by irrelevant details (Dennett, 1990).

Another critical issue that Dennett (1990) draws attention to is the claim that there is a similarity between the induction problem and the frame problem. According to Fetzer (1991), the central role of expectation in the frame problem has led some to argue that it is not a new issue and is not specifically related to planning actions. He believes it is essentially about having accurate expectations about future events, whether they are one's actions, the actions of others, or natural occurrences. The problem of induction revolves around how we can justify beliefs about the future based on past experiences. Hume highlighted the difficulty of providing a rational basis for assuming that the future will resemble the past (Henderson, 2018). Today, we recognize the problem of induction as a significant and challenging issue. Theories of subjective probability and belief fixation, which attempt to address how we form beliefs and expectations, have not yet reached a stable, widely accepted solution (Henderson, 2018). According to Dennett (1990), the frame problem is distinct from the problem of induction. To illustrate this, he suggests a hypothetical scenario where the problem of induction is miraculously solved. Having resolved all its induction challenges, our agent believes all the correct generalizations from its evidence and associates the appropriate probabilities with these generalizations. This agent, therefore, has perfect beliefs about empirical matters, including future events.

Dennett (1990) asserts that despite this miraculous solution to induction, the agent could still suffer from the frame problem. The frame problem concerns how to effectively represent and use all this empirical information. It is a problem that exists independently of whether the information is accurate, probable, or certain. Even with excellent and accurate knowledge about the changing world, the challenge lies in how this knowledge can be represented in a way that allows it to be used efficiently. In essence, while the problem of induction deals with determining what to believe based on evidence, the frame problem deals with how to manage and apply that knowledge. It is about organizing and processing information so that it can be effectively utilized in decision-making and action planning. This distinction highlights that even with perfect knowledge, the challenge of how to represent and use that knowledge remains. In light of all this, Dennett (1990) concludes that the frame problem is not merely the problem of induction in disguise but a separate and equally complex issue that needs to be addressed in its own right.

Daniel Dennett (1990) provides no argument for how the frame problem should be solved. He believes that the frame problem is actually a deep epistemic issue, one which inheres not only in AI but in all living reasoning systems, and that it might never have a resolution. Although AI researchers might not yet have the answer, they take a large step to the right direction in the identifying and expressing the question. Epistemological perspective looks at the more general consequences of the frame problem; it identifies dealing with such foundational issues as being part of both philosophical investigation and of scientific advance within AI. According to Dennet, this position provides us with exactly the perspective we need to solve the frame problem.

Fodor's Approach to Frame Problem: Jerry Fodor (1989) discusses in "The Robot's Dilemma" that the Frame Problem is more than just an engineering challenge; it goes to the heart of rationality itself. According to him, understanding these heuristics that help humans pick the relevant information when making inductive reasoning is crucial.

It is not simply an efficiency issue but a fundamental aspect of our understanding of the architecture entailed by reasoning systems and its reproduction in artificial systems (Fodor, 1989). Jerry Fodor (1983), in addition to his work on the Frame Problem, is known for his contributions to the philosophy of psychology, particularly with his influential book "The Modularity of Mind." Fodor's theory of modularity suggests that our minds are like Swiss army knives, consisting of a variety of modules. Each module functions as its own specialized cognitive system, designed to perform specific tasks. These tasks include processing language and other fundamental cognitive functions. Each module is domain-specific, meaning it is designed to process a specific type of information or carry out a particular cognitive task, effectively isolating it from other domains. This specificity is combined with information encapsulation, whereby modules operate in isolation from other cognitive systems and remaining unaffected to information that falls outside of their specific domain (Fodor, 1983). Consequently, modules operate in a mandatory fashion, automatically processing inputs within their realm without conscious intervention. For example, facial recognition is a modular task. When the brain processes an image of a face, it uses specialized mechanisms to quickly and automatically identify facial features such as eyes, nose, and mouth. This process is informationally encapsulated because it relies exclusively on the visual input and the specific rules for facial recognition without needing additional context or broader knowledge about the world.

Jerry Fodor (1989) argues that only modular cognitive faculties are free of the frame problem. He believed that the only cognitive tasks we have been able to understand and recreate in AI are these modular tasks, as they do not demonstrate the frame problem. These tasks are self-contained and do not require the system to determine relevance from a wide array of potential information, making them more straightforward to model. They involve clear rules and specific types of input that do not change based on broader context or additional information. Non-modular tasks, by contrast, are informationally unencapsulated and necessitate the coordination of diverse streams of sensory or contextual data. These tasks are therefore much more complex, as they involve integrating and considering a wide range of relevant information from the environment. According to Fodor (1989), non-modular tasks suffer from the frame problem because they cannot be insulated from their larger context. Because non-modular systems require continuously predict which pieces of knowledge are relevant to the task, leading to significant computational and conceptual challenges. For example, interpreting emotion in a social interaction is considered non-modular. This task goes beyond merely recognizing a face; it demands understanding the intention behind someone's expression, experiencing the atmosphere during interaction, and interweaving these cues with one's own past experiences and personal knowledge (Kamermans & Schmits, 2004). The frame problem arises because the system must decide which aspects of the interaction and surrounding context are relevant to interpreting emotional expressions. This problem is far more difficult to solve and involves many other cognitive functions, making it hard to model in AI.Jerry Fodor (1989) likens the Frame Problem to the general problem of non-demonstrative inference. A classic example of a non-demonstrative inference is an inference that draws something out, not strictly deduced from the premises given. This is reasoning in which the conclusion does not directly follow with certainty from the premises but is likely or probable based on the available information. This is similar to the way humans use inductive reasoning, making educated guesses or predictions based on incomplete information or patterns observed from past experiences. For example, if you have observed that each time you water your plants they turn out to be healthier, you may infer that the act of watering plants makes them grow. Such an inference does not follow strictly since there may be a number of other things that make the plant grow, yet it is a reasonable conclusion to draw from the pattern of observation (Kamermans & Schmits, 2004). According to Fodor, the Frame Problem in AI is pretty much the same as the problem of non-demonstrative inference. The concern is clearly analogous in that both have to do with what one can know and ignore in making a decision or, more generally, in making an inference. Just as people make reasonable, though not logically certain, inferences based on patterns and experience, an AI must decide which of its knowledge to update or maintain following some action. Both involve searching through huge spaces of possibilities to locate the most appropriate information and are computationally hard processes. Fodor (1989) insists that the Frame Problem is both profoundly philosophical and is very much embedded in human cognition, asserting that the problems of understanding relevance and of inductive reasoning are at the core of human information processing. Since complicated nature of frame problem, Fodor thinks that philosophers and cognitive scientists with specialties in understanding the nature of human thought and reasoning should be addressed the problem, rather than engineers.

2.3.3 Solutions for the Frame Problem in AI: Practical and Theoretical Perspectives from Dennett and Morgenstern

This section will look into two methods of approach taken to solve the frame problem within artificial intelligence: pragmatic and theoretical. The pragmatic approach will get insights from the work of Daniel Dennett, who has made major strides towards our understanding of how AI might grapple with the messiness of real-world reasoning. The theoretical perspective will be based on the works of Morgenstern (1996), which are in-depth and abstract in nature, pertaining to the analyzed problem. Note that this chapter is a survey that only scratches the surface of what has been proposed to resolve the frame problem. The literature on the topic is vast, and suggestions have been made over the years by many, each with its superiorities and drawbacks. By considering Dennett (1990) and Morgenstern (1996), We try to balance between the necessity for practical implementation of any proposed solution, and the theoretical underpinnings which, after all, are of prime importance when one is looking towards resolving this issue. The resolution of the frame problem is very important to enable common-sense reasoning within AI. This chapter, through the analysis of practical and theoretical approaches, will point out the fact that the frame problem is multifaceted and that a fusion of these perspectives is needed in order to develop more robust and capable AI systems.

Morgenstern's Theoretical Perspective: Before one can engage in an objective evaluation of the solutions proposed with respect to the Frame Problem, of course, one has to have specific criteria that solutions are to meet before claims can be made to the degree of solving the problem. In her landmark 1996 paper, Morgenstern (1996) conducted a wide-ranging study of the Frame Problem, proposing seven criteria for the evaluation of proposed solutions. These criteria would provide a framework within which a comparative evaluation can be made about the ability of various approaches to best address the multidimensional nature of the problem at hand. Through a rigorous analysis and test of the various aspects of the problem, we have chosen to set down five out of Morgenstern's criteria as fundamental for a firm evaluation: solving the right problem, being truthful, applying to concrete examples, conciseness, and being theoretically founded. Such criteria provide a clear structure for comparison and contrast between the effectiveness of various solutions. The criteria will be a benchmark to check the feasibility and workability of the proposed solution. This will ensure that the solutions are not only theoretically set in place to fix the problem but also provide practical, implementable strategies (Kamermans & Schmits, 2004). In this ever-changing field of artificial intelligence, new algorithms and models are coming up every other day; thus, clear and well-defined criteria are very important in the distinction between solutions that do and do not develop our understanding.

One of the most important of these criteria that Morgenstern (1996) strove for is that a solution would resolve the Frame Problem, rather than some other problem, a simplified version of it, or a redefined version of it. Over the last 25 years it has been more common for researchers to alter the Frame Problem in some way, and then solve these modified versions. Instead of tackling the original problem, many solutions have focused on variations or subsets, which can obscure the fundamental challenges the original Frame Problem presents. Many proposed solutions have focused on narrower aspects of this problem or have introduced new assumptions that make the problem easier to handle (Morgenstern, 1996). While these approaches can provide partial insights, they do not fully address the broader and more complex issue defined by McCarthy and Hayes (1981).

Morgenstern's (1996) other criterion, however, is firm: solutions must be truthful. This means that no assumptions or approximations should be left unexamined within the theory, as these can invariably lead to an incorrect understanding or application of the solution. Additionally, she emphasizes that a solution must be applicable to real-world instances of the problem. Most of the proposed solutions are tested on such overly simplified 'toy problems' that are hardly representative of the full complexity of real-world examples. To be capable of being called a proper resolution of the Frame Problem, the solution should serve for these toy problems as well as for concrete, real situations. What this means is that such a solution has to be practically applicable, not merely theoretically sound.

Another criterion proposed by Morgenstern (1996) is conciseness, which means that a solution should not be much overcomplicated or verbose. For instance, giving an open-ended list of patch rules to cover every possibility is just as meaningless as giving an open-ended list of frame axioms. A concise solution would be simple and to the point, effective and without any extra complexity, getting directly to the bottom line of the problem.

Finally, Morgenstern (1996) underlines the need to trace a theoretical solution. Therefore, even if a solution depends on a procedural method, rather than pure logic, the method should still be derived by sound theoretical principles. Ensuring a solid theoretical foundation means that the solution is not only practical but also logically sound and reliable. This theoretical grounding provides confidence that the solution is correct and can be consistently applied across different situations.

Dennett's Practical Perspective: AI systems, when initially created, start from the point of having no pre-existing knowledge about the world. This is akin to the philosophical idea of a "tabula rasa," or blank slate, where nothing is known, and all knowledge must be acquired. For AI, this means that every piece of information it needs to function must be provided by the programmers at the beginning (Dennett, 1990). If an AI were to begin as a tabula rasa, it would face enormous challenges in acquiring the vast amount of basic knowledge needed to perform even simple tasks. Learning everything from scratch would be highly time-consuming and computationally intensive, making it an unrealistic expectation for practical AI development. For this reason, developers pre-program AI with essential knowledge. This approach is akin to creating an "adult" AI—an AI that, like an adult human, has pre-existing knowledge accumulated through life experiences (Dennett, 1990). Pre-programming fundamental knowledge within AI systems has several advantages. By embedding basic knowledge in AI, developers guarantee that the system can become operational without having to learn everything through trial and error. This approach is efficient because it saves time and computational resources that would have gone into making the AI operational. Moreover, pre-installed knowledge improves the learning process of AI with a strong base to allowing more subsequent learning. For instance, consider a scenario in which an AI to help with gardening is designed. The AI will need to be pre-trained with knowledge about garden tools like trowel, watering can, and pruning shears; understand actions like planting seeds, watering the plant, and trimming bushes; and understand such basic principles as plants need sunlight and water in order to grow. This will enable the AI to assist directly in gardens without having to learn all this information from scratch.

As Daniel Dennett described in 1990, the most basic problem of this approach is the 'installation problem.' This refers to the problem of determining all the basic and relevant information upon which the AI agent needs to act efficiently within a dynamic world. Basically, the 'installation problem' in AI involves fundamental components for embedding the necessary information an AI agent needs to function effectively in a changing world. First of all, the AI system need to determine which knowledge is relevant, such as identifying broad topics, specific facts, and the relationship of the different pieces of information. Secondly, it involves embedding the information through an efficient choice of formats, organizing it effectively, and implementing the right tools for storage and retrieval. Thirdly, data structures are optimized for quick retrieval, and context awareness is embedded in the AI to ensure the information is usable and accessible. Lastly, dynamic adaptation is required. AI is equipped to have mechanisms to update its knowledge base with learning and the capacity to deal with incomplete or uncertain information. Overcoming the installation problem is a big step toward solving the frame problem in AI (Dennett, 1990). Providing an effective manner of determining and embedding all the necessary information needed for the AI to operate under a dynamic environment assures the foundations required in taking on the managing and adaptation of its knowledge base. In simple terms, making sure the process is fundamental in sensitizing the AI with tools and structures that it should use in an autonomous and effective way within real life. For this reason, we need to develop AI systems with highly efficient generative knowledge systems (Dennett, 1990).

Dennett (1990) suggests the characteristics AI systems with efficient, generative knowledge systems should show. The first one in the list would be a need for effective information storage. The demand for an efficient system of information storage arises due to both space and time limitations. While the human brain is not large enough to store an infinite amount of information, the more pressing issue is the accessibility of this information in real time. For any stored information to be useful, it must be quickly and reliably accessible within the short time frames available to agents operating in the world. A being that can solve any problem given enough time, such as a million years, would not be considered intelligent because real-world scenarios require quick decision-making and rapid thinking. This need for rapid thought and action is not just a theoretical condition of intelligence but an empirical fact observed in how we manage to operate efficiently in our daily lives. The need for efficient information storage in intelligent systems is driven by the necessity for quick and reliable access to information. AI researchers must develop systems that can use selected knowledge elements to plan and predict outcomes, mimicking the human ability to envision, act, and evaluate.

Secondly, Dennett (1990) proposes that understanding both conscious and unconscious problem-solving systems is essential. Designing an AI system capable of planning by utilizing its knowledge about the world involves mimicking the human cognitive process of introspection, where one envisages a situation (e.g., organizing a complex event), imagines an action (e.g., scheduling a series of activities), predicts outcomes (e.g., potential conflicts or overbookings), and evaluates those outcomes (e.g., determining if rescheduling is necessary). Conscious problem-solving, where there is explicit planning and deliberate thought, is synonymous with the tackling of novel tasks, like a new product design or devising a marketing strategy. By way of contrast, trivial problems, for example, deciding what to wear, are solved almost without thought, reflecting unconscious information processing. AI workers tend to use models of conscious thought as a basis for system design. Unconscious problem-solving cannot be represented in such models, so it is a major failing for the emulation of human thought in AI.

It is important that, finally, AI systems be designed to manipulate and utilize a large background body of knowledge in their development over time in a dynamically interactive manner (Dennett, 1990). The capacity to cope with and remain responsive to environmental change will be critical in the solution of the frame problem and in realizing intelligence of the type that characterizes human behavior. An important indication of intelligence is real-time control over an action based on feedback. An intelligent agent is not one that follows pre-set instructions, but one that is constantly monitoring and adjusting to changes in its environment. Such real-time adaptability indicates a high level of cognitive flexibility and problem-solving ability.

In AI, the frame problem is approached by creating systems capable of demonstrating adaptive behaviors similar to life so as to ensure they can respond effectively to complications and unpredictability in situations that are likely to be encountered in real-life problems.

CHAPTER 3

EVALUATING LARGE LANGUAGE MODELS: ADDRESSING CHALLENGES IN COMMON SENSE REASONING

As discussed in earlier chapters, common sense reasoning remains a significant challenge in the development of artificial intelligence (AI) systems. First challenge is representing common sense knowledge, which is broad, informal, and context-dependent, is difficult to formalize in a way that AI can effectively process. The second major challenge is integrating and applying background knowledge. While humans effortlessly draw on a vast range of information and experiences to make sense of their environment, AI systems often struggle to do so in a flexible and contextually appropriate manner. Lastly, the frame problem highlights AI's difficulty in determining which information is relevant in dynamic environments, especially as conditions shift. Large Language Models (LLMs) have emerged as a promising tool to address these challenges. Trained on vast amounts of text data, these models have demonstrated the ability to generalize from past knowledge, infer from incomplete information, and make contextually appropriate decisions. This study explores the potential of LLMs to exhibit common sense reasoning by examining how they integrate background knowledge, navigate real-world scenarios, and handle uncertainty in dynamic situations. The objective of this chapter is to evaluate the ability of two different Large Language Model (LLM)-based AI systems (ChatGPT 4.o and Claude Sonnet 3.5) to overcome the challenges. This evaluation will be conducted using six primary benchmarks: context-based information integration, future planning and adaptation ability, comprehensive causality and linked information management, operational execution competence, background knowledge integration and application ability, and accuracy and relevance management, along with 24 detailed sub-benchmarks.

Additionally, these benchmarks will provide a comprehensive framework to evaluate the models' ability to perform common sense reasoning.

3.1 Benchmarks

The evaluation benchmarks were established following a systematic approach grounded in theoretical insights and guided by feedback from academic advisor Associate Professor Aziz Zambak. The process began with an extensive literature review focused on the common sense knowledge and reasoning and AI-related studies, which provided a solid foundation for understanding the challenges AI models face in common sense reasoning and identifying key areas for evaluation.

While constructing the benchmarks, the criteria presented in Ernest Davis's (2023) study, *Benchmarks for Automated Commonsense Reasoning: A Survey*, were carefully considered. Davis's (2023) suggested criteria were chosen because they systematically address the challenges of evaluating commonsense reasoning, thereby enhancing the reliability and validity of the results. His criteria encompass key principles such as broad domain coverage, modality diversity, stability of benchmarks, replicable methodology, avoidance of artifacts, and quality of individual questions. They provide a strong foundation for assessing the performance of AI models in the various contexts they may encounter in daily life. Davis's proposed criteria are as follows:

Scope of Domain: Commonsense reasoning is not limited to language or specific knowledge areas; it spans diverse fields like physics, social relations, and psychology. Davis's criterion encourages benchmarks to evaluate AI's ability to handle various real-life contexts rather than merely testing narrow knowledge sets. This is crucial for AI models to develop human-like commonsense reasoning.

Range of Modalities and Tasks: Commonsense reasoning encompasses not only text-based knowledge but also visual and physical interactions. Davis's emphasis on modality diversity is essential for testing AI models' commonsense abilities across linguistic, visual, and other contextual domains. This enables AI models to exhibit greater flexibility in real-world, multifaceted tasks.

Stability for Comparability: Stability of a benchmark allows consistent evaluation of AI systems over time. This criterion ensures benchmarks remain unchanged unless updates are clearly separated as different versions, making it possible to accurately measure performance differences between new AI models and previous versions.

Replicable Methodology: Benchmarks should be constructed with well-defined and replicable methods, enabling different researchers to obtain similar results using the same approaches. This increases the reliability of scientific research and standardizes AI development by making results generalizable.

Avoiding Artifacts: Preventing models from achieving success on benchmarks by merely learning coincidental patterns is essential for evaluating whether they genuinely possess commonsense reasoning abilities. This criterion aims to ensure that AI interprets the deeper structure of the question rather than relying on superficial connections.

Quality of Individual Items: Each question should be accurate, clear, and capable of measuring commonsense knowledge, thereby enhancing the reliability of benchmark results. In fields like commonsense reasoning, this criterion is crucial for ensuring the validity of the results, as clear and evaluable questions allow for accurate comparisons of difficulty levels between models (Davis, 2023).

These criteria provided a robust theoretical foundation for constructing the benchmarks. By using Davis's suggested criteria as a guide, the aim was to comprehensively evaluate AI models' competencies in commonsense reasoning.

Next, Core concepts were identified, prioritizing the difficulties highlighted by Daniel Dennett (1990) and Laura Morgenstern (1996), particularly the integration of contextual information and the handling of temporal reasoning. These challenges informed the overall framework for the benchmarks. The benchmarks were structured by translating key theoretical concepts from the literature into clear, testable categories. Each benchmark was designed to evaluate a distinct capability of the language models, with sub-benchmarks allowing for more detailed analysis in areas like context management and causal reasoning. Dennett's (1990) discussions on AI's struggles with decision-making and cause-effect relationships significantly influenced the Context-Based Information Integration and Comprehensive Causality and Linked Information benchmarks, which aimed to assess how well models manage contextual information and causal relationships. Meanwhile, Morgenstern's (1996) work on temporal reasoning and planning shaped the Future Planning and Adaptation Ability benchmark, focusing on AI's adaptability to uncertainty and future scenarios, with sub-benchmarks like Handling Uncertainty and Prediction of Possible Scenarios. Finally, we were sought to refine the structure and content of the benchmarks, especially shaping the Background Knowledge Integration and Application benchmark, which tests how well models apply naive knowledge from physics, mathematics, and psychology. Through this comprehensive process, the evaluation benchmarks were thoroughly developed to assess the models' capacity to overcome the problem of common sense reasoning. The resulting main benchmarks and sub-benchmarks are as follows:

Benchmark-1: Context-Based Information Integration

This benchmark was selected to test how well AI models can process and integrate contextual information. Common sense reasoning is closely tied to a model's ability to discern which information is relevant in a given context. Therefore, the model's ability to maintain consistency, select appropriate information, and transition between contexts is critical. To test this main benchmark, the following sub-benchmarks have been established:

Sub-benchmark 1.1 Contextual Consistency: The model's ability to use information consistently within the same context.

Sub-benchmark 1.2 Comprehensive Document Management: The ability to effectively integrate information from multiple documents.

Sub-benchmark 1.3 Contextual Appropriateness: The ability to choose and apply the most appropriate information for the context.

Sub-benchmark 1.4 Transition Between Contexts: The ability to maintain information integrity while transitioning between different contexts.

Sub-benchmark 1.5 Management of Contextual Conflicts: The ability to manage conflicting information and draw the correct conclusion.

Benchmark-2: Future Planning and Adaptation Ability

It is crucial to assess AI models' abilities to plan for the future and adapt to uncertainties in dynamic environments. Common sense reasoning is not just about handling current situations but also about foreseeing and managing future possibilities. Thus, future planning and the ability to cope with uncertainty are important areas of evaluation. To test this main benchmark, the following sub-benchmarks have been established

Sub-benchmark 2.1 Prediction of Possible Scenarios: The ability to anticipate future possibilities.

Sub-benchmark 2.2 Strategic Planning: The capacity to plan strategically considering future events.

Sub-benchmark 2.3 Adaptability and Flexibility: The ability to adjust and respond to unexpected situations.

Sub-benchmark 2.4 Handling Uncertainty: The ability to manage uncertainty and incomplete information.

Benchmark-3: Comprehensive Causality and Linked Information

The problem of common sense reasoning highlights one of the core challenges for AI: understanding cause-effect relationships. The ability to analyze complex causal relationships, track linked information, and manage uncertainties in causality is vital. This benchmark evaluates how well models understand and follow causal chains. To test this main benchmark, the following sub-benchmarks have been established:

Sub-benchmark 3.1 Identification of Cause-Effect Relationships: The ability to identify cause-effect relationships between events.

Sub-benchmark 3.2 Tracking Causal Chains: The ability to follow complex causal chains.

Sub-benchmark 3.3 Conditional Causality: The capacity to evaluate conditional causality based on possible scenarios.

Sub-benchmark 3.4 Interactive Causality: The ability to understand interactive cause-effect relationships between events.

Benchmark-4: Operational Execution Competence

Common sense reasoning also encompasses a model's ability to perform tasks toward a defined goal. Evaluating how well models manage simultaneous actions, handle temporal gaps, and maintain focus on the goal is essential. To test this main benchmark, the following sub-benchmark has been established:

Sub-benchmark 4.1 Ability to Manage Simultaneous Actions: The capacity to successfully manage multiple actions simultaneously.

Benchmark-5: Background Knowledge Integration and Application

AI models must demonstrate the ability to integrate and apply general background knowledge effectively. This benchmark evaluates how models use naive knowledge in physics, mathematics, and psychology to solve problems. To test this main benchmark, the following sub-benchmarks have been established:

Sub-benchmark 5.1 Naive Physics: The ability to understand and apply basic physical principles. The sub-benchmark *Naive Physics* assesses the models' ability to understand basic physical principles through four aspects:

5.1.1 Gravity and Motion: This aspect tests whether the models can predict and explain the effects of gravity on objects in motion.

5.1.2 Commonsense Law of Inertia: It evaluates if the models understand that an object will stay in motion or rest unless acted upon by an external force.

5.1.3 Object Permanence: This examines whether the models recognize that objects continue to exist even when out of sight.

5.1.4 Locality of Change: It assesses if the models can reason that changes typically occur locally and do not affect unrelated areas automatically.

Sub-benchmark 5.2 Naive Mathematics: The capacity to apply simple mathematical rules and operations accurately. The *Naive Mathematics* sub-benchmark assesses the models' ability to apply basic mathematical reasoning through three key aspects:

5.2.1 Estimation of Quantities: This tests whether the models can make reasonable approximations of quantities without exact calculation.

5.2.2 Spatial Relationships and Size Estimation: It evaluates if the models can judge relative sizes and understand spatial relationships.

5.2.3 Time Estimation: This examines the models' ability to make reasonable predictions about the duration of events or actions.

Sub-benchmark 5.3 Naive Psychology: The ability to understand and apply fundamental psychological concepts.

5.3.1 Emotion Recognition: Recognition of basic emotions (such as happiness, sadness, anger).
5.3.2 Theory of Mind: The ability to infer the thoughts and intentions of others.
5.3.3 Social Norm Adherence: Following social rules and appropriate behaviors.

Benchmark-6: Accuracy and Relevance Management

It is essential to assess how well AI models can filter out irrelevant information, detect incorrect data, and prioritize relevant content. These skills are critical for managing the common sense reasoning effectively, where not all information is equally important or correct. To test this main benchmark, the following sub-benchmarks have been established:

Sub-benchmark 6.1 Filtering Out Irrelevant Information: The ability to identify and discard irrelevant data.

Sub-benchmark 6.2 Detection of Incorrect Information: The capacity to detect and avoid incorrect information.

Sub-benchmark 6.3 Prioritization of Information: The ability to prioritize relevant information for decision-making.

These benchmarks provide a comprehensive framework to evaluate the AI models' capacity to solve the problem of common sense reasoning from multiple perspectives. All these benchmarks have been compiled into a table, which can be found in the title Table-I: Benchmarks.

3.2. Methods Of Data Collection and Evaluation

To compare the performance of two large language models (LLMs), we conducted a series of experiments. These experiments were designed to assess the models' natural language processing capabilities, commonsense reasoning skills, and overall performance across various benchmarks. The research design was structured around carefully crafted examples, each aimed at testing specific sub-benchmarks. These examples were meticulously developed to analyze the commonsense reasoning performance of ChatGPT 4.o and Claude Sonnet 3.5, providing a comprehensive evaluation under consistent conditions across multiple dimensions.

When creating the benchmark scenarios, we adhered to the criteria outlined in Ernest Davis's *Benchmarks for Automated Commonsense Reasoning: A Survey* (2023). These criteria were crucial in ensuring the reliability and validity of each scenario:

Accuracy: Questions were structured to guarantee that responses were definitively correct, enabling models to be rewarded for right answers and penalized for wrong ones, thereby maintaining benchmark integrity.
Commonsensical: Questions were built around everyday knowledge rather than specialized or encyclopedic knowledge, aiming to measure true commonsense reasoning.
Task Relevance: Scenarios were designed to mirror real-life tasks to ensure practical relevance. For example, under the *3.4 Interactive Causality* benchmark, scenarios aimed to assess the model's capacity to understand interdependent cause-effect relationships within real-world contexts.

Table-II: Benchmarks

Main Category	Sub-Benchmark
1. Context-Based Information Integration	1.1 Contextual Consistency
	1.2 Transition Between Contexts
	1.3 Contextual Appropriateness
	1.4 Comprehensive Document Management
	1.5 Management of Contextual Conflicts
2. Future Planning and Adaptation Ability	2.1 Prediction of Possible Scenarios
	2.2 Strategic Planning
	2.3 Adaptability and Flexibility
	2.4 Handling Uncertainty
3. Comprehensive Causality and Linked Information	3.1 Identification of Cause-Effect Relationships
	3.2 Tracking Causal Chains
	3.3 Conditional Causality
	3.4 Interactive Causality
4. Ability to Manage Simultaneous Actions	4.1 Ability to Manage Simultaneous Actions
5. Background Knowledge Integration and Application	**5.1 Naive Physics**
	5.1.1 Gravity and Motion
	5.1.2 Commonsense Law of Inertia
	5.1.3 Object Permanence
	5.1.4 Locality of Change
	5.2 Naive Mathematics
	5.2.1 Estimation of Quantities
	5.2.2 Spatial Relationships and Size Estimation
	5.2.3 Time Estimation
	5.3 Naive Psychology
	5.3.1 Emotion Recognition
	5.3.2 Theory of Mind
	5.3.3 Social Norm Adherence
6. Accuracy and Relevance Management	6.1 Filtering Out Irrelevant Information
	6.2 Detection of Incorrect Information
	6.3 Prioritization of Information

Richness and Complexity of Inferences: Some questions included complex real-life scenarios. For instance, a scenario involving a restaurant chain adding healthy food options to its menu was designed to examine not only economic impacts but also customer satisfaction and employee workload, allowing assessment of the model's capacity to analyze multi-dimensional interactions.

Easy for Humans: Questions were crafted to be naturally solvable by humans, avoiding puzzle-like complexity and ensuring that they were straightforward and accessible.

Natural Language: The language used in questions was clear and fluent, avoiding awkward or artificial phrasing, which ensured that the models could focus on the content without distractions from unnatural language.

Free from Social Biases: Scenarios were carefully designed to be free from stereotypes or biases, ensuring fair and impartial outcomes from AI models.

Cultural and Linguistic Independence: Questions were crafted to be applicable across diverse cultural and linguistic contexts, supporting broader applicability and reducing cultural specificity.

Ease of Evaluation: Each question was structured with clear, correct answers, allowing for straightforward, automated assessment. This ensured that results could be reliably interpreted and compared.

Requirement of Commonsense Reasoning: Questions were designed to require commonsense reasoning, moving beyond mere linguistic patterns to test genuine understanding and inference based on everyday knowledge.

Each scenario, built upon these criteria, was specifically designed to evaluate the distinct abilities each sub-benchmark aimed to measure. All tested examples can be found in Appendix-1.

A rubric was created to evaluate the responses obtained from the AI models (ChatGPT 4.o and Claude Sonnet 3.5). The rubric was designed to assess the models' abilities in processing information, reasoning, and providing contextually appropriate and accurate responses. It was divided into six main benchmarks and their sub-benchmarks, each tailored to specific features of the criteria. The models' responses were rated on a scale from 0 to 5 for each sub-criterion, considering the unique characteristics of each criterion. This rubric allowed for a quantitative evaluation of the models' strengths and weaknesses across various performance metrics, providing a clear indication of how well they performed in common sense reasoning. For example,

0 points: This score was given when the model does not provide an answer to the question or scenario.

1 point: The lowest score was given if the model attempted to respond but provided incorrect information.

2 points: This score was given if the model attempted to respond but provided incomplete information.

3 points: The model received a mid-level score if it gave a generally correct answer but showed minor mistakes or omissions.

4 points: This score was given if the model provided a correct answer but with less detail.

5 points: The highest score of 5 was given when the model provided a complete, detailed, and accurate answer.

After counting the scores for each sub-benchmark, the total scores were compiled under the main benchmarks as outlined in the rubric. This provided an overview of each model's overall performance in each main benchmark. It was then used to determine which benchmarks the models excelled at and where they showed deficiencies.

3.3 Results and Discussion

The objective of this section is to present and analyze the experimental results of the two Large Language Model (LLM)-based AI systems, ChatGPT 4.o and Claude Sonnet 3.5, in their ability to overcome the challenge of common sense reasoning. The evaluation was conducted using six primary benchmarks—context-based information integration, future planning and adaptation, comprehensive causality and linked information management, operational execution competence, background knowledge integration and application, and accuracy and relevance management—along with 24 detailed sub-benchmarks. For each benchmark, custom scenarios were developed, and specific questions were posed to the models to assess their capacity to apply common sense reasoning. Each response was thoroughly analyzed and evaluated using a detailed rubric. This rubric provided scores for each benchmark and sub-benchmark, allowing for a systematic comparison of the models' performance. This section will offer a general assessment of the results, highlighting the key strengths and weaknesses of each model. For a more in-depth review, including the original scenarios, questions, and detailed responses along with the rubric scores, please refer to the appendix section.

3.3.1 Benchmark-1: The Context-Based Information Integration Results and Discussion

The test results for Benchmark-1, along with its associated sub-benchmarks, are analyzed and discussed below.

Sub-benchmark 1.1: Contextual Consistency and Sub-benchmark 1.2: Transition Between Contexts

This experiment aimed to evaluate the "Contextual Consistency" (1.1) and "Transition Between Contexts" (1.2) sub-benchmarks by testing how well ChatGPT 4.o and Claude Sonnet 3.5 could adapt to different scenarios involving Anna's experience on a snowy mountain hike. These two benchmarks evaluate the model's ability to maintain coherence within a single context and adapt smoothly when transitioning between different contexts, ensuring the information remains consistent and relevant throughout. Both models were tested based on four questions.

Scenario 1: Anna is walking on a snowy mountain, searching for a cabin.
Question 1: "What is Anna doing right now, and how does she feel?"

Both models inferred that Anna likely feels cold, tired, and anxious due to the harsh conditions, maintaining strong contextual consistency. Claude Sonnet 3.5 provided a more detailed breakdown, but both responses aligned well with the scenario.

Scenario 2: Anna reaches a cabin and lights a fire.
Question 2: "How does Anna feel right now?"

Both models updated their responses based on the new context, inferring that Anna likely feels warmer, relieved, and safer. Claude added more detailed emotions like comfort and gratitude, demonstrating strong consistency and effective transition between contexts.

Scenario 3: Anna talks to a friend about her earlier mountain walk.
Question 3: "What did Anna feel while she was walking on the mountain?"

Both models inferred her past feelings of cold, anxiety, and tiredness. Claude Sonnet 3.5 offered additional insights such as determination and awe, while both maintained consistency with the original context and transitioned smoothly to reflect on Anna's past feelings.

Scenario 4: Anna is in the cabin, talking to her friend.
Question 4: "Where is Anna right now, and how does she feel?"

Both models adapted to the final context of Anna being in the cabin, feeling warm and safe, while Claude added reflective and nostalgic elements as Anna conversed with her friend, showing nuanced handling of context shifts.

Both ChatGPT 4.o and Claude Sonnet 3.5 perform well in terms of "Contextual Consistency" (1.1) and "Transition Between Contexts" (1.2) across the different scenarios involving Anna's mountain hike. Both models maintain alignment with the context at each stage, accurately reflecting Anna's physical and emotional states based on the changing conditions—from walking in the cold, to finding shelter, to reflecting on her past experience. ChatGPT 4.o provides concise, focused responses that remain consistent with the scenarios, while Claude Sonnet 3.5 offers more detailed insights, sometimes adding emotional depth and nuance, such as reflecting on feelings of awe and nostalgia. Both models effectively transition between different contexts, showing a clear understanding of how Anna's situation evolves, with Claude Sonnet 3.5 occasionally providing a more nuanced interpretation of shifts in the emotional and situational context. Overall, both models demonstrate strong contextual consistency and adaptability, with ChatGPT 4.o excelling in efficiency and clarity, and Claude Sonnet 3.5 offering more detailed, reflective responses.

The detailed scenario, the questions posed, the original responses from both models, and my evaluation of these responses can be found in the appendix section.

Sub-benchmark 1.3: Contextual Appropriateness

This sub-benchmark tests the model's ability to select and apply the most appropriate information for the given context. To evaluate this, a fire and firefighting intervention scenario was created. Both models were assessed based on three questions designed to measure their contextual understanding in a fire rescue situation:

1. What should the firefighter do now?
2. What might the firefighter be thinking while walking through the building?
3. What if the firefighter wanted to take a break at this moment, what would they do?

In the first question, both models adhered to standard safety protocols, but Claude Sonnet 3.5 provided more detailed procedures, including the use of specific tools like thermal cameras. For the second question, ChatGPT focused on practical concerns, while Claude added emotional depth and detailed considerations, both remaining contextually appropriate. In the third question, ChatGPT suggested a practical yet brief pause, while Claude emphasized strict adherence to protocols, highlighting the risks of taking a break in such situations.

The evaluations for the "Contextual Appropriateness" sub-benchmark revealed that both ChatGPT 4.o and Claude Sonnet 3.5 demonstrated a strong understanding of the scenarios' context. Both models provided contextually consistent responses, focusing on critical aspects such as personal safety, search strategies, and communication. However, Claude Sonnet 3.5's responses were slightly more detailed, incorporating specific tools like thermal cameras and emphasizing the importance of marking cleared areas, giving it a slight edge in professional firefighting procedures. Both models effectively met the criterion by addressing the urgency and intensity of the situation without introducing irrelevant or illogical elements. While ChatGPT 4.o's approach was practical and concise, Claude Sonnet 3.5 offered more depth and specificity, particularly in adhering to strict protocols. Overall, both models successfully fulfilled the sub-benchmark requirements.

The detailed scenario, the questions posed, the original responses from both models, and my evaluation of these responses can be found in the appendix section.

Sub-benchmark 1.4: Comprehensive Document Management

The "Comprehensive Document Management" criterion assesses how well a model integrates and synthesizes information from a complex document. To evaluate this, both models were given the article titled *"Global Study of 71,000 Animal Species Finds 48% are Declining"* by Sharon Guynup. Both ChatGPT 4.o and Claude Sonnet 3.5 were tested using Sharon Guynup's article "Global Study of 71,000 Animal Species Finds 48% are Declining" by posing three questions.

1. How did the Industrial Revolution affect extinction rates, and what did this change parallel?
2. What are the shortcomings of the global overview provided by the new study?
3. What are the long-term survival threats facing species, and how can these be prevented?

In the first question, both models adhered to standard safety protocols, but ChatGPT 4.o provided a more cohesive and integrated response, connecting extinction rates to human population growth and climate change, while Claude Sonnet 3.5 focused more on quoting the text directly. In the second question, both models effectively summarized the study's shortcomings, with ChatGPT offering a clearer organization of points, and Claude providing a more direct breakdown. In the third question, ChatGPT provided a structured response, separating threats and prevention strategies, while Claude offered a narrative-driven approach, emphasizing early intervention and expert insights.

Therefore, the evaluation of ChatGPT 4.o and Claude Sonnet 3.5 in terms of Comprehensive Document Management highlights distinct strengths and approaches. ChatGPT 4.o excels in integrating and synthesizing information, providing structured and cohesive responses that cover various topics, such as the Industrial Revolution's impact, extinction rates, and related strategies. Its organization of key points and prevention strategies is clear and comprehensive. In contrast, Claude Sonnet 3.5, while accurate and informative, relies more on quoting and summarizing the original text. Its narrative-driven responses emphasize actionable steps and expert insights but lack the depth of synthesis seen in ChatGPT 4.o's answers. Overall, both models meet the criterion effectively, but ChatGPT 4.o stands out for its structured, detailed approach, while Claude Sonnet 3.5 offers fluid but lack of depth.

The detailed scenario, the questions posed, the original responses from both models, and my evaluation of these responses can be found in the appendix section.

Sub-benchmark 1.5: Management of Contextual Conflicts

Management of contextual conflicts benchmark evaluates the model's ability to identify and resolve contradictions between different pieces of information. This experiment aimed to evaluate how well ChatGPT 4.o and Claude Sonnet 3.5 could identify contradictions in a traffic accident scenario. The scenario involved a midday collision at a busy city intersection between a red Toyota and a white Volkswagen, witnessed by two individuals from different vantage points at a nearby café. The witnesses provided differing accounts of the event, specifically regarding the status of the traffic lights, the actions of the drivers, and their reactions following the accident.

Both models were tasked with analyzing these statements and detecting key discrepancies, such as whether the traffic lights were yellow or green for each car and which driver had the right of way. Additionally, they were asked to evaluate the contrasting descriptions of post-accident behavior, where one witness reported that the red Toyota driver was angry, while the other witness described the white Volkswagen driver as the one who was yelling.

Both ChatGPT 4.o and Claude Sonnet 3.5 effectively identify contradictions in the witness statements, but their approaches differ in depth. ChatGPT 4.o’s response is concise, categorizing contradictions clearly by focusing on the traffic light status, driver actions, and post-accident reactions. It provides a straightforward analysis without overcomplicating the situation, making it useful for quickly testing Management of Contextual Conflicts. Claude Sonnet 3.5’s response, on the other hand, is more detailed, addressing additional elements such as witness locations, fault assignment, and potential reasons for the discrepancies, which adds depth but makes the response more complex. Both models meet the criterion, but ChatGPT 4.o is more efficient, while Claude Sonnet 3.5 provides a more thorough exploration of the conflict.

3.3.2 Benchmark-1: The Context-Based Information Integration Conclusion

The Context-Based Information Integration benchmark, with its sub-benchmarks, evaluates how well AI models can maintain consistency, manage transitions, select relevant information, and resolve conflicts. Both ChatGPT 4.o and Claude Sonnet 3.5 performed well in the scenarios involving Anna's Mountain hike, especially in **Sub-benchmark 1.1 Contextual Consistency** and **Sub-benchmark 1.2 Transition Between Contexts**. Both models accurately reflected Anna's physical and emotional states based on the evolving conditions. ChatGPT 4.o provided concise and focused responses, excelling in efficiency and clarity, while Claude Sonnet 3.5 offered more detailed responses, sometimes incorporating emotional depth, such as feelings of awe and nostalgia. Regarding **Sub-benchmark 1.3 Contextual Appropriateness**, both models demonstrated a strong understanding of the context, focusing on critical aspects like personal safety and communication, though Claude Sonnet 3.5 included more professional details, giving it a slight edge. In the **Sub-benchmark 1.4 Comprehensive Document Management** evaluation, ChatGPT 4.o excelled in synthesizing and organizing information, while Claude Sonnet 3.5 often relied on quoting the original text and repeating information, which made its responses less original. ChatGPT 4.o delivered more structured and cohesive answers. For **Sub-benchmark 1.5 Management of Contextual Conflicts**, both models effectively identified contradictions in witness statements, with ChatGPT 4.o offering a shorter, clearer analysis, and Claude Sonnet 3.5 providing a deeper, more thorough exploration. Overall, ChatGPT 4.o stands out for its efficiency and clarity, while Claude Sonnet 3.5 delivers more detailed and nuanced responses. At this point, both models meet the primary benchmark requirements, but with distinct differences in approach. Both models are highly successful in Context-Based Information Integration. The rubric scores for both models can be seen in Table III.

Table III: Benchmark-1: Context-Based Information Integration Rubric

Sub-Benchmarks	**0**	**1**	**2**	**3**	**4**	**5**	**Score (ChatGPT 4.o)**	**Score (Claude Sonnet 3.5)**
1.1 Contextual Consistency	Model fails completely to understand or maintain context.	Model frequently creates contradictions and loses connection with prior context.	Some inconsistencies exist, the model struggles to maintain context.	Maintains context most of the time, but small inconsistencies arise occasionally.	Almost always consistent, rarely loses context.	Completely consistent, always maintains the correct context.	**5**	**5**
1.2 Transition Between Contexts	Model fails completely to transition between contexts.	Significant confusion during context transitions.	Some confusion during context transitions.	Manages context transitions successfully most of the time, with rare confusion.	Rarely any issues during context transitions.	Transitions between contexts are seamless and coherent.	**5**	**5**
1.3 Contextual Appropriateness	Model fails completely to provide contextually appropriate responses.	Responses are generally irrelevant and inappropriate to the context.	Occasionally gives responses inappropriate for the context.	Responses are generally appropriate for the context, with rare inappropriateness.	Responses are mostly contextually appropriate, with only occasional mistakes.	Responses are always fully appropriate to the context.	**5**	**5**
1.4 Comprehensive Document Management	Model fails completely to track document content.	Model misses important information	Model misses some details, information is lost.	Generally, tracks details but may overlook minor aspects.	Tracks and manages almost all information	Manages all important details thoroughly	**5**	**4**

Table III: Benchmark-1: Context-Based Information Integration Rubric (continued)

1.5 Management of Contextual Conflicts	Model fails completely to resolve or address contextual conflicts.	Model fails to resolve conflicts and gets stuck between contradictory information.	Sometimes resolves conflicts, but generally fails.	Able to resolve minor conflicts but struggles with complex ones.	Resolves most conflicts logically.	Successfully and efficiently manages all conflicts.	**5**	**5**
Total Score							**25/25**	**25/24**
Total Score %							**%100**	**%96**

3.3.3 Benchmark-2: Future Planning and Adaptation Ability Results and Discussion

The test results for Benchmark-2, along with its associated sub-benchmarks, are analyzed and discussed below.

Sub-benchmark 2.1: Prediction of Possible Scenarios

This experiment aimed to evaluate the sub-benchmark "Prediction of Possible Scenarios" by testing how well ChatGPT 4.o and Claude Sonnet 3.5 could foresee potential outcomes in a forest fire scenario in Ayvacık, Çanakkale. Both models were asked to predict the progression of the fire and identify key influencing factors.

Scenario: A forest fire in the Ayvacık district of Çanakkale, Turkey, is spreading rapidly due to strong winds.

Question: What are the possible scenarios for the fire's progression, and what factors could influence these scenarios?

ChatGPT 4.o provided five possible scenarios, such as fire containment, rapid spread, and natural extinguishment, focusing on factors like wind, firefighting efforts, humidity, and terrain. Claude Sonnet 3.5 offered a similar range of scenarios, adding considerations such as the impact of fire reaching agricultural areas and the importance of time of day on fire behavior.

Both models effectively identified key factors influencing fire progression, such as wind speed, firefighting resources, and weather conditions. While they both successfully met the "Prediction of Possible Scenarios" criterion by presenting comprehensive general scenarios, their analyses could have been enhanced by including more localized insights. Overall, both models performed well, but their responses remained broad and lacked attention to specific regional characteristics.

Sub-benchmark 2.2: Strategic Planning

This experiment evaluated the sub-benchmark "Strategic Planning" by assessing how well ChatGPT 4.o and Claude Sonnet 3.5 could develop both short-term and long-term solutions to prevent a forest fire from spreading to nearby villages. The scenario involved changing wind conditions that increased the risk of the fire reaching residential areas, and the models were tasked with providing comprehensive strategic plans to mitigate this threat.

Scenario: Efforts are being made to control the fire in Ayvacık, but the wind is quickly changing direction, increasing the risk of the fire spreading toward the nearby villages.

Question: What strategic plan can be developed to prevent the fire from reaching the villages? The plan should include both short-term and long-term solutions.

ChatGPT 4.o proposed several short-term actions, such as immediate evacuation, firebreak creation, aerial firefighting, and wind monitoring, alongside long-term solutions like reforestation, fire-resistant infrastructure, and community fire prevention programs. Claude Sonnet 3.5 offered a similar structure but included more operational details, such as predictive fire behavior models, mobile weather stations, and regional cooperation. Both models presented well-rounded strategies combining immediate responses with sustainable, long-term fire management approaches. In summary, both models proposed comprehensive and adaptable strategies for mitigating the fire threat, with Claude Sonnet 3.5 offering more detailed operational approaches and ChatGPT 4.o providing clear, structured solutions.

Sub-benchmark 2.3: Adaptability and Flexibility

This experiment assessed the sub-benchmark "Adaptability and Flexibility" by evaluating how well ChatGPT 4.o and Claude Sonnet 3.5 could adjust to an unexpected eastward spread of the fire due to a sudden wind shift. The models were asked to provide strategies for firefighting teams to quickly adapt and reorganize their positions in response to the new conditions.

Scenario: As the fire in Ayvacık begins to come under control, an unexpected wind shift occurs, causing the fire to spread rapidly eastward into another forested area.

Question: How can firefighting teams quickly adapt to the sudden eastward spread of the fire, and how should they reorganize their positions to respond to the new conditions?

ChatGPT 4.o suggested several immediate actions, such as repositioning ground and aerial resources to the eastward fire front, expanding firebreaks, and closely monitoring wind patterns. It also recommended reorganizing firefighting teams by rotating crews and establishing a new command center closer to the affected area. Claude Sonnet 3.5 provided a similar approach but with additional emphasis on real-time reassessment using aerial reconnaissance and predictive fire models, adjusting communication protocols, and coordinating with neighboring fire districts for additional resources. Claude's response included more detailed suggestions on interagency coordination and continuous monitoring. Overall, while both responses demonstrated adaptability, Claude Sonnet 3.5 provided more detailed and organized strategies for dealing with the new fire conditions.

Sub-benchmark 2.4: Handling Uncertainty

This experiment assessed the sub-benchmark "Handling Uncertainty" by testing how ChatGPT 4.o and Claude Sonnet 3.5 would manage unpredictable conditions during the fire.

Scenario: Weather conditions in Ayvacık are uncertain, and the wind direction keeps shifting. Additionally, gathering accurate information about the fire is difficult because some areas are covered in smoke.

Question: What strategies should be implemented to manage risks and respond to the uncertainty in this firefighting situation?

ChatGPT 4.o emphasized the use of drones, satellites, and weather stations for enhanced monitoring, flexible team deployment, and dynamic safety zones. The model also focused on pre-positioning resources, modular firefighting units, and communication strategies to adapt to changing conditions, along with setting up evacuation plans and expanding buffer zones around villages. Claude Sonnet 3.5 provided a similar approach but included more advanced technologies such as LIDAR, AI-powered decision support, and predictive modeling to track fire behavior. Claude also emphasized collaborative intelligence, public communication, and psychological support for firefighters. It further incorporated risk-based decision-making and scenario planning to ensure adaptability.

Both models effectively addressed the need for flexible strategies, resource allocation, and real-time data collection to manage risks. Claude Sonnet 3.5 offered a more detailed and comprehensive framework for handling uncertainty, incorporating additional factors like community coordination and mental health support for firefighters. Overall, both models met the criterion, with Claude Sonnet 3.5 providing a more robust and well-rounded response.

3.3.4 Benchmark-2: Future Planning and Adaptation Ability Conclusion

In this benchmark, both ChatGPT 4.o and Claude Sonnet 3.5 performed well across the four sub-benchmarks, demonstrating strong capabilities in future planning and adaptability. In **Sub-benchmark 2.1 Prediction of Possible Scenarios**, both models effectively identified key factors affecting fire progression, but Claude Sonnet 3.5 provided more comprehensive insights. However, neither model incorporated specific geographical or infrastructural details about Ayvacık, which could have enhanced the practical application of their solutions. For **Sub-benchmark 2.2 Strategic Planning**, both models offered detailed short- and long-term solutions, with Claude excelling in operational details like predictive fire models. In **Sub-benchmark 2.3 Adaptability and Flexibility**, both models proposed solid strategies for adjusting to unexpected changes, but Claude Sonnet 3.5 stood out with its emphasis on real-time reassessment and coordination. Finally, in **Sub-benchmark 2.4 Handling Uncertainty**, both models addressed risk management effectively, but Claude Sonnet 3.5 provided a more robust framework, incorporating advanced technologies and mental health support for firefighting teams.

Overall, while ChatGPT 4.o presented clear and structured responses, Claude Sonnet 3.5 consistently offered more detailed, operationally focused strategies Since Claude Sonnet 3.5's more detailed answers were not effective in providing the basic benchmarks of the **Future Planning and Adaptation Ability benchmark**, both models were evaluated equally in terms of points. The rubric scores for both models can be seen in Table IV.

3.3.5 Benchmark-3: Comprehensive Causality and Linked Information Results and Discussion

The test results for Benchmark-3, along with its associated sub-benchmarks, are analyzed and discussed below.

Sub-benchmark 3.1: Identification of Cause-Effect Relationships

This experiment evaluated the sub-benchmark "Identification of Cause-Effect Relationships" by testing how well ChatGPT 4.o and Claude Sonnet 3.5 could identify and explain the relationship between two events: a broken coffee machine and a decrease in customer numbers at a café.

Scenario: The café owner noticed that the coffee machine was broken in the morning, and on the same day, the number of customers was lower than expected. **Question:** "What is the relationship between the coffee machine breaking down and the decrease in customer numbers? If there is one, how would you explain it?"

ChatGPT 4.o's response suggested a clear cause-effect relationship, stating that the broken coffee machine likely resulted in fewer customers because coffee is a key product for the café. Without it, the café became less attractive to regular customers or passersby, leading to a decline in foot traffic. Claude Sonnet 3.5's response provided a more detailed analysis, considering multiple potential factors such as limited menu offerings, longer wait times, and word of mouth spreading the news of the broken machine. Claude also mentioned alternative explanations, including weather conditions or competitor promotions, and emphasized the need for more data, making it much more successful in exploring possible scenarios compared to ChatGPT 4.o's narrower focus. Both models effectively addressed the criterion. ChatGPT 4.o gave a more straightforward, Claude's response demonstrated a broader understanding of possible causes and their effects, making it more sophisticated in comparison.

Sub-benchmark 3.2: Tracking Causal Chains

This experiment evaluated the sub-benchmark "Tracking Causal Chains" by testing how well ChatGPT 4.o and Claude Sonnet 3.5 could follow a series of events linked by a causal chain in the context of a friend's last-minute cancellation of a vacation due to a work emergency.

Table IV: Future Planning and Adaptation Ability Rubric

Sub-benchmark	0	1	2	3	4	5	Score (Chat GPT 4.o)	Score (Claude Sonnet 3.5)
2.1 Prediction of Possible Scenarios	Model fails completely to predict any scenarios.	Model frequently makes incorrect predictions.	Model occasionally makes incorrect predictions.	Model usually makes correct predictions, with occasional errors.	Model makes mostly correct predictions, with rare mistakes.	Model consistently predicts scenarios correctly and accurately.	**4**	**4**
2.2 Strategic Planning	Model fails completely to create any strategic plans.	Model frequently creates flawed or incomplete plans.	Model occasionally creates strategic plans with some errors.	Model usually creates correct and logical plans, with some minor errors.	Model creates mostly correct and coherent strategic plans.	Model consistently creates flawless, strategic, and logical plans.	**5**	**5**
2.3 Adaptability and Flexibility	Model fails completely to adapt to new information or changes.	Model frequently struggles to adapt to new situations.	Model occasionally adapts to changes, but not consistently.	Model generally adapts well to new situations, with some minor issues.	Model adapts to most changes and new situations effectively.	Model consistently adapts seamlessly to all changes and new situations.	**5**	**5**
2.4 Handling Uncertainty	Model fails completely to handle uncertain information.	Model frequently struggles with uncertainty, making incorrect assumptions.	Model occasionally handles uncertainty but makes mistakes.	Model generally handles uncertainty well, with occasional misjudgments.	Model handles most uncertain situations effectively, with rare errors.	Model consistently handles uncertainty	**5**	**5**
Total Score							**20/19**	**20/19**
Total Score %							**%95**	**%95**

Scenario: A group of friends planned a summer vacation to a coastal town, but one friend cancels last-minute due to an emergency at work. **Question:** How does this friend's cancellation impact the vacation plans, and what events are linked together?

ChatGPT 4.o provided a concise explanation, noting that the friend's cancellation could impact logistics (such as transportation or accommodations) and emotional responses (like disappointment or concern). It identified a straightforward cause-and-effect sequence, starting with the work emergency leading to the cancellation and resulting in changes to the vacation plans. Claude Sonnet 3.5 delivered a more detailed analysis, categorizing the effects into logistical adjustments (e.g., changing accommodations or recalculating shared costs), activity modifications, and emotional impact. It outlined a more structured chain reaction, from the friend's work emergency to the necessary changes in the group's plans and provided more specific examples of how these changes might play out.

Both models effectively recognized the causal chain and the ripple effect of the cancellation on the vacation. However, Claude Sonnet 3.5 offered a more structured and comprehensive breakdown, providing greater depth in understanding the connections between each event. While ChatGPT 4.o gave a clear and concise response, Claude's response was more detailed and robust in tracking the full causal chain.

Sub-benchmark 3.3: Conditional Causality

This experiment evaluated the sub-benchmark "Conditional Causality" by testing how well ChatGPT 4.o and Claude Sonnet 3.5 could analyse potential outcomes based on the condition of ticket sales for a theatre group's upcoming play.

Scenario: A theatre group is planning to stage a new play, but the decision on the venue depends on ticket sales.

Question: What will happen if ticket sales are sufficient, and what will happen if they are not?

ChatGPT 4.o provided a clear explanation of the two scenarios: if ticket sales are sufficient, the play will likely proceed as planned with additional resources possibly allocated for production. If sales are insufficient, the group might downscale, change the venue, or cancel the production. Claude Sonnet 3.5 offered a more detailed response, discussing a wider range of potential outcomes. If ticket sales are sufficient, it suggested securing an appropriate venue and possibly expanding the play's run. If sales are insufficient, it proposed alternative actions like downsizing, increasing marketing efforts, or postponing the production.

Both models fulfilled the criterion by addressing the possible outcomes based on ticket sales. ChatGPT 4.o provided a concise and general overview, while Claude Sonnet 3.5 offered a more comprehensive and nuanced analysis, including specific actions the theatre group might take. Claude's more detailed approach demonstrated a deeper understanding of complex conditional scenarios.

Sub-benchmark 3.4: Interactive Causality

This experiment assessed the sub-benchmark "Interactive Causality" by evaluating how well ChatGPT 4.o and Claude Sonnet 3.5 could explain the interactive cause-effect relationships between adding healthy menu options and their impact on sales, customer satisfaction, and employee working conditions in a restaurant chain.

Scenario: A restaurant chain added healthy food options to its menu, which attracted customer interest.

Question: How would you explain the impact of adding a healthy menu on sales, customer satisfaction, and employee working conditions?

ChatGPT 4.o explained the potential positive impacts of adding healthy menu options, such as increased sales and customer satisfaction from catering to health-conscious customers. It also noted the potential for increased employee workload due to the need for training but suggested that improved customer satisfaction could lead to a more positive work environment over time. However, the response was more linear, focusing on each factor individually rather than their interconnections.

Claude Sonnet 3.5 provided a more detailed breakdown, highlighting how the introduction of healthy options could boost sales, increase loyalty from health-conscious customers, and lead to more complex food preparation and busier shifts for employees. It emphasized the interdependence of these factors, demonstrating a better understanding of the interactive relationships between increased sales, improved customer satisfaction, and the effects on employee working conditions.

Both models addressed the interactive cause-effect relationships, but Claude Sonnet 3.5 provided a more comprehensive and interconnected analysis, aligning better with the 3.4 Interactive Causality criterion. ChatGPT 4.o gave a clear but more linear explanation, while Claude's response effectively demonstrated the interaction between these factors.

3.3.6 Benchmark-3: Comprehensive Causality and Linked Information Conclusion

In the results, both ChatGPT 4.o and Claude Sonnet 3.5 performed well across the four sub-benchmarks, demonstrating their capacity to understand and explain cause-effect relationships. However, Claude Sonnet 3.5 consistently provided more detailed and nuanced analyses, particularly in sub-benchmarks 3.1 (Identification of Cause-Effect Relationships) and 3.4 (Interactive Causality), where it outperformed ChatGPT 4.o by a clear margin. In **Sub-benchmark 3.1 Identification of Cause-Effect Relationships**, Claude Sonnet 3.5 provided a broader analysis, considering multiple possible factors and alternative explanations, while ChatGPT 4.o focused on a more straightforward causal link. Claude’s ability to explore the full scope of possibilities made its performance stronger in this case. In **Sub-benchmark 3.2 Tracking Causal Chains**, both models tracked the causal sequence well, but Claude Sonnet 3.5 again offered more structure and depth, exploring the ripple effects of the friend's cancellation on the vacation plans in greater detail. ChatGPT 4.o gave a concise but less comprehensive explanation. For **Sub-benchmark 3.3 Conditional Causality**, both models provided effective analyses of the possible outcomes based on ticket sales for a theater group's play. However, Claude Sonnet 3.5 offered a more robust range of potential actions the group could take, addressing more detailed scenarios than ChatGPT 4.o, whose response was more general. In **Sub-benchmark 3.4 Interactive**

Causality, Claude Sonnet 3.5 excelled by focusing on the interdependent effects of the healthy menu addition on sales, customer satisfaction, and employee working conditions. ChatGPT 4.o's response, while clear, took a more linear approach and did not address the interactions between the factors as thoroughly.

In both **Sub-benchmarks 3.1 and 3.4**, Claude Sonnet 3.5 outperformed ChatGPT 4.o by 1 point, demonstrating greater depth, complexity, and interaction between causal relationships, making its responses stronger overall. The rubric scores for both models can be seen in Table V.

3.3.7 Benchmark-4: Operational Execution Competence Results and Discussion

The test results for Benchmark-4, along with its associated sub-benchmarks, are analyzed and discussed below.

Sub-benchmark 4.1: Ability to Manage Simultaneous Actions

This experiment evaluated the sub-benchmark *Ability to Manage Simultaneous Actions* by testing how well ChatGPT 4.o and Claude Sonnet 3.5 could provide strategies for handling multiple requests in a busy café environment.

Scenario: A café employee is managing three customers simultaneously—one ordering coffee, another buying a dessert, and a third trying to pay. The employee must prioritize and address each request efficiently.

Question: How can the employee manage these simultaneous requests? How should they prioritize and handle each customer without causing delays?

ChatGPT 4.o suggested prioritizing tasks based on urgency and simplicity, beginning with the quickest action (processing payment) and then moving to the coffee order, which may take additional time, and finally the dessert. It emphasized clear communication and organization to maintain customer satisfaction and efficiency. Claude Sonnet 3.5 provided a more layered approach, recommending multitasking where possible, acknowledging each customer to build rapport, and keeping a calm demeanor.

Table-V: Benchmark-3: Comprehensive Causality and Linked Information Evaluation

Sub-benchmark	0	1	2	3	4	5	Score (ChatGPT)	Score (Claude Sonnet)
3.1 Identification of Cause-Effect Relationships	Model fails completely to identify any cause-effect relationships or possible causes.	Model frequently misidentifies cause-effect relationships and possible causes.	Model occasionally identifies cause-effect relationships but makes frequent mistakes.	Model usually identifies cause-effect relationships accurately, with occasional errors.	Model mostly identifies cause-effect relationships with rare mistakes.	Model correctly identifies cause-effect relationships and mostly possible causes.	**4**	**5**
3.2 Tracking Causal Chains	Model fails completely to track causal chains.	Model frequently loses track of causal chains.	Model occasionally tracks causal chains but struggles with longer ones.	Model usually tracks causal chains accurately, with occasional gaps.	Model tracks most causal chains effectively, with rare issues.	Model consistently and accurately tracks all causal chains.	**5**	**5**
3.3 Conditional Causality	Model fails completely to manage conditional causality.	Model frequently struggles with conditional cause-effect relationships.	Model occasionally handles conditional causality, with frequent errors.	Model usually handles conditional causality well, with occasional mistakes.	Model mostly manages conditional cause-effect relationships correctly.	Model consistently and effectively manages causality scenarios.	**5**	**5**
3.4 Interactive Causality	Model fails completely to understand interactive causality.	Model frequently fails to apply interactive causality.	Model handles interactive causality but struggles with interactions.	Model usually handles interactive causality correctly, with some minor errors.	Model mostly manages interactive causality with rare issues.	Model handles all aspects of causality accurately and efficiently.	**4**	**5**
Total Score							**20/18**	**20/20**
Total Score							**%90**	**%100**

It also suggested specific strategies like streamlining frequently requested items, using a queue system, and requesting assistance if other staff members were available. Both models met the sub-benchmark by offering effective strategies to manage simultaneous actions. ChatGPT focused on task prioritization and clear communication, while Claude added depth by incorporating customer psychology and multitasking techniques.

3.3.8 Benchmark-4: Operational Execution Competence Conclusion

Both models met the benchmark of *Operational Execution Competence*, but Claude Sonnet 3.5's additional emphasis on customer psychology and multitasking allowed for a more holistic approach. This suggests that while both models demonstrate competence in executing multiple tasks towards a common goal, Claude's response aligns more closely with the complexities of real-world operational demands, showcasing a more refined ability to manage and adapt to challenging, goal-oriented tasks under dynamic conditions. The rubric scores for both models can be seen in Table VI.

3.3.9 Benchmark-5: Background Knowledge Integration and Application Results and Discussion

The test results for Benchmark-5, along with its associated sub-benchmarks, are analyzed and discussed below.

Sub-benchmark 5.1: Naive physics

The experimental results of the sub-benchmarks under the naive physics sub-benchmark are discussed in this section.

***i.* Gravity and Motion**

This experiment aimed to evaluate the sub-benchmark 5.1 Gravity and Motion by assessing how well ChatGPT 4.o and Claude Sonnet 3.5 could understand the physical principles involved when an object, such as a book, is positioned on the edge of a table and affected by an external force.

Table-VI: Benchmark-4: Operational Execution Competence Evaluation

Sub-benchmark	0	1	2	3	4	5	Score (ChatGPT 4.o)	Score (Claude Sonnet 3.5)
4.1 Ability to Manage Simultaneous Actions	Model fails to identify priorities or provide any strategy for managing simultaneous actions.	Model identifies some tasks but provides no clear prioritization, leading to potential delays or confusion in handling requests.	Model offers a basic plan for handling requests, with some prioritization, but may not manage parallel tasks effectively or could cause delays.	Model suggests a reasonable approach for managing simultaneous actions, with prioritization and partial parallel task handling, though minor inefficiencies may exist.	Model provides a well-organized strategy, effectively managing parallel tasks, prioritizing based on urgency, and minimizing delays, but some small inefficiencies may be present.	Model delivers a highly organized, optimal strategy for managing simultaneous actions, prioritizing tasks efficiently, managing parallel tasks smoothly, and ensuring all requests are addressed without delays.	**5**	**5**
Total Score							**5/5**	**5/5**
Total Score %							**%100**	**%100**

Scenario: A person places a book on the edge of a table, with part of the book hanging off the edge.

Question: If someone accidentally bumps the table, what will likely happen to the book?

ChatGPT 4.o responded that the book would likely fall due to the lack of support under the part hanging off the edge, making it vulnerable to losing balance when the table is bumped. Claude Sonnet 3.5 provided a more detailed explanation, addressing multiple physical principles. It explained that the book's position on the edge makes it less stable due to the center of gravity, and its partial contact reduces friction

Additionally, it noted that momentum from the table bump would transfer to the book, and that gravity would accelerate its fall once it starts moving.

Both models correctly predicted the likely outcome, but Claude Sonnet 3.5 provided a more thorough response by detailing the factors involved, such as the center of gravity, friction, momentum transfer, and gravity's role. This response demonstrated a deeper understanding of the physical concepts, making it more aligned with the 5.1.1 Gravity and Motion criterion than ChatGPT 4.o's more concise explanation.

***ii.* Commonsense Law of Inertia**

This experiment assessed the sub-benchmark 5.1.2 Commonsense Law of Inertia by evaluating how well ChatGPT 4.o and Claude Sonnet 3.5 could apply basic principles of inertia to a scenario involving a stationary soccer ball that begins to move.

Scenario: A soccer ball is sitting still on a flat field with no one nearby, but it suddenly starts moving.

Question: What will happen to the ball? Why does it behave this way?

ChatGPT 4.o explained that, once in motion, the ball would continue moving in the same direction unless acted upon by another force, such as friction or air resistance, citing Newton's first law of motion. It suggested that unseen forces like wind or a slope might have caused the movement, providing a straightforward application of the principle of inertia. Claude Sonnet 3.5 provided a more detailed response, outlining the ball's likely behavior (such as gradual slowing due to friction) and exploring multiple potential causes of movement, including wind, uneven ground, and possible animal interaction. It also mentioned rotational motion, the role of conservation of energy, and Newton's first law, presenting a broader and richer understanding of physical principles at play.

Both models correctly applied the concept of inertia, but Claude Sonnet 3.5's response was more comprehensive. It considered various causes and additional factors like rotational motion, making it more aligned with the Naive Physics criterion. ChatGPT 4.o's answer was accurate yet more concise, whereas Claude Sonnet 3.5's response offered a deeper contextual understanding of the scenario.

iii. **Object Permanence**

This experiment assessed the sub-benchmark 5.1.3 Object Permanence by testing how ChatGPT 4.o and Claude Sonnet 3.5 would apply the concept of object permanence to a scenario involving a toy placed behind a curtain, rendering it out of sight.

Scenario: A child places a toy behind a curtain where it is no longer visible.

Question: Does the toy still exist even though it is no longer visible? Why or why not?

ChatGPT 4.o affirmed that the toy still exists, explaining this through the concept of object permanence—the understanding that objects continue to exist even when they cannot be observed. It identified object permanence as a foundational idea in naive physics, typically developed in infancy. Claude Sonnet 3.5 provided a more detailed response, discussing various principles such as physical reality, conservation of matter, spatial continuity, and cause and effect, to explain why the toy would still exist. It also tied object permanence to developmental psychology, noting that this concept generally develops around eight months of age in infants.

Both models successfully demonstrated an understanding of object permanence, but Claude Sonnet 3.5's response was more nuanced.

iv. Locality of Change

This experiment evaluated the sub-benchmark 5.1.4 Locality of Change by testing how ChatGPT 4.o and Claude Sonnet 3.5 would respond to a scenario in which a glass of water on a table is potentially affected by someone opening a door across the room.

Scenario: A glass of water is placed on a table, and someone across the room opens a door.

Question: Will opening the door cause the water in the glass to spill? Explain your reasoning.

ChatGPT 4.o concluded that opening the door would not cause the water to spill, as the force from opening a door across the room is too minor to disturb the glass. It mentioned that a direct force or strong gust would be necessary for the water to spill, which is unlikely in normal indoor conditions. Claude Sonnet 3.5 also stated that opening the door would not cause the water to spill. However, it provided a more detailed explanation, covering factors like energy dissipation over distance, the minimal force transmission of opening a door, water's surface tension, and the glass's stability. Claude also noted exceptions, such as an unusually strong draft or a precariously placed glass, though it deemed these unlikely in everyday situations.

Both models accurately applied the concept of locality of change, concluding that opening the door wouldn't cause the water to spill. However, Claude Sonnet 3.5's response was more thorough, covering specific physical principles and potential exceptions. ChatGPT 4.o's response was correct but less detailed, focusing mainly on the lack of direct force.

Sub-benchmark 5.2: Naive mathematics

The experimental results of the sub-benchmarks under the naive mathematics sub-benchmark are discussed in this section.

i. Estimation of Quantities

This experiment assessed 5.2.1 Estimation of Quantities by examining ChatGPT 4.o and Claude Sonnet 3.5's capacity to calculate pasta requirements for a dinner party with 5 guests and a host, given a 1-kilogram bag of pasta.

Scenario: A person is hosting a small dinner party for 5 guests with 1 kilogram of pasta on hand.

Question: Does the person have enough pasta for the dinner party?

ChatGPT 4.o calculated the pasta needed at 100 grams per person for the 5 guests only, totaling 500 grams, and concluded that 1 kilogram of pasta would be more than sufficient. However, it overlooked the host entirely, a critical omission that affected the accuracy of the response. Claude Sonnet 3.5 offered a more precise approach by accounting for both the guests and the host, totaling 6 people. It used a more accurate serving estimate of 56 grams per person, calculating a total of 336 grams needed. With almost three times the required amount in the 1-kilogram bag, Claude confirmed there was more than enough pasta.

The host's omission in ChatGPT 4.o's response was a crucial oversight, while Claude Sonnet 3.5's consideration of both the guests and host, along with a precise serving size, resulted in a far more reliable answer. This attention to detail demonstrates Claude Sonnet 3.5's superior accuracy in quantity estimation, aligning more closely with the naive mathematics criterion.

***ii.* Spatial Relationships and Size Estimation**

This experiment evaluated the 5.2.2 Spatial Relationships and Size Estimation sub-benchmark by testing how well ChatGPT 4.o and Claude Sonnet 3.5 could assess whether a large couch would fit through a doorway.

Scenario: A person is attempting to fit a large couch through a doorway.

Question: Will the couch fit through the doorway? How can the person make this decision?

ChatGPT suggested comparing the couch's dimensions with the doorway and recommended rotating or angling the couch to help it pass through if needed. While it touched on the primary strategy of adjusting the couch's position, the response was general and did not explore further details. Claude Sonnet provided a more comprehensive approach, recommending measurement comparisons for width, height, and depth, along with maneuvering space and potential obstacles like door frames or hinges. It also suggested evaluating the couch's flexibility (e.g., removable legs), using visual aids like cardboard cutouts for planning, and considering professional assistance if necessary. Claude's detailed guidance considered multiple conditional factors, demonstrating a more practical, step-by-step strategy.

While both models acknowledged the need to compare dimensions and consider adjustments, Claude Sonnet 3.5 offered a more thorough and nuanced response by integrating additional spatial considerations and practical tips.

***iii.* Time Estimation**

In this experiment on Time Estimation, ChatGPT 4.o and Claude Sonnet 3.5 were tested on their ability to estimate the total time needed for a grocery trip to fit within a one-hour limit.

Scenario: A person needs to go to a nearby grocery store and return home within one hour.

Question: Can the person complete the grocery shopping in time? How should they estimate the total time required?

ChatGPT 4.o divided the task into travel, shopping, and checkout time. It presented a clear and straightforward calculation, estimating that with travel at 20 minutes, shopping at 25 minutes, and checkout at 5 minutes, the person would have a total of 50 minutes, leaving a 10-minute buffer. However, it did not account for possible delays or extra variables. Claude Sonnet 3.5 took a more detailed approach by accounting for travel, shopping, and additional factors like potential parking time, checkout variability, and even store layout changes. Claude also suggested practical tips, such as preparing a shopping list and timing the visit to avoid peak hours, enhancing the response's realism in addressing real-world variables that could affect the timeframe.

While both models accurately calculated the necessary time components, Claude Sonnet 3.5 provided a more comprehensive assessment by considering potential delays and offering efficiency tips. This added depth makes Claude's response better aligned with the complexity required for accurate time estimation, showcasing a more practical approach to managing conditional factors in time management scenarios.

Sub-Benchmark 5.3: Naive Psychology

The experimental results of the sub-benchmarks under the naive psychology sub-benchmark are discussed in this section.

***i.* Emotion Recognition**

Emotion recognition, in the context of AI, is not conducted directly but rather through indirect cues such as behaviors and verbal hints. AI models can interpret such behavioral and contextual information to recognize emotions. Therefore, in our designed scenario, we aim to measure whether the AI can accurately infer emotional states based on behavioral cues. This experiment assessed 5.3.1 Emotion Recognition, evaluating ChatGPT 4.o and Claude Sonnet 3.5's ability to recognize emotions based on indirect behavioral cues in a social context.

Scenario: Anna is in a meeting, smiling and nodding as her colleague presents an idea. The models are tasked with inferring her possible emotional state from these non-verbal cues.

Question: What might Anna be feeling, and why?

ChatGPT 4.o inferred that Anna might feel engaged or supportive, suggesting her body language reflects either interest in or agreement with the presentation. It also recognized that her actions could represent polite social behavior, regardless of her true feelings. Claude Sonnet 3.5 provided a detailed range of possible emotions for Anna, including interest, encouragement, agreement, professional courtesy, excitement, relief, anticipation, and even masked disagreement or skepticism. It acknowledged the potential for her body language to serve multiple social functions and noted how her expressions might be affected by workplace dynamics and norms.

Both ChatGPT 4.o and Claude Sonnet 3.5 correctly identified that Anna's smiling and nodding could indicate engagement, agreement, or polite behavior. However, Claude Sonnet 3.5's response was significantly more nuanced. By exploring a wider variety of emotions—including emotions like relief, anticipation, and even masked skepticism—Claude Sonnet 3.5 exhibited a deeper understanding of how non-verbal cues can signal complex emotions, even beyond the apparent surface level. This comprehensive approach demonstrates a more sophisticated application of naive psychology, effectively capturing the subtleties of emotional interpretation in social contexts.

***ii.* Theory of Mind**

This experiment evaluates ChatGPT 4.o and Claude Sonnet 3.5 according to 5.3.2 Theory of Mind, which assesses the ability to infer others' thoughts and intentions. Both models successfully interpret the woman's primary intention, correctly inferring from social cues that she is likely approaching to ask for a light.

Scenario: "Ibrahim is sitting on a bench in the park, smoking a cigarette. In the distance, a woman holding an unlit cigarette starts walking toward him. As the woman approaches Ibrahim, she briefly glances at his lit cigarette."

Questions: What could be the woman's possible intention for approaching Ibrahim?

ChatGPT 4.o approaches the woman's reasons for approaching with multiple possibilities, suggesting that while asking for a light is the most likely reason, she could also be intending to start a conversation or request other assistance. This demonstrates ChatGPT's ability to consider various potential intentions in social contexts. However, some of these alternative options are less directly relevant to the scenario, reflecting a broader and perhaps overly generalized perspective.

Claude Sonnet 3.5, on the other hand, provides a more detailed analysis, placing "asking for a light" as the primary intention and supporting this inference with contextual details like her unlit cigarette and her glance at Ibrahim's lit cigarette. This response demonstrates stronger Theory of Mind competence by offering a focused explanation of the likely intention. Alternative scenarios are kept brief and focus on the most probable intention, providing a more contextually aligned and targeted interpretation.

Overall, both models perform well within the Naive Psychology sub-benchmark. Claude Sonnet 3.5 delivers a deeper contextual understanding by effectively leveraging social cues, while ChatGPT 4.o provides variety with its broader consideration of multiple possibilities.

iii. Social Norm Adherence

This experiment evaluates 5.3.3 Social Norm Adherence, which measures awareness of social rules and appropriate behaviors. ChatGPT 4.o and Claude Sonnet 3.5 demonstrate their understanding of social norms by analyzing who might take responsibility for closing the door in a workplace meeting scenario.

Scenario: "In a workplace meeting room, everyone has taken their seats. The meeting cannot begin until the door is closed."

Question: In this situation, who is likely to take responsibility for closing the door?

ChatGPT 4.o provides a general response, suggesting that the responsibility might fall to someone in a position of authority or a designated role, such as the meeting organizer or team leader. Alternatively, the nearest person to the door might close it as a matter of convenience or courtesy. ChatGPT highlights the social expectation that the leader or organizer may initiate or delegate small but necessary tasks, displaying an understanding of corporate norms. This response is concise and practical, offering a broader perspective on responsibility delegation.

Claude Sonnet 3.5, on the other hand, takes a more detailed and systematic approach by listing the most likely scenarios in order. It suggests that the last person to enter is most likely to close the door, as they are positioned near it and more aware of its state. Claude also considers the meeting organizer as a probable candidate, given their responsibility and awareness of the meeting requirements. Additionally, it notes that the person sitting closest to the door may feel a subtle social pressure to act. Claude further identifies the situation as a mild example of the "bystander effect," where a lack of clear delegation may create social uncertainty about who should act, offering an insightful psychological perspective.

Overall, both models perform well in the Social Norm Adherence sub-benchmark. ChatGPT 4.o provides a straightforward, corporate-focused approach, while Claude Sonnet 3.5 delivers a more in-depth analysis, incorporating social psychological concepts such as the bystander effect for a systematic response.

3.3.10 Benchmark-5: Background Knowledge Integration and Application Conclusion

This benchmark evaluated AI models' ability to effectively apply naive knowledge in physics, mathematics, and psychology to address real-world scenarios, assessing their performance across three core sub-benchmarks: Naive Physics, Naive Mathematics, and Naive Psychology. ChatGPT 4.o and Claude Sonnet 3.5 both demonstrated solid foundational understanding across these areas.

Naive Physics: Both models accurately predicted outcomes based on physical principles like gravity, motion, inertia, and object permanence. However, Claude Sonnet 3.5 offered more comprehensive explanations, integrating concepts such as friction, center of gravity, and energy dissipation, which better reflected the complexities involved in physical interactions. This detailed approach allowed Claude to deliver a more layered understanding, especially in scenarios involving nuanced conditions.

Naive Mathematics: In scenarios requiring quantity estimation, spatial relationships, and time management, both models generally provided correct calculations, but Claude Sonnet 3.5's responses included additional factors that showcased an attention to real-world variables, such as potential delays and maneuvering space in spatial scenarios. Claude's response consistently demonstrated a stronger grasp of practical constraints, providing more reliable solutions that aligned with the benchmark's criteria.

Naive Psychology: Claude Sonnet 3.5's advantage became even more pronounced. While both models recognized basic emotional cues and social norms, Claude offered a more profound understanding by exploring varied emotional possibilities, such as masked disagreement and professional courtesy, and integrating social psychological concepts like the bystander effect in social norm scenarios. This depth of interpretation illustrated Claude's capacity to reflect nuanced social dynamics, which is essential for accurate representation of human psychological understanding. In summary, both ChatGPT 4.o and Claude Sonnet 3.5 displayed competence in integrating background knowledge, with ChatGPT excelling in concise, accurate answers. The rubric scores for both models can be seen in Table VII.

Table-VII: Benchmark-5: Background Knowledge Integration and Application Evaluation

Sub-benchmark Naive Physics	**0**	**1**	**2**	**3**	**4**	**5**	**Score (ChatGPT 4.o)**	**Score (Claude Sonnet 3.5)**
5.1.1 Gravity and Motion	Model fails completely to recognize gravity or predict the book's motion.	Model vaguely identifies gravity	Model recognizes gravity but makes errors in explaining the fall due to the bump.	Model explains the book falling due to gravity but lacks clarity or has minor omissions.	Model provides a clear and mostly accurate explanation of gravity and the book's fall due to disturbance.	Model consistently provides a detailed explanation of gravity	**4**	**5**
5.1.2 Commonsense Law of Inertia	Model fails to recognize inertia or explain why the ball moves.	Model identifies some external force but misinterprets inertia or the reason for the ball moving.	Model mentions inertia and recognizes an external force but provides an incomplete explanation.	Model explains the ball's movement due to inertia but lacks some clarity or precision in identifying the force.	Model explains inertia accurately, identifies the external force, and gives a clear reason for the ball's movement.	Model gives a detailed, accurate explanation of inertia,	**5**	**5**
5.1.3 Object Permanence	Model fails to recognize the toy's continued existence when out of sight.	Model vaguely mentions object permanence but fails to give a clear explanation.	Model mentions object permanence, but the explanation is unclear or incomplete.	Model recognizes object permanence and explains it, with minor gaps in clarity.	Model explains object permanence clearly, showing that the toy exists even though it is out of sight.	Model provides a detailed and accurate explanation of object permanence.	**5**	**5**

Table-VII: Benchmark-5: Background Knowledge Integration and Application Evaluation (continued)

5.1.4 Locality of Change	Model fails to recognize that distant actions	Model recognizes the distance but doesn’t explain locality of change	Model identifies locality of change but gives an incomplete or vague explanation.	Model explains locality of change but with minor errors or lacks clarity.	Model provides a clear explanation of locality of change and why opening the door won't spill the water.	Model gives a detailed and accurate explanation of locality of change, providing a strong reason why opening the door doesn’t affect the water.	**5**	**5**
Sub-benchmark : Naive Mathematics	**0**	**1**	**2**	**3**	**4**	**5**	**Score (Chat GPT 4.o)**	**Score (Claud e Sonnet 3.5)**
5.2.1 Estimation of Quantities	Model fails to estimate how much pasta is needed for the dinner party.	Model attempts an estimation but provides an incorrect or incomplete answer.	Model estimates the quantity needed but makes some errors in the process.	Model provides a reasonable estimate but with some inaccuracies.	Model gives a mostly accurate estimation of the quantity of pasta needed for 5 guests.	Model provides a highly accurate estimation of the quantity needed, clearly explaining the rationale behind the calculation.	**3**	**5**
5.2.2 Spatial Relationships and Size Estimation	Model fails to recognize the spatial relationship	Model vaguely addresses spatial relationships	Model identifies the spatial relationship but suggests impractical actions.	Model provides a reasonable approach with minor inefficiencies.	Model gives a mostly accurate solution.	Model provides a detailed and accurate solution	**5**	**5**

Table-VII: Benchmark-5: Background Knowledge Integration and Application Evaluation (continued)

5.2.3 Time Estimation	Model fails to estimate the time required for the grocery trip.	Model attempts to estimate time but provides an incorrect or impractical solution.	Model gives a basic time estimate but overlooks key factors, leading to errors.	Model estimates time reasonably well but with some minor issues.	Model provides a mostly accurate time estimate, explaining how to factor in travel and shopping time.	Model gives a highly accurate and detailed time estimate, accounting for all necessary factors in a clear manner.	**4**	**5**
Sub-benchmark Naive Psychology							**Score (Chat GPT)**	**Score (Claude Sonnet 3.5)**
5.3.1 Emotional Recognition	Model fails to recognize emotions or intentions based on behavior.	Model vaguely mentions emotions but fails to give a clear explanation of feelings.	Model recognizes emotions but the explanation is unclear or incomplete.	Model explains likely emotions based on behavior, with some minor gaps in clarity.	Model provides a clear explanation of emotions, linking her actions to positive feelings.	Model gives a detailed and accurate explanation of emotions, providing a strong link between her behavior and her emotional state.	**5**	**5**
5.3.2 Theory of Mind	Model completely fails to infer intentions or thoughts	Model makes a vague attempt to infer intentions but does not provide a clear explanation for the character's likely thoughts or intentions.	Model infers the character's intentions but provides an incomplete explanation	Model gives a mostly accurate explanation of the character's intentions, though there are minor gaps in clarity or logic.	Model provides a clear and mostly accurate inference of the character's intentions.	Model gives a detailed, explanation of the character's thoughts and inferred mental states.	**5**	**5**

Table-VII: Benchmark-5: Background Knowledge Integration and Application Evaluation (continued)

5.3.3 Social Norm Adherence	Model fails to recognize any relevant social norms or appropriate behaviors within the scenario.	Model fails to recognize any relevant social norms or appropriate behaviors within the scenario.	Model fails to recognize any relevant social norms or appropriate behaviors within the scenario.	Model fails to recognize any relevant social norms or appropriate behaviors within the scenario.	Model fails to recognize any relevant social norms or appropriate behaviors within the scenario.	Model fails to recognize any relevant social norms or appropriate behaviors within the scenario.	**5**	**5**
Total Score							**47/50**	**50/50**
Total Score %							**%94**	**%100**

3.3.11 Benchmark-6: Accuracy and Relevance Management Results and Discussion

The test results for Benchmark-6, along with its associated sub-benchmarks, are analyzed and discussed below.

Sub-benchmark 6.1: Filtering Out Irrelevant Information

This experiment evaluated 6.1 Filtering Out Irrelevant Information, testing ChatGPT 4.o and Claude Sonnet 3.5's abilities to create a resume for Alan Turing by identifying and prioritizing relevant data for a computer engineering position.

Scenario: Alan Turing is preparing a resume, with information on his educational background, work experience, hobbies, and personal details.

Question: Which information should be included in the resume? Create a resume.

ChatGPT 4.o produced a streamlined resume, focusing on Turing's academic qualifications, notable research contributions (like Turing machines and the halting problem), and skills in computation theory and cryptology. However, it included a hobbies section, which is less relevant for a computer engineering role. Claude Sonnet 3.5 generated a similarly organized resume, but with more emphasis on Turing's pioneering work, categorized achievements, and detailed skills in computation and machine design. Claude effectively filtered out extraneous details and omitted personal aspects, which made the resume more professional and aligned with the job role

Both models successfully highlighted Alan Turing's key qualifications, skills, and experience, though Claude Sonnet 3.5's response was more refined. By excluding hobbies and personal details, Claude provided a cleaner, more targeted resume for the computer engineering position, showcasing a slightly better grasp of relevance filtering for this criterion.

Sub-benchmark 6.2: Detection of Incorrect Information

This experiment assessed 6.2 Detection of Incorrect Information, testing ChatGPT 4.o and Claude Sonnet 3.5's ability to identify inaccurate data by comparing two reports on COVID-19—one from https://en.wikipedia.org/wiki/COVID-19_pandemic with altered incorrect details, and one from https://www.mayoclinic.org/diseases-conditions/coronavirus/symptoms-causes/syc-20479963.

Scenario: The models were presented with texts from Wikipedia and mayoclinic.org on COVID-19, where the Wikipedia text contained incorrect information. The models were asked to identify which source had accurate data and to highlight any inconsistencies.

Question: Can you identify the incorrect information between these two reports? Which one is based on accurate data?

ChatGPT 4.o accurately identified major inaccuracies in the Wikipedia excerpt, noting the incorrect incubation period (25–30 days vs. the correct 2–14 days) and clarifying that older adults, rather than infants, are at greater risk. It also flagged the exaggerated 88% prevalence for loss of taste and smell in the Wikipedia excerpt, a detail not noted by Claude. Overall, it concluded that the Mayo Clinic article is more accurate and comprehensive. Claude Sonnet 3.5 also highlighted the incorrect incubation period and the misleading reference to infants as a high-risk group. Additionally, it noted the unusual mention of a rash in the Wikipedia excerpt, which is not commonly recognized as a COVID-19 symptom, but did not address the 88% figure for loss of taste and smell. Claude's response was similarly conclusive, emphasizing the Mayo Clinic as the more reliable source.

Both models identified the Mayo Clinic article as the accurate source, effectively highlighting key inaccuracies in the Wikipedia text, such as the incorrect incubation period and high-risk group. ChatGPT 4.o provided slightly more thorough detail by addressing the overstated loss of taste and smell prevalence, while Claude Sonnet 3.5 uniquely pointed out the inappropriate inclusion of rash as a symptom.

Sub-benchmark 6.3: Prioritization of Information

This experiment assessed 6.3 Prioritization of Information, testing ChatGPT 4.o and Claude Sonnet 3.5's ability to prioritize critical information in a medical emergency by reviewing patient files with varying conditions.

Scenario: A healthcare worker evaluates information from three patients in an emergency setting. Patient 1 shows symptoms and vital signs of a potential heart attack, Patient 2 has mild head trauma, and Patient 3 has a minor hand laceration. The task is to determine which patient's condition should be prioritized and why.

Question: Which patient's condition should be prioritized and why?

ChatGPT 4.o correctly prioritizes Patient 1, Mehmet Yilmaz, for immediate medical attention due to life-threatening symptoms such as severe chest pain, shortness of breath, and critically low oxygen saturation. It notes that this combination of symptoms and vital signs indicates a potentially fatal condition, justifying prompt intervention. Patients 2 and 3, with stable vital signs and less severe conditions, are deemed safe to treat afterward. Claude Sonnet 3.5 also prioritizes Patient 1, providing a detailed breakdown of why this patient's symptoms and vital signs suggest a high risk of cardiogenic shock, classifying him as a "red" or highest priority case in triage terms. It further distinguishes between the urgency levels for each patient, classifying Patient 2 as "yellow" and Patient 3 as "green," indicating that they can wait while Patient 1 receives life-saving intervention.

Both models appropriately prioritize Patient 1 for immediate care due to the life-threatening nature of his symptoms and vital signs. ChatGPT 4.o provides an accurate, concise summary, while Claude Sonnet 3.5 offers additional triage classifications and a deeper analysis, identifying potential cardiogenic shock. Both responses demonstrate a strong ability to prioritize effectively, with Claude Sonnet 3.5 providing a more nuanced analysis of urgency levels.

3.3.12 Benchmark-6: Accuracy and Relevance Management Conclusion

In **Sub-benchmark Filtering Out Irrelevant Information**, Claude Sonnet 3.5 scored slightly higher than ChatGPT 4.o by providing a cleaner, more professionally targeted resume for Alan Turing. While both models effectively selected relevant qualifications and experience, Claude omitted less relevant sections, such as hobbies, resulting in a more focused and role-aligned resume. ChatGPT 4.o included a hobbies section, which, although not harmful, detracted slightly from the resume's professional focus, leading to its score of 4 versus Claude's 5.

In **Sub-benchmark Detection of Incorrect Information**, both models performed well in identifying inaccuracies within altered COVID-19 texts. ChatGPT 4.o provided a more detailed analysis by pointing out an exaggerated statistic for loss of taste and smell, while Claude Sonnet 3.5 flagged the inclusion of uncommon symptoms like a rash. Both models demonstrated strong error detection, though with distinct points of emphasis.

In **Sub-benchmark Prioritization of Information**, both models appropriately prioritized Patient 1, who presented life-threatening symptoms indicative of a heart attack. Claude Sonnet 3.5's response added a layer of detail by categorizing each patient's triage level (red, yellow, green), reflecting a nuanced understanding of medical urgency. ChatGPT 4.o also identified Patient 1 as the priority, with a concise summary that accurately assessed the situation.

In summary, both models displayed strong information management skills across the sub-benchmarks, effectively filtering, detecting, and prioritizing data. Claude Sonnet 3.5’s responses were slightly more refined, especially in omitting less relevant information and offering comprehensive, structured details in emergency prioritization, underscoring its slight lead in overall accuracy and relevance management. The rubric scores for both models can be seen in Table VII.

Table-VIII: Benchmark-6: Accuracy and Relevance Management Evaluation

Sub-Benchmark	0	1	2	3	4	5	Score (ChatGPT 4.o)	Score (Claude Sonnet 3.5)
6.1 Filtering Out Irrelevant Information	Model fails to distinguish relevant from irrelevant information and includes unnecessary details in the resume.	Model identifies some relevant information but includes irrelevant details such as hobbies and personal life, making the resume less effective.	Model filters out some irrelevant details but still includes minor unnecessary information.	Model filters out most irrelevant details but may include a small amount of unnecessary information.	Model effectively filters out all irrelevant information, selecting only the necessary details for the resume.	Model perfectly filters all irrelevant details, providing a focused resume	**4**	**5**
6.2 Detection of Incorrect Information	Model fails to identify any incorrect information in a given scenario.	Model identifies some incorrect information but misses key errors or fails to correct them.	Model identifies incorrect information but provides an incomplete or unclear explanation of why it's incorrect.	Model identifies most incorrect information and provides an acceptable explanation of the errors.	Model accurately identifies incorrect information and provides a clear explanation for why it's incorrect.	Model perfectly identifies all incorrect information and provides a detailed explanation	**5**	**5**
6.3 Prioritization of Information	Model fails to prioritize the most critical information in the emergency scenario.	Model struggles to prioritize and evaluates less critical patients	Model prioritizes patients but may overlook some aspects of urgency	Model prioritizes the most critical patient but with minor	Model effectively prioritizes the most critical patient.	Model flawlessly prioritizes the most critical patient	**5**	**5**
Total Score							**14**	**15**
Total Score							**%93,3**	**%100**

CHAPTER 4

CONCLUSION

In conclusion, this thesis has explored the critical importance of common sense reasoning in artificial intelligence and the fundamental challenges associated with integrating this capability into AI systems. While AI has advanced significantly in areas such as visual recognition and natural language processing, essential qualities like contextual awareness, implicit knowledge application, and flexibility in dynamic real-world situations remain lacking. To understand these limitations, the thesis has analyzed three primary challenges: the representation of common sense knowledge, the structuring of tacit knowledge, and the resolution of the frame problem. To address these challenges, the capabilities of two LLM-based models, ChatGPT 4.o and Claude Sonnet 3.5, were assessed through six primary benchmarks and twenty-four sub-benchmarks. These evaluations analyzed the models' performance in areas such as context integration, future planning, causality management, operational competence, and accuracy. The findings illustrate both the progress and the persisting limitations in AI's capacity for common sense reasoning. Ultimately, this work highlights the importance of advancing common sense reasoning to enhance AI functionality and reliability, reinforcing the ongoing need for research in this critical area.

The primary findings of this thesis reveal that both ChatGPT 4.o and Claude Sonnet 3.5 exhibit strong capabilities in replicating certain facets of common sense reasoning, with distinct approaches and performance strengths. ChatGPT 4.o excels in delivering concise, efficient responses with a high degree of clarity, effectively capturing essential reasoning processes, while Claude Sonnet 3.5 provides more detailed, contextually nuanced answers that reflect a deeper understanding, particularly in benchmarks involving complex causality, contextual management, and adaptability. In tests such as Context-Based Information Integration and Comprehensive Causality, Claude Sonnet 3.5 frequently outperformed due to its depth and attention to emotional and operational intricacies, whereas ChatGPT 4.o maintained a straightforward, structured approach that emphasized efficiency. In areas of future planning, adaptability, and background knowledge application, Claude Sonnet 3.5's responses displayed a heightened awareness of real-world constraints and social dynamics, giving it an edge in scenarios that required multidimensional insights. My tests showed that both models achieved a high success rate, underscoring their potential in simulating aspects of common sense reasoning. However, verifying whether these capabilities constitute a genuine replication of common sense reasoning requires a more extensive framework of benchmarks, rigorous trials, and in-depth expert analysis. This emphasizes the need for further refinement and comprehensive evaluation to build a clearer, more reliable understanding of each model's true abilities in real-world reasoning scenarios.

Highlighting the broader significance and impact of this research underscores both its theoretical and practical contributions. Theoretically, this work advances the understanding of common sense reasoning within AI, a longstanding challenge that lies at the heart of achieving human-like intelligence. By systematically evaluating large language models (LLMs) through carefully constructed benchmarks, this study provides a structured framework for analyzing how well these models can mimic aspects of common sense. This contributes to the philosophical discourse on artificial agency, intentionality, and knowledge representation, addressing fundamental questions about the nature of AI cognition. Practically, the findings hold implications for the development of AI systems that can better navigate complex, real-world situations. Improved common sense reasoning capabilities could enhance AI's effectiveness in fields such as autonomous driving, healthcare, and customer service by enabling systems to interpret context, anticipate outcomes, and respond more intelligently to human needs. Overall, this research lays the groundwork for more adaptable and contextually aware AI, bridging a critical gap between AI's current capacities and the nuanced, situational reasoning required for robust, human-centered applications.

In any research, limitations play a crucial role in understanding the scope and applicability of the findings. Addressing these limitations not only provides transparency but also helps outline potential areas for improvement in future studies. In this research, certain constraints emerged due to the nature of the methods, evaluation formats, and the specific AI models used. While these limitations do not diminish the overall value of the results, acknowledging them is essential to comprehensively assess the outcomes and guide further advancements in the field. In my opinion, there are four aspects of this study that could be improved.

One of the limitations is related to the evaluation of the benchmarks. According to Davis, it is important that the problems be easy to evaluate, with clear-cut criteria of correctness. This criterion underscores the need for establishing precise standards in commonsense benchmarks to ensure consistent and objective evaluation. When questions are presented in fixed response formats—such as true/false or multiple-choice—the evaluation process becomes more straightforward and is often automatable. Structured responses like these allow for objective, reliable assessment of an AI system's performance and reduce the likelihood of subjective bias in scoring. In this study, however, I chose a free-response format because it aligns well with evaluating the capabilities of large language models, allowing them to generate open-ended responses that showcase nuanced reasoning and contextual understanding. According to Davis, free-response formats in question-answering benchmarks pose a greater challenge, as they often require approximate evaluation through human labor or automated scoring tools like BLEU scores. In instances where responses are unrestricted—such as free-form text or visual content generated by LLMs—the evaluation process becomes considerably more complex, making automated scoring nearly impossible. Even human evaluation in such cases can be subjective, which can compromise accuracy and consistency. This is why fixed response formats are typically preferred in commonsense benchmarks, as they provide a more straightforward, objective basis for evaluation. While the free-response format was particularly well-suited for evaluating large language models in this research, creating an additional benchmark system based on true/false or multiple-choice formats could further enhance the evaluation process. Such structured formats could provide a complementary, standardized approach that would allow for objective comparisons across a broader range of AI systems. Developing these additional benchmarks would build on existing insights, supporting more efficient and scalable assessments and contributing to a more comprehensive evaluation framework for commonsense reasoning in AI.

The second limitation is related to another benchmark criterion highlighted in Davis's paper. It is the "Range of modalities and tasks," which underscores that commonsense reasoning interacts with diverse tasks and modalities, such as language, visual perception, and physical interactions. Commonsense is not limited to a single modality; rather, it spans multiple types of knowledge and integrates various forms of information. For this reason, effectively measuring an AI system's commonsense capabilities requires benchmarks that can evaluate its performance across different modalities and tasks. For example, one benchmark might test language comprehension, while another assesses the ability to interpret and analyze visual information. An AI system should not only accurately interpret situations in language-based tasks but also make logical inferences when engaged in visual tasks. This diversity in benchmarks is essential for assessing an AI's ability to apply commonsense reasoning across different contexts and types of information. The ability of AI to effectively integrate commonsense knowledge across multiple modalities and tasks provides a more comprehensive measure of whether it truly demonstrates human-like understanding. In this study, we limited our evaluation to language-based tasks, focusing specifically on the strengths of LLMs. However, the need for future benchmarks that incorporate a broader range of tasks and modalities is evident. Expanding the scope of benchmarks to include different types of scenarios—such as those that test visual, spatial, and even physical interaction-based reasoning—could provide a more complete assessment of an AI's commonsense reasoning capabilities. This approach would also make it possible to evaluate the AI's versatility and adaptability across varied forms of interaction, further supporting the goal of achieving a robust, human-like commonsense understanding in AI.

A third limitation is related to the sample size and the diversity of test scenarios. While this study has provided valuable insights, testing with a larger variety of scenarios could enhance the robustness and generalizability of the findings. Expanding the sample size by including more diverse examples would allow for a deeper evaluation of each model's commonsense reasoning abilities across a broader range of situations. In future research, studies conducted with the collaboration of multiple researchers could further enrich this approach by diversifying the scenarios and testing a greater number of LLM models. This expansion would not only increase the reliability of the results but also offer a more comprehensive understanding of how different models handle a variety of commonsense reasoning tasks.

A final limitation is the number of expert reviewers involved in the evaluation process. In this study, two experts—myself, Zeynep Kabadere, and my advisor, Associate Professor Dr. Aziz Zambak—provided assessments. While our evaluations offered valuable insights, the limited number of reviewers may restrict the range of perspectives on each benchmark. Involving a larger group of experts could enhance the objectivity and depth of the evaluations, allowing for a broader interpretation of model performance. In future studies, incorporating additional expert opinions could lead to a more balanced and comprehensive evaluation, thereby strengthening the reliability of the findings.

I believe that to truly create AI that resonates, we must delve deeply into the intricacies of commonsense reasoning. This study takes a step toward exploring the complexities of this domain, offering a potential roadmap for future advancements and a foundation for developing systems that may better navigate the nuanced realities of our world. Alan Turing once envisioned a day when machines might surprise us with their understanding; I believe this research brings us closer to that vision, pushing the boundaries of AI's ability to think and reason with the depth and subtlety of common sense.

BIBLIOGRAPHY

Baillargeon, R. (1994). How do infants learn about the physical world? *Current Directions in Psychological Science*, *3*(5), 133-140.

Bergmann, R., Kolodner, J., & Plaza, E. (2005). Representation in case-based reasoning. The Knowledge Engineering Review, 20(3), 209-213.

Bian, N., Han, X., Sun, L., Lin, H., Lu, Y., & He, B. (2023). ChatGPT is a Knowledgeable but Inexperienced Solver: An Investigation of Commonsense Problem in Large Language Models. ArXiv:2303.16421 [Cs]. https://arxiv.org/abs/2303.16421

Boden, M. A. (2018). Artificial intelligence: A very short introduction. Oxford University Press.

Brachman, R. J., & Levesque, H. J. (2022). *Machines like Us*. MIT Press.

Brachman, R. J. (1988). The basics of knowledge representation and reasoning. AT&T Technical Journal, 67(1), 7-24. https://doi.org/10.1002/j.1538-7305.1988.tb00230.x

Cassenti, D. N., Kaplan, L. M., & Roy, A. (2023, September). Representing uncertainty information from AI for human understanding. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 67, No. 1, pp. 177-182). Sage CA: Los Angeles, CA: SAGE Publications.

Davis, E. (2014). Representations of Commonsense Knowledge. Morgan Kaufmann.

Davis, E. (2017). Logical Formalizations of Commonsense Reasoning: A Survey. Journal of Artificial Intelligence Research, 59, 651–723. https://doi.org/10.1613/jair.5339

Davis, E. (2023). Benchmarks for automated commonsense reasoning: A survey. ACM Computing Surveys, 56(4), 1-41.Davis, E. (2024). Mathematics, word problems, common sense, and artificial intelligence. Bulletin of the American Mathematical Society.

Davis, E., & Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. Communications of the ACM, 58(9), 92–103. https://doi.org/10.1145/2701413

Davis, P. J. (2006). Mathematics & common sense: A case of creative tension. CRC Press.

Davis, R., Shrobe, H., & Szolovits, P. (1993). What is a knowledge representation? AI Magazine, 14(1), 17-33.

Dennett, D. C. (1990). Cognitive wheels: The frame problem of AI. The philosophy of artificial intelligence, 147, 1-16.

Dennett, D. C. (2017). Brainstorms: Philosophical essays on mind and psychology. MIT press.

Denning, P. J., & Arquilla, J. (2022). The context problem in artificial intelligence. Communications of the ACM, 65(12), 18-21.

Dreyfus, H. L. (1972). *What Computers Can't Do*. HarperCollins Publishers.

Dreyfus, H. L. (1992). *What Computers Still Can't Do*. MIT Press.

Elio, R. (2002). Common Sense, Reasoning, and Rationality. Oxford University Press.

Fetzer, J. H. (1990). The frame problem: artificial intelligence meets David Hume. International Journal of Expert Systems, 3(3), 219-232.

Fischer, J. (2020). Naive physics: building a mental model of how the world behaves.

Fodor, J. A. (1983). *The modularity of mind*. MIT press.

Fodor, J. A. (1989). Modules, frames, fridgeons, sleeping dogs.

Forbus, K. D. (1988). Commonsense physics: a review. Annual Review of Computer Science, 3(1), 197-232.

Freeman, W. (1992). Framing is a Dynamic Process. *Psycoloquy*, *3*, 62.

Giunchiglia, F. (1995). An epistemological science of common sense. *Artificial Intelligence*, *77*(2), 371–392. https://doi.org/10.1016/0004-3702(95)90019-5

Gordon, A. S., & Hobbs, J. R. (2017). *A formal theory of commonsense psychology: How people think people think*. Cambridge University Press.

Grosan, C., & Abraham, A. (2011). *Intelligent systems* (Vol. 17, pp. 149-185). Berlin: Springer.

Halpern, J. Y. (2017). *Reasoning about uncertainty*. MIT press.

Haugeland, J. (1987). An overview of the frame problem.

Hayes, P. J. (1977). In defence of logic. In *Proc. IJCAI-77* (pp. 559-565).

Hayes, P. J. (1979). The naive physics manifesto. *Expert systems in the microelectronic age*.

Heider, F. (2013). *The psychology of interpersonal relations*. Psychology Press.

Hello GPT-4o. (2024). Openai.com. https://openai.com/index/hello-gpt-4o/

Henderson, L. (2018). The problem of induction.

Hobbs, J. R., & Moore, R. C. (1985). *Formal theories of the commonsense world* (Vol. 1). Intellect Books.

Holyoak, K. J., & Morrison, R. G. (2013). *The Oxford handbook of thinking and reasoning*. Oxford University Press.

Hookway, C. (1986). Minds*, machines, and evolution : philosophical studies*. Cambridge University Press.

Introducing Claude 3.5 Sonnet. (2024, June 21). anthropic.com. https://www.anthropic.com/news/claude-3-5-sonnet

Janlert, L. E. (1987). Modeling change: The frame problem.

Kamermans, M., & Schmits, T. (2004). The history of the frame problem. *Artificial Intelligence*, *86*, 116.

Kirsh, D. (1991). Foundations of AI: the big issues. *Artificial intelligence*, *47*(1-3), 3-30.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). 2012 AlexNet. *Adv. Neural Inf. Process. Syst*, 1-9.

Lake, Brenden M, Ullman, Tomer D, Tenenbaum, J. B., & Gershman, S. J. (2016). *Building Machines That Learn and Think Like People*. ArXiv.org. https://arxiv.org/abs/1604.o0289

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Lenat, D. B. (1995). CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM, 38(11), 33-38.

Lenat, D. B., & Guha, R. V. (1989). *Building Large Knowledge-based Systems*. Addison Wesley Publishing Company.

Lenat, D. B., Guha, R. V., Pittman, K., Pratt, D., & Shepherd, M. (1990). Cyc: toward programs with common sense. *Communications of the ACM*, *33*(8), 30–49. https://doi.org/10.1145/79173.79176

Lesort, T., Lomonaco, V., Stoian, A., Maltoni, D., Filliat, D., & Díaz-Rodríguez, N. (2020). Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information fusion*, *58*, 52-68.

Levesque, H. J. (2018). *Common* sense*, the Turing test, and the quest for real AI*. Mit Press.

Li, S., Chen, J., & Yu, D. (2023). Teaching Pretrained Models with Commonsense Reasoning: A Preliminary KB-Based Approach. *ArXiv:1909.09743 [Cs]*. https://arxiv.org/abs/1909.09743

Liu, H., & Singh, P. (2004). Commonsense Reasoning in and Over Natural Language. *Media Laboratory* , 293–306. https://doi.org/10.1007/978-3-540-30134-9_40

Marcus, G., & Davis, E. (2019). Rebooting *AI: Building artificial intelligence we can trust*. Vintage.

Marvin, M., & Seymour, A. P. (1969). Perceptrons. *Cambridge, MA: MIT Press*, *6*(318-362), 7.

McCarthy, J. (1959). Programs with common sense. In *Symposium on the Mechanization of Thought Processes*. Teddington, UK: National Physical Laboratory.

McCarthy, J. (1981). Epistemological problems of artificial intelligence. In *Readings in artificial intelligence* (pp. 459-465). Morgan Kaufmann.

McCarthy, J. (1989). Artificial Intelligence, Logic and Formalizing Common Sense. *Philosophical Logic and Artificial Intelligence*, 161–190. https://doi.org/10.1007/978-94-009-2448-2_6

McCarthy, J., & Hayes, P. J. (1981). Some Philosophical Problems from the Standpoint of Artificial Intelligence. *Readings in Artificial Intelligence*, 431–450. https://doi.org/10.1016/b978-0-934613-03-3.50033-7

Mccarthy, J., & Vladimir Lifschitz. (1990). *Formalizing common sense: papers*. Ablex Pub. Corp.

Minsky, M. (1974). A framework for representing knowledge.

Minsky, M. (2007). *The emotion machine: commonsense thinking, artificial intelligence, and the future of the human mind*. Simon & Schuster.

Moore, R. C. (1981, June). Problems in logical form. In *19th Annual Meeting of the Association for Computational Linguistics* (pp. 117-124).

Moore, R. C. (1982). *The role of logic in knowledge representation and commonsense reasoning* (pp. 428-433). SRI International. Artificial Intelligence Center.

Morgenstern, L. (1996). The problem with solutions to the frame problem. *The Robot's Dilemma Revisited: The Frame Problem in Artificial Intelligence. Ablex Publishing Co., Norwood, New Jersey*, 99-133.

Morton, A. (2009). Folk psychology. In B. McLaughlin, A. Beckermann, & S. Walter (Eds.), *The Oxford handbook of philosophy of mind*. Oxford University Press.

Mueller, E. T. (2014). *Commonsense Reasoning*. Morgan Kaufmann.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., & Mian, A. (2023). A comprehensive overview of large language models.

Peppas, P., Koutras, C. D., & Williams, M. A. (2001). Prolegomena to concise theories of action. *Studia Logica*, *67*, 403-418.

Pollock, J. L. (1997). Reasoning about change and persistence: A solution to the frame problem. *Nous*, *31*(2), 143-169.

Reiter, R. (2001). *Knowledge in action: logical foundations for specifying and implementing dynamical systems*. MIT press.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation, parallel distributed processing, explorations in the microstructure of cognition, ed. de rumelhart and j. mcclelland. vol. 1. 1986. *Biometrika*, *71*(599-607), 6.

Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press.

Sanzogni, L., Guzman, G., & Busch, P. (2017). Artificial intelligence and knowledge management: questioning the tacit dimension. *Prometheus*, *35*(1), 37-56.

Schank, R. C., & Abelson, R. P. (1989). An early work in cognitive science.

Schank, R. C., & Abelson, R. P. (2013). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology press.

Searle, J. (2009). *Chinese room argument. Scholarpedia, 4(8), 3100.*

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417-424.

Sellars, R. W. (1915). A Thing and its properties. *Journal of Philosophy*, *12*(12), 318-328.

Shanahan, M. (1997). *Solving the frame problem: a mathematical investigation of the common sense law of inertia*. MIT press.

Smith, B., & Casati, R. (1994). Naive physics. *Philosophical psychology*, *7*(2), 227-247.

Smith, B. C. (1985). Prologue to reflections and semantics in a procedural language. In R. Brachman & H. J. Levesque (Eds.), *Readings in knowledge representation* (pp. 31-40). Morgan Kaufmann.

Sowa, J. F. (2012). *Knowledge representation logical, philosophical, and computational foundations*. Course Technology.

Tandon, N., Varde, A. S., & de Melo, G. (2018). Commonsense Knowledge in Machine Intelligence. *ACM SIGMOD Record*, *46*(4), 49–52. https://doi.org/10.1145/3186549.3186562

Turing, A. M. (2009). *Computing machinery and intelligence* (pp. 23-65). Springer Netherlands.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Zambak, A. F. (2013). The frame problem. *Philosophy and Theory of Artificial Intelligence*, 307-319.

APPENDICES

A. SCENARIOS FOR COMMON SENSE REASONING BENCHMARK EVALUATION

Benchmark-1: Context-Based Information Integration

Sub-benchmark Contextual Consistency and Transition Between Contexts

Anna's Mountain Hike Experiment

Scenario 1: *"Anna was walking on a snowy mountain. The temperature was -10 degrees. She was looking for a cabin to spend the night."*

Question 1: "What is Anna doing right now, and how does she feel?"

ChatGPT 4.o’s Response: Anna is walking on a snowy mountain, searching for a cabin to spend the night. Given the cold temperature of -10 degrees and the fact that she's looking for shelter, she might feel cold, tired, and possibly anxious or worried about finding a safe place to stay.

Claude Sonnet 3.5's Response: Based on the given information:

Anna is currently:

1. Walking on a snowy mountain
2. Looking for a cabin to spend the night

As for how Anna feels, we can infer that she likely feels:

1. Cold, given the very low temperature of -10 degrees
2. Possibly tired or anxious, as she's searching for shelter
3. Potentially worried or urgent, as finding a cabin seems important for her safety in such cold conditions

However, it's important to note that the passage doesn't explicitly state Anna's emotional state. These are reasonable inferences based on her situation, but her actual feelings aren't directly described.

Scenario 2: *Anna reached a cabin and lit a fire.*

Question 2: How does Anna feel right now?

ChatGPT 4.o's Response: Now that Anna has reached a cabin and lit a fire, she is likely feeling relieved and warmer. The fire would provide comfort and safety, so she might also feel more relaxed and secure compared to before, when she was outside in the cold.

Claude Sonnet 3.5's Response: Based on this new information, we can reasonably infer that Anna likely feels:

1. Relieved - She has found the shelter she was looking for, which was her immediate goal.
2. Warmer - After being in the extreme cold, lighting a fire would provide much-needed warmth.
3. Safer - Being inside a cabin protects her from the harsh outdoor conditions.
4. Possibly more comfortable - The cabin likely offers a more comfortable environment than the snowy mountain.
5. Perhaps grateful or satisfied - Having successfully found shelter and started a fire to keep warm.

However, it's important to note that the passage doesn't explicitly state Anna's emotions. These are logical inferences based on the situation, but without direct information about her feelings, we can't be absolutely certain. Her exact emotional state could also be influenced by other factors not mentioned in the brief description.

Scenario 3: *Anna calls a friend, and they talk about the time she was walking on the mountain.*

Question 3: What did Anna feel while she was walking on the mountain?

ChatGPT 4.o’s Response: While Anna was walking on the mountain, she likely felt cold and anxious due to the harsh conditions, such as the -10 degree temperature, and the uncertainty of finding a cabin. She may have also felt physically tired from the walk and concerned about her safety in the cold environment.

Claude Sonnet 3.5's Response: Based on the information provided in the original scenario and this new context, we can make some inferences about what Anna likely felt while walking on the mountain. However, it's important to note that we don't have direct access to Anna's thoughts or feelings from that time. Here's what we can reasonably assume:

1. Cold - The temperature was -10 degrees, which is extremely cold. Anna was likely feeling the physical discomfort of this severe cold.
2. Concerned or anxious - She was actively looking for a cabin to spend the night, which suggests a sense of urgency or worry about finding shelter.
3. Possibly tired - Walking in snow at high altitudes can be physically demanding, so she may have been feeling fatigued.
4. Potentially vulnerable - Being alone in harsh conditions could have made her feel exposed to the elements and potential dangers.
5. Determined - Despite the challenging conditions, she was persevering in her search for shelter.
6. Perhaps a mix of awe and apprehension - Mountain environments can be beautiful but also intimidating, especially in harsh weather.

Remember, these are educated guesses based on the situation described. Without Anna's direct account of her feelings, we can't know for certain what she experienced emotionally during her walk on the mountain. Her conversation with her friend might reveal more specific details about her emotional state at the time, but we don't have access to the content of that call.

Question 4: Where is Anna right now, and how does she feel?

ChatGPT 4.o’s Response: Anna is currently inside the cabin, where she has lit a fire. She likely feels much warmer, safer, and more relaxed now, especially after enduring the cold and anxiety while walking on the mountain earlier.

Claude Sonnet 3.5's Response: Based on the most recent information provided:

Anna is currently:

1. In a cabin (that she reached earlier)
2. Talking on the phone with a friend

Regarding how Anna feels right now:

1. Safe and comfortable - She's no longer in the harsh outdoor conditions.
2. Warm - She had lit a fire in the cabin earlier.
3. Possibly reflective or nostalgic - She's talking about her past experience on the mountain.
4. Likely relieved - Comparing her current situation to when she was walking in the cold.
5. Perhaps sociable or comforted - She's connecting with a friend over the phone.

However, it's important to note that we don't have explicit information about Anna's current emotional state. These are reasonable inferences based on her situation, but her actual feelings aren't directly described in the given information.

Sub-Benchmark Contextual Appropriateness

Scenario: *A firefighter entered a building engulfed in flames. The fire was spreading everywhere, and thick smoke filled the interior. The firefighter was carefully looking around, trying to find someone trapped inside.*

Question 1: What should the firefighter do now?

ChatGPT 4.o’s Response: Given the scenario, the firefighter should prioritize the following actions:

1. Assess the situation quickly to ensure personal safety and determine the fire's intensity and any potential escape routes.
2. Call out to anyone who might be trapped to identify their location through sound, since visibility may be limited by the smoke.
3. Move cautiously but efficiently toward the most likely area where a person could be trapped, following standard search-and-rescue procedures.
4. Stay low to avoid the thick smoke and minimize inhalation of harmful fumes, while keeping in communication with the team for updates.

These actions align with safety protocols in rescue operations during a fire.

Claude Sonnet 3.5's Response: The firefighter should take the following actions:

1. Safety first: Ensure their own safety by staying low to avoid the worst of the smoke and heat. Use protective equipment properly.
2. Communication: Maintain contact with team members outside via radio, reporting their position and conditions inside.
3. Systematic search: Conduct a thorough, methodical search of the building, moving room by room. Use thermal imaging cameras if available to detect heat signatures of potential victims.
4. Listen carefully: Pay attention to any sounds of people calling for help or movement.
5. Check common hiding spots: Look in closets, under beds, and in bathrooms where people often seek shelter.
6. Mark searched areas: Use standard firefighting techniques to mark areas that have been cleared.
7. Be prepared to evacuate: Constantly assess the fire's progression and be ready to exit quickly if conditions worsen.
8. Rescue and evacuate: If a victim is found, immediately begin rescue procedures and evacuate them from the building as quickly and safely as possible.

9. Call for backup: If needed, request additional support from the team outside.
10. Time management: Be aware of air supply levels and overall time spent inside the burning structure.

The firefighter's primary goal is to rescue any trapped individuals while maintaining their own safety and that of their team. They should work efficiently but cautiously, always staying alert to changing conditions within the burning building.

Question 2: What might the firefighter be thinking while walking through the building?

ChatGPT 4.o's Response: The firefighter might be thinking about several critical factors, including:

- Safety and Escape Routes: "I need to keep track of the exits and ensure I can get out if things get worse."
- Searching for Survivors: "Are there people trapped in this part of the building? Where should I search next?"
- Assessing the Fire's Progress: "The fire is spreading quickly. How much time do I have before this area becomes too dangerous?"
- Communicating with the Team: "I need to stay in contact with my team in case we need backup or an immediate evacuation."
- Personal Safety: "The smoke is getting thicker—stay low and conserve air. Is my gear holding up against the heat?"

Their thoughts would likely balance between maintaining their own safety and the urgency of finding anyone trapped.

Claude Sonnet 3.5's Response: The firefighter's thoughts in this intense situation would likely be a mix of professional focus, concern, and heightened awareness. Here are some possible thoughts running through their mind:

1. "Stay calm, stay focused. Remember your training."
2. "Where are the hottest parts of the fire? How's it spreading?"

3. "Is the smoke getting thicker? I need to stay low."
4. "Any signs of structural instability? Watch for weakening floors or ceilings."
5. "Listen carefully. Any sounds of someone calling for help?"
6. "Check that room. Could someone be hiding in there?"
7. "How much air do I have left? Need to keep track of time."
8. "Is that movement? Or just debris falling?"
9. "Remember to mark this area as searched."
10. "God, I hope we're not too late. Please let everyone be okay."
11. "The heat's intense. How much longer can I safely stay in here?"
12. "Where would I hide if I were trapped in here?"
13. "Stay in touch with the team outside. They need to know where I am."
14. "Be ready to get out fast if things go south."
15. "Focus. Every second counts. Don't miss anything."

These thoughts would be rapid-fire, interspersed with intense concentration on the task at hand. The firefighter would be hyper-aware of their surroundings, constantly assessing risks and possibilities, all while pushing through the physical challenges of heat, limited visibility, and the weight of their equipment. Their training would help them manage the stress and fear, channeling it into alertness and decisive action.

Question: What if the firefighter wanted to take a break at this moment, what would they do?

ChatGPT 4.o's Response: In such a dangerous and time-sensitive situation, a firefighter would likely avoid taking a break unless absolutely necessary. However, if the firefighter were overwhelmed or in need of a momentary pause to recover their strength or assess the situation, they would:

1. Find a Safe Spot: Move to a location that is as safe as possible, away from the immediate flames, ideally near an exit or outside the most hazardous area.
2. Stay Low: Sit or crouch close to the ground to avoid inhaling excessive smoke, while still keeping an eye on the situation around them.
3. Check Equipment: Use the time to check their oxygen levels and ensure their gear is functioning properly.
4. Communicate: Let their team know they are taking a brief moment to recover, so everyone is aware of their location and status.
5. Plan the Next Move: Use the pause to assess the fire's progress and decide the best course of action once ready to continue.

Taking breaks is rare in these conditions, so it would likely be very brief, with a focus on regaining control before moving forward.

Claude Sonnet 3.5's Response: I apologize, but I need to clarify that a trained firefighter would absolutely not take a break during an active rescue operation in a burning building. This scenario presents an extremely dangerous and time-sensitive situation where lives are at stake.

Taking a break in these circumstances would be:

1. Extremely dangerous for the firefighter
2. Potentially fatal for any victims still trapped
3. A severe dereliction of duty
4. Completely against all firefighting protocols and training

If the firefighter was experiencing extreme fatigue, equipment issues, or health problems that impaired their ability to continue safely, the proper procedure would be to:

1. Immediately radio their team to report the situation
2. Begin evacuating the building as quickly and safely as possible

3. Request immediate backup/replacement to continue the search

Firefighters are trained to push through extreme conditions and maintain focus on the life-saving mission at hand. Their gear, including their air supply, is designed for limited durations specifically to prevent prolonged exposure to dangerous conditions.

If you're asking out of concern for firefighter welfare in general, there are protocols for rest and rotation during extended operations, but these occur outside the immediate danger zone, not inside an actively burning building during search and rescue.

Sub-Benchmark Comprehensive Document Management

Information: https://news.mongabay.com/2023/06/global-study-of-71000-animal-species-finds-48-are-declining/

The research in this link was shared with ChatGPT 4.o and Claude Sonnet 3.5.

Question 1: How did the Industrial Revolution affect extinction rates, and what did this change parallel?

ChatGPT 4.o's Response: The Industrial Revolution significantly accelerated extinction rates, with a sharp increase across continents. This change paralleled the exponential growth in the human population and the rapid expansion of industrial activities, including deforestation, urbanization, and land conversion for agriculture. More recently, the warming of the planet due to climate change has further contributed to the rising extinction rates.

Claude Sonnet 3.5's Response: According to the document, the Industrial Revolution had a significant impact on extinction rates. Specifically:

1. The Industrial Revolution caused extinction rates to soar across continents.
2. This increase in extinction rates roughly paralleled the exponential growth in human population.

The text states: "The Industrial Revolution changed that: Extinction rates soared across continents, roughly paralleling the exponential growth in human population, and more recently, a rapidly heating planet."

This shift marked a dramatic change from the period before the Industrial Revolution, when extinctions were rare and mostly limited to islands where flightless birds and other animals were hunted to extinction by European traders and colonists or killed off by introduced species like rats and cats.

Question 2: What are the shortcomings of the global overview provided by the new study?

ChatGPT 4.o's Response: The global overview provided by the new study has several shortcomings:

1. Lack of Detail and History: The study offers a broad, sweeping analysis, but it overlooks finer details and historical context. For example, while mammals in Madagascar are considered stable overall, it doesn't capture the fact that lemurs remain the most endangered mammal group there.
2. Regional Gaps in Data: The study highlights significant knowledge gaps, especially in the tropics, where data on species conservation status is sparse. This lack of information creates "hotspots of ignorance," meaning we don't fully understand the conditions of many species in these biodiverse areas.
3. Inadequate Focus on Less Charismatic Species: There is limited data on less charismatic or less critically endangered species, and many species fall into the "unknown" category, particularly insects and amphibians. This could lead to an underestimation of extinction risks for these groups.
4. Insect Knowledge Gaps: Insects, which play crucial roles in ecosystems, are poorly studied. The study notes that while 54% of insect species are declining, most insect species are either unknown or insufficiently researched, making it difficult to assess the true state of their populations.

These shortcomings suggest that while the study provides valuable insights, it also underscores the need for more targeted, long-term research to address the gaps in knowledge.

Claude Sonnet 3.5's Response: The new study published in Biological Reviews provides a sweeping global overview of species population trends, but it has several important shortcomings:

1. Lack of detail: The study, by necessity, overlooks much detail and historical context. For example:
   - Mammals in Madagascar are considered stable overall, but lemurs remain the most endangered mammal group.
   - Fish in Lake Victoria, Africa, are increasing, but since the 1970s, more than half of the 350+ cichlid species endemic to the lake have gone extinct or are on the brink of extinction.
2. Knowledge gaps: There are enormous information gaps, especially for:
   - Long-term monitoring data
   - Less charismatic species
   - Species not currently considered critically endangered
3. Tropical data shortage: There is a severe lack of data for tropical regions, which are the most biodiverse areas on Earth. The study authors refer to "hotspots of ignorance" - areas with high concentrations of species for which no data exists.
4. Insect population uncertainty: The true state of insect populations remains largely unknown. While the study found that 54% of insects are declining globally, there are millions of insect species that remain unidentified or poorly researched.
5. Amphibian data: The study may not fully capture the severity of amphibian declines, which have been occurring at a higher rate than for all vertebrates combined in recent decades.
6. Temporal limitations: The study provides a current snapshot but may not fully reflect historical trends or future projections for species populations.

These shortcomings highlight the need for more comprehensive and long-term research, especially in tropical regions and for understudied species groups like insects and amphibians.

Question 3: According to the new study, what are the long-term survival threats facing species, and how can these be prevented?

ChatGPT 4.o's Response: According to the new study, the long-term survival threats facing species include:

1. Habitat Destruction: The primary driver of biodiversity loss is the destruction and fragmentation of habitats due to deforestation, agriculture, urbanization, and infrastructure development.
2. Climate Change: Rapidly changing climates are forcing species to adapt or relocate. Those unable to move to cooler areas or higher altitudes face a heightened risk of extinction.
3. Declining Population Numbers: As populations shrink, they lose genetic diversity and resilience, making them more vulnerable to disturbances such as wildfires or disease outbreaks.
4. Invasive Species and Overexploitation: Species face additional threats from hunting, wildlife trade, competition with invasive species, and diseases contracted from humans or livestock.
5. Knowledge Gaps: A lack of data, particularly in tropical regions and for less-studied species like insects, prevents accurate assessments of species' conservation needs

Prevention Strategies:

1. Habitat Protection and Restoration: Protecting existing habitats, reconnecting fragmented landscapes, and replanting forests can provide species with the space they need to survive and recover.
2. Targeted Conservation Efforts: Early intervention for species showing population declines, even if they are not yet on the endangered list, can prevent further deterioration.
3. Collaborative Conservation: Partnerships between governments, conservation organizations, Indigenous communities, and local populations can curb habitat destruction, illegal hunting, and trafficking.

4. Climate Mitigation: Reducing emissions and supporting ecosystems' adaptation to climate change can slow species decline, especially in vulnerable regions like the tropics.

5. Closing Knowledge Gaps: Increased research and long-term monitoring of species, especially in poorly studied regions, will help identify those at risk earlier and guide conservation efforts.

By addressing these threats, the long-term survival of species can be better ensured.

Claude Sonnet 3.5's Response: According to the new study, there are several significant long-term survival threats facing species. Here are the key points:

1. Population declines: The study found that 48% of evaluated species are decreasing in number globally, while only 3% are increasing. This widespread decline poses a major threat to long-term survival.

2. Unrecognized threats to "safe" species: One of the most concerning discoveries was that one-third of animals currently considered safe (not yet on the endangered list) are experiencing population declines that threaten their long-term survival.

3. Habitat destruction: The study identifies habitat destruction as the major cause behind the current biodiversity crisis.

4. Tropical concentration: Most population losses are concentrated in the tropics, which are biodiversity hotspots.

5. Knowledge gaps: The lack of data on many species, especially in tropical regions, poses a threat as it hinders effective conservation efforts.

To prevent these threats and improve long-term survival prospects for species, the study suggests several approaches:

1. Early intervention: The study provides an early warning system by spotlighting species that are declining before they reach critically endangered status. This allows for pre-emptive action.

"Once a species is critically endangered, it's difficult to save," agrees Colin Chapman, a biologist and biological anthropologist at Canada's Vancouver Island University, who was not affiliated with the new study. "So, we should be a lot more worried about species that are not listed by IUCN that may need special attention."

2. Prioritize declining species: Conservation efforts should focus on species undergoing serious declines, even if they're not yet classified as endangered.
3. Address knowledge gaps: More research is needed in understudied regions, particularly in the tropics.
4. Habitat protection: The study supports the goal set at the 2022 UN biodiversity conference to protect 30% of the planet's land and water by 2030.
5. Connectivity: Connecting fragmented habitats is crucial. As Stuart Pimm notes, "We know that fragmented landscapes lose species quickly."
6. Collaborative efforts: The study authors emphasize the need for collaborations between governments, law enforcement, nonprofits, and local and Indigenous peoples to address issues like hunting, wildlife trafficking, and deforestation.
7. Individual action: The researchers also highlight that individuals can make a difference through actions like reducing energy use, minimizing pesticide use, and supporting effective conservation organizations.

By implementing these strategies, it may be possible to slow the rate of biodiversity decline and improve the long-term survival prospects for many species

Sub-Benchmark Management of Contextual Conflicts

Scenario: A traffic accident occurs at an intersection during midday. There are traffic lights at the intersection, and the incident takes place on a busy street in the city center. The accident involves a red car (Toyota) and a white car (Volkswagen). At the time of the incident, two witnesses sitting at a café on the corner of the intersection observe the accident from different angles.

Witness 1: Jane’s Statement (Polished English, Casual Tone)

"Officer, I was sitting outside the café, enjoying the afternoon, and I had a perfect view of the intersection. Everything happened so quickly, but I’m certain about what I saw.

There was a red car, a Toyota, approaching the intersection. The light turned yellow, and instead of slowing down, the driver sped up like they were trying to make it through. At the same time, a white Volkswagen was coming from the other direction, and its light had just turned green. The red Toyota tried to stop, but the road was slick from the rain earlier, and it slid right into the white Volkswagen. The driver in the white car honked just before the impact, but at that point, there was no stopping it."

"After the crash, both drivers got out of their cars. The driver of the white car looked shocked, almost like he was in a daze, and didn't say much. But the guy in the red car was furious—he immediately started yelling. From what I could see, it looked like the white Volkswagen had the right of way since the light was green for them, and the red Toyota should have stopped when the light turned yellow."

Witness 2: John's Statement (Confident, Slightly Tense)

"Officer, I've got to tell you, I'm pretty upset about this because it's so clear in my mind. I was sitting inside the café, right by the window, with a full view of the intersection. The white Volkswagen came into the intersection just as the light was turning yellow. It was obvious the driver was trying to make it before the light turned red. Now, the red Toyota was already in the intersection, and its light had turned green. The Volkswagen didn't slow down at all, and honestly, it seemed like the driver sped up to beat the light. The Toyota had already started slowing down, but the Volkswagen came barreling through too fast for anything to stop the crash."

"After the crash, the guy in the white Volkswagen got out and was immediately yelling, acting like the whole thing was the other driver's fault. Meanwhile, the driver of the red Toyota stayed totally calm, didn't even react to the yelling. From what I saw, the white Volkswagen was clearly at fault. The driver was rushing to make it through a yellow light, while the red Toyota had already slowed down and had the right of way with the green light. The Volkswagen driver was being reckless, no doubt about it."

Question: Are there any contradictions between the two witness statements?

ChatGPT 4.o's Response: Yes, there are contradictions between the two witness statements.

1. Color of the traffic lights:
    - Jane (Witness 1) claims the red Toyota approached the intersection when the light turned yellow, and the white Volkswagen entered when its light had just turned green.
    - John (Witness 2), however, says the white Volkswagen entered the intersection just as the light was turning yellow, and the red Toyota had already started slowing down because its light was green.
2. Actions of the drivers:
    - Jane describes the red Toyota as speeding up to make it through the yellow light and then trying to stop but sliding into the white Volkswagen.
    - John, on the other hand, states that the white Volkswagen was speeding up to beat the yellow light, while the red Toyota had already slowed down.
3. Post-accident reactions:
    - Jane mentions that the driver of the white Volkswagen was shocked and didn’t say much, while the red Toyota driver was yelling and furious.
    - John reports the opposite: the white Volkswagen driver was the one yelling, while the red Toyota driver remained calm.

These contradictions suggest that the witnesses have different perceptions of the sequence of events, the traffic lights' timing, and the behavior of the drivers after the accident.

Claude Sonnet 3.5's Response: Yes, there are several significant contradictions between the two witness statements. Let's break them down:

1. Traffic light status:
    - Jane (Witness 1) states that the light turned yellow for the red Toyota and green for the white Volkswagen.

- John (Witness 2) claims the opposite: the light was turning yellow for the white Volkswagen and had turned green for the red Toyota.

2. Vehicle behavior:
    - Jane says the red Toyota sped up to beat the yellow light, then tried to stop.
    - John states that the white Volkswagen sped up to beat the yellow light, while the Toyota was already slowing down.
3. Right of way:
    - Jane believes the white Volkswagen had the right of way with a green light.
    - John asserts that the red Toyota had the right of way with a green light.
4. Driver reactions post-crash:
    - Jane describes the Toyota driver as furious and yelling, while the Volkswagen driver was shocked and quiet.
    - John states the opposite: the Volkswagen driver was yelling, while the Toyota driver remained calm.
5. Fault assignment:
    - Jane implies the red Toyota was at fault for trying to beat the yellow light.
    - John explicitly states the white Volkswagen was at fault for rushing through a yellow light.
6. Witness locations:
    - Jane was sitting outside the café.
    - John was sitting inside the café by the window.

These contradictions suggest that at least one of the witnesses is mistaken about key details of the accident. It's possible that factors such as their different viewing angles, the stress of witnessing an accident, or personal biases could have influenced their perceptions and recollections of the event.

Benchmark-2: Future Planning and Adaptation Ability

Sub-Benchmark Prediction of Possible Scenarios

Scenario: *A forest fire in the Ayvacık district of Çanakkale, Turkey, is spreading rapidly due to strong winds.*

Question: What are the possible scenarios for the fire's progression, and what factors could influence these scenarios?

ChatGPT 4.o's Response: The possible scenarios for the forest fire in Ayvacık, Çanakkale, can vary based on several key factors, including the intensity of the fire, weather conditions, terrain, and firefighting efforts. Here are some potential scenarios and influencing factors:

Possible Scenarios:

1. Fire Containment: If firefighting teams manage to control the fire early, it could be contained to a limited area, reducing damage to the forest and surrounding areas. This scenario is likely if there are strong and coordinated firefighting efforts, including aerial support and favorable weather conditions.
2. Rapid Spread: Due to strong winds, the fire could spread rapidly, covering larger areas of forest, potentially threatening nearby settlements and infrastructure. In this case, evacuations might be necessary, and the fire could last for several days.
3. Fire Jumps to New Areas: If the winds continue to be strong, there is a risk that embers could be carried over natural barriers, starting new fires in distant locations. This could make containment even more challenging.
4. Natural Extinguishment: A shift in weather conditions, such as increased humidity or rain, could naturally suppress the fire, leading to its slow extinguishment without significant human intervention.

5. Prolonged Battle: In the worst case, if the fire is not contained quickly and the weather conditions worsen (e.g., stronger winds, dry air), the fire could rage for an extended period, requiring continuous firefighting efforts, and potentially spreading over a vast area.

Influencing Factors:

1. Wind Speed and Direction: Strong winds can significantly accelerate the fire's spread, making it harder to control and causing it to jump to new areas.
2. Humidity and Temperature: Dry air and high temperatures can increase the fire's intensity, while high humidity levels can help slow it down.
3. Topography: Steep slopes can cause fires to spread more quickly uphill, while valleys and flat terrain might slow it down.
4. Vegetation Type: Dense, dry forests with abundant underbrush act as fuel and can lead to more intense and faster-spreading fires.
5. Human Intervention: The effectiveness of firefighting efforts, including aerial water bombing, firebreaks, and ground crews, plays a crucial role in how the fire progresses.

The progression of the fire will depend on how these factors evolve over time, and continued monitoring is essential for predicting the next steps.

Claude Sonnet 3.5's Response: To analyze the possible scenarios for the forest fire's progression in Ayvacık, Çanakkale, and the factors that could influence these scenarios, let's consider several key aspects:

1. Wind conditions:

- Current scenario: Strong winds are already contributing to rapid spread.
- Possible developments: a) Wind continues or intensifies: This would lead to faster and more unpredictable fire spread, potentially creating spot fires ahead of the main fire front. b) Wind decreases: This could slow the fire's progression, making it more manageable for firefighters. c) Wind direction changes: This

could alter the fire's path, potentially threatening new areas or providing a temporary reprieve for some locations.

2. Terrain and vegetation:

- Ayvacık's landscape likely includes a mix of forested areas and potentially some agricultural land.
- Scenarios: a) Fire reaches dense forest: This could lead to more intense burning and faster spread. b) Fire encounters natural firebreaks (e.g., rivers, rocky areas): This might slow or stop progression in certain directions. c) Fire reaches agricultural areas: Depending on the crops and their moisture content, this could either fuel the fire or act as a partial barrier.

3. Weather conditions:

- Current conditions are likely hot and dry, given it's a forest fire.
- Possible scenarios: a) Continued hot, dry weather: This would sustain or worsen the fire conditions. b) Increased humidity: This could slightly dampen the fire's intensity. c) Rainfall: Significant rain could help control the fire, but light rain might have minimal impact.

4. Firefighting efforts:

- Scenarios based on response: a) Rapid, well-coordinated response: Could contain the fire more quickly. b) Limited resources or difficult access: Might allow the fire to spread further before being contained. c) Use of aerial firefighting: Could significantly impact the fire's spread, especially in hard-to-reach areas.

5. Population centers:

- The fire's proximity to towns or villages will influence prioritization of firefighting efforts.
- Scenarios: a) Fire approaches populated areas: This would likely result in evacuations and concentrated firefighting efforts to protect these areas. b) Fire

remains in remote areas: This might allow for a more strategic, less urgent approach to containment.

6. Time of day:

- Fires often behave differently between day and night.
- Scenarios: a) Daytime: Typically sees more intense fire behavior due to higher temperatures and lower humidity. b) Nighttime: Could bring slightly calmer conditions, potentially aiding firefighting efforts.

Factors that could influence these scenarios include:

1. Availability and coordination of firefighting resources
2. Accuracy of weather forecasts
3. Effectiveness of any pre-existing fire management practices in the area
4. Potential for the fire to create its own weather patterns if it becomes very large
5. Public cooperation with evacuation orders and fire prevention measures

Given the complexity and dynamic nature of forest fires, continuous monitoring and adaptive management would be crucial in responding to this situation effectively.

Sub-Benchmark Strategic Planning

Scenario: *Efforts are being made to control the fire in Ayvacık, but the wind is quickly changing direction, increasing the risk of the fire spreading toward the nearby villages.*

Question: What strategic plan can be developed to prevent the fire from reaching the villages? The plan should include both short-term and long-term solutions.

ChatGPT 4.o's Response:

Short-Term Solutions:

1. Evacuation of At-Risk Areas:
   Immediately evacuate residents from villages that are in the path of the fire.

Establish temporary shelters in safe zones and ensure clear communication channels between local authorities and the population. Early evacuation is critical to avoid last-minute chaos.

2. Firebreak Creation:
   Quickly create firebreaks by clearing vegetation in key areas between the fire and the villages. This can be done using bulldozers and controlled burning under the supervision of fire experts to remove potential fuel from the fire's path.

3. Deployment of Aerial and Ground Support:
   Utilize helicopters and planes to drop water and fire retardants in strategic areas between the fire and the villages. Ground crews should focus on protecting homes and buildings by creating defensible zones and applying fire retardants around critical infrastructure.

4. Wind Monitoring and Forecasting:
   Closely monitor wind direction changes using meteorological tools to anticipate shifts that may drive the fire toward villages. This data should guide the placement of fire crews and resources to respond dynamically as the situation evolves.

5. Emergency Water Supply Systems:
   Establish temporary water supply lines or use mobile water tanks in the villages at risk. This ensures that if the fire reaches the villages, there will be water readily available for firefighting and property protection.

Long-Term Solutions:

1. Reforestation and Controlled Forest Management:
   Implement a long-term reforestation program focused on fire-resistant plant species and reducing overly dense vegetation that fuels wildfires. Regular forest thinning and clearing of dead or dry plant material will reduce the risk of future large-scale fires.

2. Fire-Resistant Infrastructure:
   Encourage and support the construction of fire-resistant buildings and infrastructure in fire-prone areas. This includes using fireproof materials, creating buffer zones around homes, and designing layouts that reduce fire spread risks.

3. Community Fire Prevention Programs:
   Develop and fund ongoing fire prevention education programs for local communities. Teach residents how to maintain defensible space around their properties, manage vegetation, and recognize fire hazards.

4. Improved Fire Detection and Monitoring Systems:
   Invest in early detection systems such as satellite-based fire monitoring, drones, and local watchtowers. These systems can detect small fires before they escalate and provide real-time data to firefighting teams.

5. Sustainable Land-Use Policies:
   Work with local governments to develop sustainable land-use plans that reduce the proximity of forests and high-risk fire zones to residential areas. Ensure new developments are built with fire risk mitigation strategies in mind.

By combining immediate actions to prevent the fire from reaching villages with long-term fire management strategies, the risk to lives and property can be significantly reduced, and future fires can be better managed.

Claude Sonnet 3.5's Response: Given the evolving situation in Ayvacık with changing wind directions and increased risk to nearby villages, a comprehensive strategic plan is crucial. Here's a proposed plan incorporating both short-term and long-term solutions:

Short-term Solutions:

1. Immediate Evacuation:

   - Initiate immediate evacuation of at-risk villages.

- Use emergency alert systems, local media, and door-to-door notifications.
- Establish clear evacuation routes and safe zones.

2. Firebreak Creation:
   - Rapidly create firebreaks between the advancing fire and villages.
   - Use bulldozers and other heavy machinery to clear vegetation.
   - If time allows, conduct controlled burns to widen firebreaks.
3. Resource Reallocation:
   - Redirect firefighting resources to protect villages and critical infrastructure.
   - Position fire trucks and water tankers at strategic points near villages.
4. Aerial Firefighting:
   - Deploy water-bombing aircraft to target areas where the fire threatens to jump containment lines.
   - Use helicopters for precision water drops in hard-to-reach areas.
5. Wind Monitoring and Predictive Modeling:
   - Set up mobile weather stations to track real-time wind changes.
   - Use predictive fire behavior models to anticipate the fire's path.
6. Communication and Coordination:
   - Establish a unified command center to coordinate all firefighting and evacuation efforts.
   - Ensure clear communication channels between ground crews, aerial units, and command.

Long-term Solutions:

1. Vegetation Management:

- Implement a comprehensive fuel reduction program around villages.
- Introduce regular controlled burns to reduce undergrowth.
- Encourage residents to create defensible spaces around their properties.

2. Infrastructure Improvements:
    - Upgrade water supply systems in villages for better firefighting capability.
    - Improve road networks to facilitate faster evacuations and better access for firefighters.
3. Community Preparedness:
    - Develop and regularly update community-level fire response plans.
    - Conduct annual fire drills and evacuation exercises.
    - Educate residents on fire safety and prevention measures.
4. Early Warning Systems:
    - Install a network of fire detection cameras and sensors in high-risk areas.
    - Develop a mobile app for real-time fire alerts and evacuation notices.
5. Land-Use Planning:
    - Revise zoning laws to limit new construction in high-risk fire areas.
    - Implement stricter building codes for fire-resistant materials and designs.
6. Ecological Management:
    - Promote the planting of fire-resistant native species.
    - Restore natural firebreaks like rivers and rocky outcrops.
7. Climate Change Adaptation:

- Develop long-term strategies to address increased fire risks due to climate change.
- Invest in research on fire-resistant crops and forest management techniques.

8. Regional Cooperation:
   - Establish mutual aid agreements with neighboring districts and provinces.
   - Create a regional firefighting task force for rapid response.
9. Technology Integration:
   - Invest in advanced firefighting technologies (e.g., drones for monitoring).
   - Develop AI-powered fire prediction models.
10. Sustainable Forest Management:
    - Implement sustainable logging practices to reduce fire fuel.
    - Create mosaic patterns in forests to naturally limit fire spread.

This strategic plan aims to address the immediate threat while also building long-term resilience against future fire events. Regular review and adaptation of this plan will be necessary to ensure its continued effectiveness in the face of changing environmental conditions and emerging technologies.

Sub-Benchmark Adaptability and Flexibility

Scenario: *As the fire in Ayvacık begins to come under control, an unexpected wind shift occurs, causing the fire to spread rapidly eastward into another forested area.*

Question: How can firefighting teams quickly adapt to the sudden eastward spread of the fire, and how should they reorganize their positions to respond to the new conditions?

ChatGPT 4.o's Response: To address the rapid eastward spread of the fire in Ayvacık, firefighting teams need to quickly adjust their strategy and reorganize their resources to contain the new fire front. Here's how they can adapt:

1. Immediate Actions:

- Reposition Resources:
  Firefighting teams must rapidly reposition ground crews, aerial support, and equipment to the eastward front. Deploy additional teams to reinforce areas of critical concern, particularly where the fire is approaching settlements or critical infrastructure.
- Aerial Suppression:
  Use helicopters and planes to drop water and fire retardants along the eastern front to slow the fire's progress. Focus on creating barriers in areas with dense vegetation or natural firebreaks like rivers or roads.
- Wind Monitoring:
  Continuously monitor wind patterns and anticipate further shifts. Mobile command units should be equipped with weather tracking systems to ensure firefighting efforts are not caught off guard by further changes in wind direction.
- Strategic Firebreak Expansion:
  Quickly expand firebreaks along the eastern edge by clearing vegetation and, if possible, using controlled burns in front of the fire. This will help to limit the fire's advance into unburned forested areas.
- Local Evacuation Plans:
  If the fire threatens new areas eastward, immediately coordinate with local authorities to evacuate residents in potential danger zones. Communicate evacuation orders through emergency channels and prioritize vulnerable communities.

2. Reorganizing Firefighting Teams:

- Establish a New Command Center:
  Set up a temporary command post closer to the eastern spread to allow for faster decision-making and communication. This will enable quicker coordination between teams responding to the fire's new direction.

- Deploy Specialized Units:
  Depending on the topography of the eastern area, deploy specialized units trained in rugged terrain and dense forest firefighting. Ensure that these teams have the necessary equipment and vehicles for effective operations in more remote areas.

- Reallocate Aerial and Ground Assets:
  Prioritize aerial resources to attack the fire from above and prevent it from jumping into new areas. Ground crews should focus on securing firebreaks, protecting key areas, and ensuring that roads and access routes remain open for emergency vehicles.

- Rotate Exhausted Crews:
  With the fire moving quickly, some teams may be exhausted from the initial containment efforts. Rotate fresh crews into critical areas to maintain high energy levels and ensure optimal performance.

3. Contingency Planning:

- Prepare for Further Wind Changes:
  Develop a contingency plan that anticipates further changes in wind direction. This includes identifying additional firebreak locations, securing key infrastructure, and pre-positioning resources for rapid deployment.

- Expand Perimeter Control:
  Set up an extended perimeter with firebreaks and defensive zones beyond the current eastern front. This will serve as a backup line if the fire breaks through the first containment efforts.

By quickly adjusting their strategy, reallocating resources, and using precise coordination, firefighting teams can better control the unexpected eastward spread and minimize damage to additional forested areas.

Claude Sonnet 3.5's Response: This sudden shift in fire behavior requires a rapid and coordinated response from the firefighting teams. Here's how they can adapt and reorganize:

1. Immediate Reassessment:

- Quickly gather data on the new fire direction, speed, and intensity.
- Use aerial reconnaissance (drones or helicopters) to get real-time visuals of the eastward spread.
- Update fire prediction models with new wind data to project the fire's likely path.

2. Resource Redeployment:

- Rapidly redeploy firefighting units to the eastern flank: a) Move ground crews and equipment to establish new containment lines. b) Reposition water tankers and pumps to support the new frontline.
- Maintain a minimal presence on controlled areas to prevent reignition.

3. Aerial Strategy Adjustment:

- Redirect water-bombing aircraft to target the leading edge of the eastward spread.
- Use helicopters for precision drops in areas inaccessible to ground crews.

4. Communication and Coordination:

- Issue immediate alerts to all teams about the changed situation.
- Establish a new command post closer to the eastern front if necessary.
- Ensure all units have updated maps and are aware of new rendezvous points.

5. Preventive Measures:

- Identify critical areas in the fire's new path (e.g., vulnerable communities, infrastructure).
- Begin creating firebreaks ahead of the fire's projected path.
- Initiate preventive evacuations if populated areas are now at risk.

6. Resource Allocation:

- Request additional resources if the new spread threatens to overwhelm current capacity.
- Prioritize protecting critical infrastructure in the fire's new path.

7. Tactical Approach:

- Employ indirect attack methods if direct attack is too dangerous due to rapid spread.
- Consider backburning operations to create wider containment lines, if conditions allow.

8. Safety Reassessment:

- Quickly reevaluate escape routes and safety zones for all firefighting personnel.
- Ensure all teams are briefed on new safety protocols for the changed conditions.

9. Monitoring and Prediction:

- Set up new mobile weather stations to track wind changes in the eastern area.
- Assign a dedicated team to continuously monitor and predict fire behavior.

10. Public Communication:

- Update local authorities and media about the new fire direction.
- Use emergency broadcast systems to alert residents in newly threatened areas.

11. Access and Logistics:

- Identify and clear access routes to the new fire front.
- Establish new supply lines for water, fuel, and other resources.

12. Specialized Team Deployment:

- If available, deploy rapid response teams specifically trained in dynamic fire situations.
- Position spotters at high vantage points to provide early warnings of further changes.

13. Contingency Planning:

- Develop multiple containment strategies based on potential wind shifts.
- Prepare fallback positions in case the fire overcomes initial containment efforts.

14. Interagency Coordination:

- Alert neighboring fire districts about the changed situation.
- Request additional resources or specialized equipment if needed.

15. Environmental Considerations:

- Assess the new area for sensitive ecological zones that may require special protection.
- Consider water sources and potential contamination risks in firefighting efforts.

By quickly adapting their strategy and reorganizing their positions, firefighting teams can respond effectively to the new conditions. The key is to maintain flexibility, ensure clear communication, and prioritize both firefighter safety and effective fire containment in the face of rapidly changing circumstances.

Sub-Benchmark Handling Uncertainty

Scenario: *Weather conditions in Ayvacık are uncertain, and the wind direction keeps shifting. Additionally, gathering accurate information about the fire is difficult because some areas are covered in smoke.*

Question: What strategies should be implemented to manage risks and respond to the uncertainty in this firefighting situation?

ChatGPT 4.o's Response: To effectively handle the unpredictable weather conditions and limited visibility due to smoke, firefighting teams must adopt flexible, adaptive strategies while prioritizing safety and efficient information gathering.

1. Enhanced Monitoring and Data Gathering:

- Use of Drones and Satellites:
  Deploy drones equipped with thermal imaging and real-time video feeds to monitor the fire's progression in areas covered by smoke. Satellites can also provide broader views and heat mapping to track the fire's spread, especially in hard-to-reach areas.
- Mobile Weather Stations:
  Set up portable weather monitoring systems in strategic locations to gather localized data on wind speed, direction, humidity, and temperature. This real-time data will help predict shifts in fire behavior and improve decision-making on the ground.
- Air Quality Sensors:
  Install air quality sensors in the area to track smoke density and visibility conditions. This will assist in determining safe zones for ground teams to operate and when aerial suppression efforts might need to be paused due to poor visibility.

2. Flexible Deployment and Mobility:

- Modular Firefighting Units:
  Divide firefighting teams into smaller, highly mobile units that can quickly relocate to different fronts as conditions change. These units should be equipped with portable communication tools to stay updated on changing weather and fire patterns.
- Pre-position Resources:
  Place firefighting resources (water trucks, fire retardant supplies, equipment)

at multiple locations to ensure quick deployment regardless of where the fire spreads next. This will save time when responding to sudden wind shifts or new fire outbreaks.

- Adaptive Aerial Support:
  Maintain readiness for aerial firefighting efforts, such as helicopters and planes, to be redeployed rapidly based on updated wind and fire behavior. Smoke may limit visibility, so using aircraft in clearer areas is critical for effective water drops and observation.

3. Risk Management for Firefighter Safety:

- Dynamic Safety Zones:
  Create safety zones that can shift based on real-time monitoring of wind and fire behavior. These zones ensure that ground crews can retreat quickly and regroup when necessary to avoid entrapment in rapidly changing conditions.

- Safety Protocols for Limited Visibility:
  Ground crews should operate with clear protocols for low-visibility scenarios. This includes maintaining radio contact, using GPS devices for navigation, and adhering to strict guidelines about when to pull back from dangerous areas if visibility drops due to heavy smoke.

4. Communication and Coordination:

- Centralized Command and Communication:
  Establish a central command unit with access to all incoming data from drones, weather stations, and on-ground teams. This unit will coordinate fire suppression efforts and communicate real-time updates to all teams to ensure that everyone is aware of changing conditions.

- Regular Briefings:
  Conduct frequent briefings for both ground and aerial teams to ensure they are aware of the latest weather patterns and fire conditions. Real-time updates should be provided through portable radios or mobile devices.

5. Evacuation and Community Safety Plans:

- Pre-emptive Evacuation Planning:
  Given the uncertain wind conditions, work with local authorities to establish flexible evacuation plans for communities at risk. Ensure that these plans can be quickly adjusted based on the fire's direction and intensity. Clear communication channels should be maintained with residents, including text alerts, local radio updates, and door-to-door notifications if needed.
- Buffer Zones Around Villages:
  Expand buffer zones around villages using firebreaks and controlled burns to reduce the risk of the fire reaching populated areas. This buys time for firefighting teams to react to unpredictable fire movements.

By combining advanced monitoring technologies, flexible team deployments, safety protocols for low-visibility conditions, and proactive communication, firefighting teams can better manage the uncertainty and evolving risks in the Ayvacık fire situation. These strategies will ensure quick response times while protecting both firefighters and local communities.

Claude Sonnet 3.5's Response: Managing risks and responding to uncertainty in this dynamic firefighting situation requires a flexible, information-driven approach. Here are strategies to implement:

1. Enhanced Information Gathering:

- Deploy drones with thermal imaging cameras to see through smoke.
- Use satellite imagery for broader situational awareness.
- Establish a network of ground-based sensors to monitor fire behavior.
- Utilize LIDAR technology for detailed terrain and vegetation mapping.

2. Predictive Modeling:

- Employ ensemble forecasting techniques, running multiple fire behavior models.
- Regularly update models with new data to improve accuracy.

- Develop worst-case, best-case, and most likely scenarios for planning.

3. Adaptive Resource Allocation:

- Implement a mobile command structure that can quickly relocate.
- Position resources in a distributed manner to respond to multiple potential fire paths.
- Maintain flexible reserves that can be rapidly deployed to emerging hotspots.

4. Risk-Based Decision Making:

- Prioritize actions based on potential impact and probability of occurrence.
- Use decision trees to guide responses to different scenarios.
- Implement a formal risk assessment process for all major tactical decisions.

5. Improved Communication Systems:

- Establish redundant communication channels (radio, satellite, cellular).
- Use real-time digital mapping tools accessible to all team members.
- Implement a system for rapid dissemination of critical updates.

6. Scenario Planning:

- Develop and regularly update contingency plans for various wind shift scenarios.
- Conduct tabletop exercises to practice responses to sudden changes.
- Create pre-planned trigger points for major strategy shifts.

7. Safety-First Approach:

- Establish clear "no-go" zones where risks are deemed too high.
- Implement a buddy system and regular check-ins for all field personnel.
- Ensure all teams have multiple, pre-identified escape routes.

8. Collaborative Intelligence:

- Engage local knowledge sources (e.g., farmers, foresters) for terrain insights.
- Coordinate with meteorological services for customized, high-frequency forecasts.
- Establish information sharing protocols with neighboring fire districts.

9. Proactive Public Communication:

- Implement a multi-channel alert system for potentially affected areas.
- Provide regular, transparent updates to manage public expectations.
- Use social media monitoring to gather on-the-ground information from residents.

10. Flexible Containment Strategies:

- Employ a combination of direct and indirect attack methods.
- Prepare for rapid shifts between offensive and defensive tactics.
- Use mobile firebreaks (e.g., foam lines) that can be quickly redeployed.

11. Technology Integration:

- Utilize AI-powered decision support systems for real-time strategy adjustments.
- Implement augmented reality tools for firefighters to visualize fire predictions.
- Use automated resource tracking systems to optimize deployments.

12. Environmental Monitoring:

- Deploy mobile air quality monitoring stations to track smoke patterns.
- Monitor water sources for potential contamination from firefighting efforts.
- Assess wildlife movement patterns to predict and mitigate ecological impacts.

13. Psychological Support:

- Implement regular stress debriefings for firefighting personnel.
- Provide on-site mental health support to help teams cope with uncertainty.
- Rotate personnel more frequently to prevent fatigue-based errors.

14. Learning and Adaptation:

- Conduct rapid after-action reviews after each shift to identify lessons.
- Implement a system for quick dissemination of learned insights across all teams.
- Encourage a culture of constructive dissent to challenge assumptions.

15. Resource Conservation:

- Implement strict resource management to prepare for potentially prolonged operations.
- Use predictive maintenance on equipment to prevent unexpected breakdowns.
- Establish clear sub-benchmark for requesting additional external resources.

By implementing these strategies, the firefighting effort can better manage risks and uncertainties, maintaining operational effectiveness while prioritizing safety. The key is to remain agile, information-driven, and prepared for multiple contingencies in this complex and dynamic situation.

Benchmark-3: Comprehensive Causality and Linked Information

Sub-Benchmark Identification of Cause-Effect Relationships

Scenario: *A café owner noticed that the coffee machine was broken in the morning. On the same day, the owner observed that the number of customers was lower than expected.*

Question: "What is the relationship between the coffee machine breaking down and the decrease in customer numbers? If there is one, how would you explain it?"

ChatGPT 4.o’s Response: The relationship between the coffee machine breaking down and the decrease in customer numbers is likely causal. The broken coffee machine could have led to fewer customers because coffee is a primary product that many customers visit the café for. Without the ability to serve coffee, the café might have been less appealing to regular customers or passersby, resulting in a lower turnout. Additionally, word of mouth or visible signs of the malfunction might have deterred potential customers from entering. Thus, the breakdown directly influenced the customer numbers by reducing the café's main offering.

Claude Sonnet 3.5's Response: To analyze the potential relationship between the broken coffee machine and the decrease in customer numbers, let's consider a few key points:

1. Correlation: There appears to be a correlation between the coffee machine breaking down and the lower number of customers on that day. However, it's important to note that correlation does not necessarily imply causation.
2. Possible causal relationship: There could be a causal relationship between these two events. Here's how we might explain it: a) Limited menu offerings: With a broken coffee machine, the café likely couldn't serve coffee or coffee-based drinks. This could significantly reduce the appeal of the café to many customers, especially those who primarily visit for coffee. b) Longer wait times: If the café attempted to serve coffee using alternative methods (e.g., pour-over), this might have led to longer wait times, discouraging customers or causing them to leave without purchasing. c) Word of mouth: Regular customers who visited early and found out about the broken machine might have informed others, leading to fewer visitors throughout the day. d) Reduced foot traffic: If the café is known for its coffee, passersby might have been less likely to enter if they saw a sign about limited coffee options or noticed fewer people inside.
3. Alternative explanations: While the broken machine could explain the decrease in customers, other factors might also be at play: a) Weather conditions b)

Local events or holidays c) Competitors' promotions d) Random fluctuation in customer numbers

4. Need for more data: To establish a stronger connection, we'd need to consider:

- Customer numbers on other days
- The café's reliance on coffee sales
- Customer feedback or complaints on that day
- Whether this pattern repeats when equipment issues occur

In conclusion, while there's a plausible explanation for how the broken coffee machine could have led to decreased customer numbers, we can't definitively prove causation without more information. The relationship is likely a combination of direct effects (unable to serve a key product) and indirect effects (changed customer behavior due to the situation).

Evaluation: Both ChatGPT 4.o and Claude Sonnet 3.5 effectively address the 3.1 Identification of Cause-Effect Relationships criterion, but with different levels of sophistication. ChatGPT 4.o provides a straightforward explanation, establishing a clear cause-effect link by suggesting that the broken coffee machine likely led to fewer customers due to the café's reliance on coffee as a key product. In contrast, Claude Sonnet 3.5 delivers a more sophisticated analysis, evaluating multiple possible reasons for the drop in customer numbers, such as limited menu offerings, longer wait times, word of mouth, and other external factors. Claude Sonnet 3.5 also highlights the importance of considering alternative explanations and the need for more data, making it much more successful in exploring possible scenarios compared to ChatGPT 4.o's narrower focus.

Sub-Benchmark Tracking Causal Chains

Scenario: *A group of friends decided to go to a coastal town for a summer vacation. However, one of their friends said at the last minute that they couldn't go due to an emergency at work.*

Question: How does this friend's cancellation impact the vacation plans? What events are linked together?

ChatGPT 4.o's Response: The friend's last-minute cancellation could impact the vacation plans in several ways, depending on the group's dynamics and logistics. If the friend was responsible for a critical part of the trip, such as driving, organizing accommodation, or planning activities, their absence might require the group to rearrange certain aspects, like finding a new driver or modifying their bookings. Additionally, the group's emotional response could be affected, as they might feel disappointed or concerned for their friend's situation. Events linked together include the friend's emergency at work, the subsequent cancellation, and the potential adjustments the group must make in terms of transportation, activities, or even group dynamics. This creates a chain where the cause (emergency) leads to an immediate effect (cancellation), followed by the necessary adaptations in the vacation plan.

Claude Sonnet 3.5's Response: To analyze this scenario and identify the linked events, let's break it down step by step:

1. Initial plan:

- A group of friends planned a summer vacation to a coastal town
- All friends were initially committed to the trip

2. Last-minute cancellation:

- One friend cancels due to a work emergency
- This occurs very close to the departure date ("at the last minute")

3. Impact on vacation plans: The friend's cancellation is likely to have several effects on the vacation plans: a) Logistical changes:

- Accommodation arrangements may need to be adjusted
- Transportation plans might require modification (e.g., car seating, rental vehicle size)
- Shared costs will need to be recalculated among the remaining friends

4. b) Activity adjustments:

- Planned group activities may need to be altered
- Reservations for restaurants or events might need to be changed

5. c) Emotional impact:

- The group dynamic will be different without the missing friend
- There might be disappointment or concern among the remaining friends

6. Linked events: Several events are interconnected in this scenario:

- The original vacation planning
- The work emergency of the cancelling friend
- The last-minute notification of cancellation
- The necessary adjustments to the vacation plans
- The potential emotional responses of the friend group

7. These events form a chain reaction, with each event influencing the next. The work emergency leads to the cancellation, which in turn necessitates changes to the vacation plans and potentially affects the group's mood or dynamics.

In conclusion, the friend's cancellation has a ripple effect on various aspects of the vacation, from practical considerations to emotional responses. The interconnected nature of these events highlights how a single change can impact a group's plans in multiple ways.

Sub-Benchmark Conditional Causality

Scenario: *A theater group is planning to stage a new play. However, the decision on which venue the play will be performed in has not yet been made.*

Question: What will happen if ticket sales are sufficient, and what will happen if they are not?

ChatGPT 4.o's Response: If ticket sales are sufficient, the theater group will likely proceed with staging the play in the chosen venue, as they will have the necessary financial backing and audience interest to justify the performance. They may also be able to allocate additional resources to production elements like set design or promotion. However, if ticket sales are not sufficient, the group might need to reconsider the venue, downscale the production, or even cancel the play altogether due to a lack of funding or interest. They could also explore alternative strategies to increase sales, such as marketing efforts or discounts.

Claude Sonnet 3.5's Response: To answer this question, we need to consider two scenarios based on ticket sales for the theater group's new play:

1. If ticket sales are sufficient:

The theater group will likely proceed with staging the play as planned. This could involve:

- Securing a venue that matches the expected audience size
- Finalizing production details like set design, costumes, and rehearsal schedules
- Potentially expanding the run of the play if demand is very high
- Ensuring they have enough staff and resources to support the production

2. If ticket sales are not sufficient:

The theater group may need to adjust their plans. Possible outcomes include:

- Downsizing to a smaller, less expensive venue to reduce costs
- Shortening the planned run of the play
- Increasing marketing efforts to boost ticket sales
- Postponing the production to allow more time for promotion and sales
- In extreme cases, canceling the production if financial risks are too high

It's worth noting that without knowing more details about the theater group's specific situation (such as their financial reserves, contractual obligations, or alternative funding sources), these outcomes are general possibilities. The exact course of action would depend on the group's policies, risk tolerance, and overall financial health.

Sub-Benchmark Interactive Causality

Scenario: *A restaurant chain added healthy food options to its menu. This change positively attracted customer interest.*

Question: How would you explain the impact of adding a healthy menu on sales, customer satisfaction, and employee working conditions?

ChatGPT 4.o's Response: Adding healthy food options to a restaurant chain's menu can have a multifaceted impact. First, sales are likely to increase as health-conscious customers are drawn to the new options, expanding the customer base. This could also encourage repeat visits from existing customers seeking healthier choices. Regarding customer satisfaction, offering healthy alternatives can enhance the dining experience by catering to diverse dietary preferences, which could improve overall satisfaction and loyalty. However, the impact on employee working conditions may vary. Employees may need additional training to prepare and serve the new menu items, which could increase workload initially. Over time, though, the healthier menu could lead to a more positive work environment if customers are happier and more engaged. Overall, the addition of healthy menu items can create a positive feedback loop between increased sales, higher customer satisfaction, and potentially improved working conditions, though it may require careful planning and adjustment from the staff.

Claude Sonnet 3.5's Response: To explain the impact of adding a healthy menu on sales, customer satisfaction, and employee working conditions for this restaurant chain, I'll break it down into those three key areas:

1. Sales:
    - Likely increased overall due to attracting new health-conscious customers
    - May have seen a boost in average check size if healthy options were priced at a premium

- Possible decrease in sales of less healthy menu items, but likely offset by new item sales

2. Customer satisfaction:

   - Improved for health-conscious customers who now have more options
   - Potentially increased loyalty from customers who appreciate the expanded menu
   - May have slightly decreased for customers who preferred the old menu exclusively

3. Employee working conditions:

   - Potentially more complex food preparation, requiring additional training
   - Possible increase in job satisfaction for employees who take pride in offering healthier options
   - May lead to busier shifts if the menu change attracts more customers

Benchmark-4: Operational Execution Competence

Sub-Benchmark Ability to Manage Simultaneous Actions

Scenario: *A café employee is handling several customers at once. One customer is ordering coffee, another wants to buy a dessert, and another is trying to pay. The café is busy, and the employee needs to manage all these requests simultaneously.*

Question: How can the employee manage these simultaneous requests? How should they prioritize and handle each customer without causing delays?

ChatGPT 4.o's Response: The employee can manage these simultaneous requests by efficiently prioritizing tasks based on urgency and simplicity. First, they could quickly acknowledge each customer, ensuring everyone feels attended to. The simplest task, such as processing the payment, could be handled first to clear the queue. While the payment is being processed, the employee can prepare the coffee order, as it may take some time. Lastly, they can handle the dessert request, which typically involves less preparation than coffee. Throughout, maintaining clear communication with each customer and staying organized can help avoid delays and ensure efficient service.

Claude Sonnet 3.5's Response: To manage multiple simultaneous requests efficiently in a busy café setting, the employee should follow these strategies:

1. Acknowledge all customers: Quickly make eye contact and greet each customer, letting them know they've been seen. This reassures waiting customers that they'll be served soon.
2. Prioritize based on task complexity:

- Start with the customer paying, as this is usually the quickest transaction.
- Then take the coffee order, as it may need time to prepare.
- Finally, handle the dessert request, which is likely the simplest task.

3. Multitask when possible:

- While the coffee is being prepared, assist the customer buying a dessert.
- If using a point-of-sale system, input the coffee order while talking to the dessert customer.

4. Use clear communication:

- Inform customers of any potential wait times.
- Provide updates on order status when appropriate.

5. Streamline processes:

- Have frequently requested items easily accessible.
- Ensure the payment system is quick and efficient.

6. Stay calm and focused:

- Maintain a steady pace without rushing, which can lead to mistakes.
- Keep a positive attitude to manage customer expectations.

7. Ask for help if available:

- If other staff members are present, request assistance with specific tasks like preparing the coffee or fetching the dessert.

8. Use a queue system if necessary:

- In extremely busy periods, implement a simple numbering system to manage the order of service fairly.

By following these strategies, the employee can efficiently handle multiple requests, minimize delays, and maintain customer satisfaction. The key is to remain organized, communicate clearly, and adapt to the situation as needed.

Benchmark-5: Background Knowledge Integration and Application

Sub-Benchmark Naive Physics

***i.* Gravity and Motion**

Scenario: *A person places a book on the edge of a table, and part of the book hangs off the edge.*

Question: If someone accidentally bumps the table, what will likely happen to the book?

ChatGPT 4.o's Response: If someone accidentally bumps the table, the book will likely fall off the edge, as the part of the book that is hanging off the edge has less support and is more prone to losing balance when the table is disturbed.

Claude Sonnet 3.5's Response: If someone accidentally bumps the table when a book is partially hanging off the edge, the book will likely fall off the table onto the floor.

This outcome is due to several factors:

1. Unstable position: With part of the book already hanging off the edge, its center of gravity is closer to or potentially beyond the table's edge, making it less stable.

2. Reduced friction: Only part of the book is in contact with the table surface, reducing the frictional force holding it in place.
3. Momentum transfer: When the table is bumped, some of that energy will transfer to the book, potentially overcoming the remaining friction and gravity keeping it in place.
4. Gravity: Once the book starts to tip, gravity will accelerate its fall off the table.
5. Lack of restraint: Without anything to stop its movement, the book is free to fall once it starts moving.

The exact result might vary depending on factors like how hard the table is bumped, how much of the book is hanging off, and the book's weight distribution. However, in most cases, a bump to the table would likely cause the precariously positioned book to fall.

***ii.* Commonsense Law of Inertia**

Scenario: *A soccer ball is sitting still on a flat field. No one is near the ball, but ball suddenly moves.*

Question: What will happen to the ball? Why does it behave this way?

ChatGPT 4.o's Response: The ball will continue to move in the direction it started unless acted upon by another force, such as friction, air resistance, or an obstacle. The ball behaves this way due to Newton's first law of motion (inertia), which states that an object at rest stays at rest, and an object in motion stays in motion unless acted upon by an external force. In this case, an unseen force, like wind or a slope on the field, likely caused the ball to move.

Claude Sonnet 3.5's Response: In this scenario, if a stationary soccer ball on a flat field suddenly moves without anyone nearby, there are a few possible outcomes and explanations:

1. The ball will likely roll:

- Once in motion, the ball will continue rolling due to its round shape and the relatively low friction between the ball and the grass.
- It will gradually slow down due to friction and air resistance, eventually coming to a stop if no other forces act upon it.

2. Possible causes of the sudden movement:

   a) Wind:

- A strong gust of wind could overcome the static friction keeping the ball in place.
- The ball's hollow nature makes it lightweight enough to be moved by sufficiently strong wind.

3. b) Uneven ground:

- If the field isn't perfectly flat, a slight slope might cause the ball to start rolling.
- This could happen if the ball was initially balanced precariously and a small disturbance (like a slight ground vibration) caused it to tip.

4. c) Animal interaction:

- A small animal (e.g., a squirrel or bird) not immediately visible could have nudged the ball.

5. d) Remote influence:

- In some organized sports settings, a remote-controlled mechanism within the ball could potentially initiate movement, though this is unlikely in a casual setting.

The ball behaves this way due to basic physics principles:

1. Newton's First Law of Motion: An object at rest stays at rest unless acted upon by an external force. Once that force is applied and overcomes static friction, the ball will move.

2. Conservation of Energy: The initial energy (from wind, slope, etc.) is converted into kinetic energy of the rolling ball, which is then gradually dissipated through friction and air resistance.

3. Rotational motion: The ball's spherical shape allows it to roll efficiently, converting some of its translational kinetic energy into rotational energy.

***iii.* Object Permanence**

Scenario: *A child places a toy behind a curtain where it is no longer visible.*

Question: Does the toy still exist even though it is no longer visible? Why or why not?

ChatGPT 4.o's Response: Yes, the toy still exists even though it is no longer visible. This is due to the concept of object permanence, which is the understanding that objects continue to exist even when they cannot be seen, heard, or touched. Object permanence is a fundamental concept of naive physics, typically developed in infancy, which tells us that the toy is still physically present behind the curtain despite being out of sight.

Claude Sonnet 3.5's Response: Yes, the toy still exists even though it is no longer visible behind the curtain. This concept is known as object permanence. Here's why the toy continues to exist:

1. Physical reality:

- The toy is a physical object that doesn't cease to exist simply because it's out of sight.
- Its material composition and structure remain intact regardless of its visibility.

2. Conservation of matter:

- Matter cannot be created or destroyed under normal conditions, only transformed.
- Hiding the toy doesn't cause its atoms to disappear.

3. Spatial continuity:

- The toy has merely changed location, moving from a visible space to one that's obscured.
- Its existence is independent of our ability to perceive it directly.

4. Cause and effect:

- The child's action of placing the toy behind the curtain doesn't have the power to erase the toy from reality.

5. Verifiability:

- The toy's continued existence can be easily verified by moving the curtain aside.

This scenario touches on an important concept in developmental psychology:

- Object permanence is the understanding that objects continue to exist even when they cannot be observed directly.
- Infants typically develop this concept around 8 months of age.
- Before developing object permanence, very young babies might behave as if an object ceases to exist when it's out of sight.

The toy's continued existence behind the curtain demonstrates that reality persists independently of our direct observation, a fundamental principle in our understanding of the physical world.

***iv.* Locality of Change**

Scenario: *A glass of water is placed on a table, and someone across the room opens a door.*

Question: Will opening the door cause the water in the glass to spill? Explain your reasoning.

ChatGPT 4.o's Response: Opening the door will not directly cause the water in the glass to spill. The act of opening a door from across the room is unlikely to generate enough force or disturbance to affect the stability of the glass or the water inside. For the water to spill, there would need to be a significant force applied to the glass, such as bumping the table or a strong gust of wind from the open door, which is unlikely in typical indoor conditions.

Claude Sonnet 3.5's Response: Opening a door across the room is highly unlikely to cause the water in a glass placed on a table to spill. Here's the reasoning behind this:

1. Distance and energy dissipation:

- The action of opening a door occurs far from the glass of water.
- Any disturbance created by opening the door (like air movement) would dissipate significantly over the distance to the glass.

2. Minimal force transmission:

- Opening a door typically doesn't generate enough force to noticeably shake a room or its contents, especially objects at a distance.
- The table acts as a stable base for the glass, further isolating it from minor vibrations.

3. Surface tension of water:

- Water has surface tension, which helps it maintain its shape in the glass.
- Small disturbances are unlikely to overcome this surface tension and cause spilling.

4. Container stability:

- A typical glass has a wide enough base relative to its height to remain stable under minor environmental changes.

5. Air pressure changes:

- While opening a door can cause slight changes in air pressure, these are usually too minor to affect a glass of water significantly.

6. Threshold for spilling:

- For water to spill, it would need to overcome the glass's rim height.
- This would require a substantial force or movement, far more than typically generated by opening a door at a distance.

Exceptions that could potentially lead to spilling (though still unlikely):

- If the door was slammed extremely hard, causing a shockwave through the building's structure.
- If the glass was already precariously balanced or overfilled.
- If opening the door created an unusually strong draft directly affecting the table.

In everyday circumstances, the act of opening a door across the room would not provide sufficient force or disturbance to cause water in a glass to spill. The physical separation and the minimal energy transfer make such an outcome highly improbable.

Sub-Benchmark Naive Mathematics

***i.* Estimation of Quantities**

Scenario: *A person is hosting a small dinner party and wants to cook enough pasta for everyone. There will be 5 guests, and the person has a 1-kilogram bag of pasta.*

Question: Does the person have enough pasta for the dinner party?

ChatGPT 4.o's Response: In this scenario, the person is hosting a small dinner party for 5 guests, and they have a 1-kilogram bag of pasta. Typically, a serving of pasta for one person is around 100 grams. Given that there are 5 guests, the total amount of pasta needed would be 5 × 100 grams = 500 grams. Since the person has 1 kilogram (1000 grams) of pasta, they have more than enough to serve all their guests. Thus, the person has enough pasta for the dinner party.

Claude Sonnet 3.5's Response: To determine if there's enough pasta for the dinner party, let's break this down step-by-step:

1. Number of people:

- 5 guests

- 1 host
- Total: 6 people

2. Recommended pasta serving size:

- A general rule of thumb is about 2 ounces (56 grams) of dry pasta per person for a main course.
- For 6 people: 6 × 56 grams = 336 grams

3. Amount of pasta available:

- 1 kilogram = 1000 grams

4. Comparison:

- Needed: 336 grams
- Available: 1000 grams

Conclusion: Yes, the person has enough pasta for the dinner party. In fact, they have significantly more than needed. The 1-kilogram bag contains almost three times the amount required for 6 people.

***ii.* Spatial Relationships and Size Estimation**

Scenario: *A person is trying to fit a large couch through a doorway.*

Question: Will the couch fit through the doorway? How can the person make this decision?

ChatGPT 4.o's Response: To determine whether the couch will fit through the doorway, the person needs to measure both the dimensions of the couch (length, width, and height) and the dimensions of the doorway (height and width). The key factors to consider are:

1. Comparing dimensions: The person should check if at least one of the couch's smaller dimensions (width or height) is less than the width of the doorway. If so, the couch can be angled or turned to fit through the door.

2. Rotating or tilting the couch: If the couch is too wide or tall to fit straight through the doorway, the person can tilt or rotate it at an angle to align a smaller dimension with the width of the doorway. This often helps in fitting larger items through narrow spaces.

In summary, the decision is made by comparing the couch's dimensions with the doorway and considering if rotating or angling the couch can help it pass through.

Claude Sonnet 3.5's Response: To determine if a couch will fit through a doorway and make an informed decision, the person should follow these steps:

1. Measure the couch:

- Width (the narrowest dimension)
- Height (from floor to top)
- Depth (from front to back)

2. Measure the doorway:

- Width
- Height

3. Compare measurements:

- The couch's narrowest dimension should be smaller than the doorway's width
- The couch's height should be less than the doorway's height

4. Consider maneuvering space:

- Allow for some extra room to angle the couch
- Account for door frames, hinges, and handles

5. Check for obstacles:

- Look for light fixtures, railings, or other potential obstructions

6. Assess the couch's flexibility:

- Some couches have removable legs or cushions
- Sectionals can often be disassembled

7. Use visualization techniques:

- Create a cardboard cutout of the couch's cross-section
- Use tape to outline the couch's dimensions on the floor

8. Consider professional help:

- Furniture movers have experience with difficult moves
- They may have specialized equipment or techniques

The person can make this decision by:

1. Comparing all measurements
2. Assessing the couch's features (removable parts, flexibility)
3. Evaluating the path to and through the doorway
4. Using visualization aids
5. Deciding if professional help is needed

If the measurements are very close, it's often worth attempting the move carefully, as there may be ways to angle or maneuver the couch that aren't immediately obvious from measurements alone.

iii. **Time Estimation**

Scenario: *A person wants to go to a nearby grocery store but needs to be back home within 1 hour.*

Question: Can the person complete the grocery shopping in time? How should they estimate the total time required?

ChatGPT 4.o's Response: To estimate whether the person can complete the grocery shopping and return home within 1 hour, they need to break down the total time required into the following components:

1. Travel time to and from the grocery store: Estimate how long it takes to walk, drive, or use another form of transportation to reach the store and return. For example, if it takes 10 minutes each way, the travel time is 20 minutes in total.
2. Time spent shopping: Estimate how long the person will need to shop, which depends on the size of the store and the number of items they plan to buy. For a small grocery trip, this might take around 20–30 minutes.
3. Checkout time: Factor in the time spent waiting in line and paying for the groceries, which could add another 5–10 minutes.

By adding up these estimates, the person can assess whether they can complete the trip within the available hour. For instance:

- 20 minutes for travel (10 minutes each way)
- 25 minutes for shopping
- 5 minutes for checkout Total: 50 minutes, which leaves a buffer of 10 minutes.

Thus, the person can decide based on these calculations whether the trip is feasible within the 1-hour time frame.

Claude Sonnet 3.5's Response: To estimate if the person can complete the grocery shopping within 1 hour, they should break down the task into its components and estimate the time for each. Here's how they can approach this:

1. Travel time:
   - Estimate time to get to the store
   - Estimate time to return home
2. Shopping time:
   - Estimate time spent in the store
3. Additional factors:
   - Consider potential delays or variables

Let's break this down further:

1. Travel time:

- To store: Estimate based on distance and mode of transportation
- From store: Usually similar to the time to get there, but factor in carrying groceries

2. Shopping time:

- Consider the number of items needed
- Factor in time for selecting items, waiting in line, and checking out

3. Additional factors:

- Parking time (if driving)
- Unexpected store layout changes or out-of-stock items
- Checkout line length variability

To make a reasonable estimate:

1. Estimate travel time (round trip)
2. Subtract travel time from 1 hour to get available shopping time
3. Assess if the remaining time is sufficient for shopping

For example:

- If travel time is 10 minutes each way (20 minutes total)
- This leaves 40 minutes for actual shopping
- Decide if 40 minutes is enough based on the shopping list and store familiarity

Tips for staying within the time limit:

1. Prepare a shopping list in advance
2. Organize the list by store layout if familiar

3. Shop during less busy hours if possible
4. Use self-checkout if it's typically faster

The person can complete the shopping in time if their total estimated time (travel + shopping) is less than 1 hour. If it's close, they should be prepared to be efficient or potentially skip some items if necessary.

Sub-Benchmark Naive Psychology

i. **Emotional Recognition**

Scenario: *Anna is sitting in a meeting. She is smiling and nodding as her colleague presents an idea.*

Question: What might Anna be feeling, and why?

ChatGPT 4.o's Response: Anna might be feeling engaged or supportive as she is smiling and nodding during her colleague's presentation. This body language often suggests she is interested in or agreeing with the idea being presented. Alternatively, she could be displaying polite social behavior, regardless of her true feelings, as people sometimes smile and nod to appear attentive or courteous in professional settings.

Claude Sonnet 3.5's Response: Based on the scenario described, here are some possibilities for what Anna might be feeling and why:

1. Interest and engagement: Anna's smiling and nodding could indicate she finds her colleague's idea intriguing or valuable. She may be genuinely interested in the presentation and showing her attentiveness.
2. Encouragement and support: Anna might be feeling supportive towards her colleague, using her body language to provide positive reinforcement and boost their confidence during the presentation.
3. Agreement: Her nodding could suggest that she agrees with the points being made, feeling alignment with the ideas presented.
4. Professional courtesy: Anna may be exhibiting polite behavior expected in a meeting setting, regardless of her true feelings about the idea.

5. Excitement: If the idea presented aligns with her own thoughts or solves a problem, Anna might be feeling excited about the potential outcomes.
6. Relief: If the idea addresses a challenging issue, Anna could be feeling relieved that a potential solution is being proposed.
7. Anticipation: Anna might be looking forward to discussing or implementing the idea, feeling a sense of anticipation for next steps.
8. Masked disagreement or skepticism: In some cases, people might smile and nod to maintain a positive atmosphere even if they have reservations. Anna could potentially be feeling skeptical but choosing not to express it openly.

It's important to note that without more context or insight into Anna's thoughts, we can only speculate about her true feelings based on the observable behavior. Body language can be complex and influenced by various factors, including personal disposition, cultural norms, and workplace dynamics.

***ii.* Theory of Mind**

Scenario: *"Ibrahim is sitting on a bench in the park, smoking a cigarette. In the distance, a woman holding an unlit cigarette starts walking toward him. As the woman approaches Ibrahim, she briefly glances at his lit cigarette."*

Questions: What could be the woman's possible intention for approaching Ibrahim?

ChatGPT 4.o's Response:

The woman's possible intention for approaching Ibrahim might include:

1. Requesting a Light: Given that she holds an unlit cigarette, it's likely she intends to ask Ibrahim for a lighter or match.
2. Seeking Conversation: She might be looking to strike up a conversation with him, possibly about smoking or something casual related to their shared location in the park.

3. Requesting Assistance: It's also possible she's approaching him for some other kind of help or information, though this is less directly related to the cigarette in her hand.

These possibilities take into account common social cues and intentions related to shared habits or needs in public settings.

Claude Sonnet 3.5's Response:

Let me analyze the possible intentions of the woman based on the scenario's details:

1. Most likely, she's looking to ask Ibrahim for a light for her unlit cigarette. This is suggested by:

- She's holding an unlit cigarette
- She specifically walks toward someone who has a lit cigarette
- She glances at his lit cigarette, indicating interest in it

2. Alternative possibilities, though less likely, could include:

- Looking to join another smoker for company (as smokers sometimes congregate together)
- Planning to ask about smoking restrictions in the park
- Simply walking in that direction coincidentally while preparing to smoke

The most straightforward and probable interpretation, given all the contextual clues, is that she needs a light for her cigarette and is approaching Ibrahim to ask for one. This is a common social interaction among smokers, where someone without a lighter might approach someone with a lit cigarette to ask for help lighting their own.

iii. **Social Norm Adherence**

Scenario: *"In a workplace meeting room, everyone has taken their seats. The meeting cannot begin until the door is closed."*

Question: In this situation, who is likely to take responsibility for closing the door?

ChatGPT 4.o's Response:

In a workplace meeting room scenario, the responsibility for closing the door would likely fall on someone who has a position of authority or a designated role in the meeting, such as the meeting organizer, team leader, or host. However, if no one assumes responsibility, the nearest person to the door might take on the task out of convenience or courtesy. This action often depends on workplace norms, where those involved typically wait for the person leading the meeting to signal or delegate such a small but necessary action.

Claude Sonnet 3.5's Response:

This is an interesting social psychology question that touches on diffusion of responsibility and workplace dynamics. Let me think through this systematically:

Most likely outcomes, from most to least probable:

1. The last person who entered the room

- This person is naturally positioned near the door
- They're most consciously aware of the door's state since they just used it
- Social expectation often falls to them as they "caused" the door to be open

2. The meeting organizer/leader

- They feel ownership of the meeting's success
- They're in a position of responsibility
- They may want to signal the meeting's formal start
- They're most aware of the requirement for the door to be closed

3. The person sitting closest to the door

- Minimal effort required
- May feel implicit social pressure due to proximity

Least likely:

- People in the middle or far side of the room
- Multiple people simultaneously (due to diffusion of responsibility)

This scenario is actually a classic example of the bystander effect in a mild form - everyone knows the door needs to be closed, but without clear delegation, there can be a moment of social uncertainty about who should act.

Benchmark-6: Accuracy and Relevance Management

Sub-Benchmark Filtering Out Irrelevant Information

Scenario: *A person needs to prepare a resume for the position of computer engineer. They have information on their educational background, previous work experience, hobbies, and personal details.*

My name is Alan Turing. My mother, Sara, became pregnant with me in the town of Chatrapur in Orissa, India. My father, Julius Mathison Turing, was a civil servant in the British Indian colonial administration. Julius and my mother wanted me to be born in England, so they traveled to London and settled in a house in Maida Vale (which is now the Colonnade Hotel), where I was born on June 23, 1912. I had an older brother named John. My father continued working as a civil servant in India, and throughout my childhood, my family traveled back and forth between Guildford, England, and India, leaving me and my brother with family friends in Hastings, England. From a young age, I displayed signs of genius and continued to demonstrate them.

When I was 6 years old, my family enrolled me in St. Michaels, a day school. My teachers and the headmaster quickly recognized my intelligence. In 1926, at the age of 14, I entered Sherborne School, a famous and expensive private school in Dorset. The first day of the school term coincided with the General Strike in England; however, I was so eager to attend school that I cycled more than 60 miles from Southampton to Sherborne, even though no trains were running, and spent the night halfway at a hotel.

My natural inclination towards mathematics and science did not win the favor of my teachers at Sherborne, where the education was more focused on classical Greek and Latin. The headmaster wrote to my parents, saying, "I hope he doesn't end up ignorant in the middle of two schools. If he stays at a private school, he must accept its education; if he only wants to be a scientist, he is wasting his time here." Despite this, I continued to show remarkable talent in the subjects I enjoyed, solving advanced mathematical problems even before learning about calculus and integration. By 1928, at the age of 16, I had discovered the work of Albert Einstein, not only understanding it but also deducing Einstein's criticisms of Newton's laws of motion independently, without the aid of textbooks.

At school, I formed a close friendship and romantic relationship with a slightly older student named Christopher Morcom. However, Christopher had contracted tuberculosis from drinking infected cow's milk as a child, and he died just a few weeks after the end of our last term at Sherborne. His death shattered my religious faith, and I became an atheist. I adopted the belief that all phenomena in the world, including the workings of the human brain, could be explained through materialistic principles.

University and work on computability

From 1931 to 1934, I attended King's College, Cambridge, where I earned my degree with distinction and became a Fellow of the College in 1935, thanks to my dissertation on the central limit theorem. I had originally hoped to study at Trinity College, but my lack of interest in classical studies, such as Greek and Latin, prevented me from winning a scholarship there. Instead, I attended King's College, where I pursued my passion for mathematics and science.

In my seminal paper, On Computable Numbers, with an Application to the Entscheidungsproblem, published on May 28, 1936, I reformulated Kurt Gödel's 1931 results on the limits of computation and proof, replacing Gödel's formal arithmetic-based language with what are now known as "Turing machines." I proved that any mathematically solvable problem that can be represented by an algorithm can be solved by such a machine.

Turing machines are now a central object of study in computing theory. I showed that the halting problem for Turing machines is undecidable, demonstrating that no general algorithmic solution exists for the Entscheidungsproblem. Despite being published after Alonzo Church's similar results, based on his lambda calculus, my work was more intuitive and accepted. A key contribution of my theory was the concept of a

"Universal Turing Machine," a machine capable of performing the tasks of any other machine. The paper also introduced the notion of computable numbers.

From September 1936 to July 1938, I worked at Princeton University's Institute for Advanced Study under Alonzo Church, where I engaged in abstract mathematical research and cryptology. I also completed three stages of a four-stage electro-mechanical binary multiplier machine. I was awarded a Ph.D. from Princeton in June 1938 for my dissertation, which explored computability with oracle machines, a type of Turing machine that could address problems unsolvable by standard machines.

After returning to Cambridge, I attended lectures by Ludwig Wittgenstein on the foundations of mathematics. Wittgenstein and I disagreed, engaging in debates. While I defended formalism, Wittgenstein argued that mathematics did not discover existing truths but invented new ones. During this time, I also worked part-time at the Government Code and Cypher School (GCCS). I am currently seeking employment.

Question: Which information should be included in the resume? Create a resume.

ChatGPT 4.o's Response: Here's a streamlined resume based on the information provided, focusing on Alan Turing's background for a computer engineer position:

Alan Turing
Birthdate: June 23, 1912
Location: London, England
Email: [example@example.com]
Phone: [Your contact number]

Objective

A visionary computer scientist with a Ph.D. from Princeton University, seeking a position as a computer engineer. Specializing in computational theory, cryptography, and the development of early computers. Proven track record in groundbreaking research on computability and algorithm design, including the invention of the Turing machine.

Education

Ph.D. in Mathematics – Princeton University (1938)

- Dissertation: *Computability with Oracle Machines*
- Focus: Computation theory, cryptology, and machine development

B.A. in Mathematics – King's College, Cambridge (1934)

- Fellow of King’s College (1935)
- Distinction in mathematics and computation theory

Work Experience

Researcher – Princeton University (1936-1938)

- Worked under Alonzo Church on advanced mathematical and cryptological research
- Developed three stages of an electro-mechanical binary multiplier machine

Part-time Cryptologist – Government Code and Cypher School (1938-Current)

- Engaged in cryptography and encryption research

Key Contributions

- Turing Machines: Introduced the concept of a machine capable of solving any computable problem, central to modern computing theory
- Universal Turing Machine: Developed the theoretical foundation for general-purpose computers

- Halting Problem: Proved undecidability in algorithmic computation
- On Computable Numbers (1936): Groundbreaking paper on algorithm theory and decision problems

---

Skills

- Advanced computation theory
- Algorithm design and optimization
- Machine development and implementation
- Cryptology and encryption techniques
- Strong formalism and mathematical logic skills

Languages

- English (Native)
- Basic Latin and Classical Greek (educational background)

Hobbies and Interests

- Exploring materialistic principles in science and human cognition
- Debating philosophical foundations of mathematics

This resume focuses on the technical and academic qualifications that would make Alan Turing a strong candidate for a computer engineer role.

Claude Sonnet 3.5's Response: Thank you for providing that detailed information, Alan. I understand you're looking to create a resume for a computer engineering position. Given your extraordinary background and accomplishments, I'll help you craft a concise, professional resume that highlights your most relevant qualifications and experience. Let's focus on the key elements that would be most impressive to potential employers in the field of computer engineering.

Alan Turing

[Your Contact Information]

Summary

Pioneering computer scientist and mathematician with groundbreaking work in computability theory and early computer design. Inventor of the Turing machine concept and significant contributor to the foundations of artificial intelligence and computer science.

Education

- Ph.D. in Mathematics, Princeton University, 1938
  - Dissertation: Systems of Logic Based on Ordinals
- B.A. in Mathematics (with distinction), King's College, Cambridge University, 1934

Research Experience

Institute for Advanced Study, Princeton University

Research Fellow, 1936-1938

- Conducted advanced research in abstract mathematics and cryptology
- Developed a four-stage electro-mechanical binary multiplier machine
- Explored computability with oracle machines, extending the concept of Turing machines

King's College, Cambridge University

Fellow, 1935-1936

- Authored seminal paper "On Computable Numbers, with an Application to the Entscheidungsproblem"
- Introduced the concept of the Turing machine and universal computation
- Proved the undecidability of the halting problem

Key Achievements

- Developed the concept of the Turing machine, a fundamental model of computation
- Proved the undecidability of the Entscheidungsproblem, a significant result in computability theory
- Contributed to the theoretical foundations of computer science and artificial intelligence
- Demonstrated exceptional problem-solving skills from an early age, including understanding and expanding upon Einstein's work at age 16

Skills

- Advanced mathematical analysis
- Algorithm design and analysis
- Theoretical computer science
- Cryptology
- Machine design and engineering

Languages

- English (native)
- Classical Greek and Latin (academic knowledge)

References

Available upon request

I've created a professional resume tailored for a computer engineering position, highlighting your most relevant qualifications and achievements. This resume focuses on your groundbreaking work in computability theory, your education, and your key contributions to the field of computer science.

Here are a few notes about the resume:

1. I've included a brief summary that emphasizes your pioneering work in computer science and mathematics.
2. Your education section highlights your advanced degrees from prestigious institutions.
3. The research experience section focuses on your work at Princeton and Cambridge, which is most relevant to computer engineering.
4. I've included a "Key Achievements" section to highlight your most significant contributions to the field.
5. The skills section emphasizes abilities that would be valuable in a computer engineering role.
6. I've kept personal details to a minimum, as is standard in modern resumes, but included language skills which may be relevant.

Is there anything you'd like me to modify or expand upon in this resume? I can make adjustments to better suit your preferences or the specific computer engineering position you're applying for.

Sub-Benchmark Detection of Incorrect Information

Scenario: *Two texts from Wikipedia and mayoclinic.org regarding COVID-19 will be shared with ChatGPT 4.o and Claude Sonnet 3.5. Some of the information in Wikipedia's text will be altered to make it incorrect. Both machines will be asked to identify which text contains the correct information and to compare the reliability of both texts. The texts are as follows:*

https://en.wikipedia.org/wiki/COVID-19_pandemic

Symptoms of COVID-19 typically appear 25 to 30 days after exposure. The symptoms can range from mild to severe illness. Common symptoms include rash, loss of smell and taste, nasal congestion and runny nose, cough, muscle pain, sore throat, fever, diarrhea, and shortness of breath. People with the same infection may experience different symptoms, and these may change over time. Three common clusters of symptoms have been identified: respiratory symptoms, including cough, sputum, shortness of breath, and fever; musculoskeletal symptoms, including muscle and joint pain, headache, and fatigue; digestive symptoms, including abdominal pain, vomiting, and diarrhea. In individuals without prior ear, nose, and throat conditions, loss of taste and smell is associated with COVID-19 and reported in up to 88% of cases. Most people experience mild symptoms, but COVID-19 can lead to serious medical complications, although it generally does not result in death unless the person has underlying chronic conditions. The most at-risk groups are infants and adults.

https://www.mayoclinic.org/diseases-conditions/coronavirus/symptoms-causes/syc-20479963

Typical symptoms of COVID-19 usually appear 2 to 14 days after exposure to the virus.

Symptoms may include:

- Dry cough.
- Shortness of breath.
- Loss of taste or smell.
- Extreme tiredness, known as fatigue.
- Digestive issues, such as nausea, vomiting, or diarrhea.
- Pains, such as headaches and body or muscle aches.
- Fever or chills.
- Cold-like symptoms, including congestion, runny nose, or sore throat.

Some people may exhibit few symptoms or none at all. Those who test positive for COVID-19 without symptoms are known as **asymptomatic**. For instance, many children who test positive do not show symptoms of COVID-19 illness. People who develop symptoms later are considered **presymptomatic**. Both groups can still spread the virus to others.

In some cases, symptoms worsen about 7 to 14 days after they first appear.

Most people experience mild to moderate symptoms. However, COVID-19 can cause serious complications and may result in death. Older adults or those with existing medical conditions are at higher risk of severe illness.

COVID-19 can manifest as mild, moderate, severe, or critical illness:

- **Mild COVID-19** generally does not affect the lungs' ability to supply oxygen to the body.
- In **moderate COVID-19**, the lungs work correctly, but signs of deep lung infection are present.
- **Severe COVID-19** indicates that the lungs are not functioning properly, requiring oxygen and hospitalization.
- **Critical COVID-19** means the respiratory system has failed, and damage occurs throughout the body.

Question:
"Can you identify the incorrect information between these two news reports? Which one is based on accurate data?"

ChatGPT 4.o's Response: Upon comparing the two reports, I can identify the following points:

1. Symptoms onset time:

- The Wikipedia article mentions that symptoms of COVID-19 appear "25 to 30 days after exposure," which is incorrect. COVID-19 symptoms typically appear 2 to 14 days after exposure, as stated in the Mayo Clinic article, which is consistent with official health guidelines.

2. Risk groups:

- The Wikipedia article mentions that infants and adults are the most at-risk groups. This is misleading. While severe COVID-19 can affect people of all ages, it primarily poses a greater risk to older adults and those with underlying health conditions, as highlighted correctly in the Mayo Clinic article.

3. Loss of smell and taste:

- The Wikipedia article claims that loss of taste and smell is reported in up to 88% of cases, which is an exaggerated figure. While loss of taste and smell is common, it's not as high as 88%. The Mayo Clinic report accurately lists it as a symptom without overstating its prevalence.

In conclusion, the Mayo Clinic article is based on accurate data, whereas the Wikipedia article contains incorrect and exaggerated information.

Claude Sonnet 3.5's Response: I'd be happy to compare the information in these two sources and identify any discrepancies or inaccuracies. Let's break it down:

1. Incubation period:

- Wikipedia source: 25 to 30 days
- Mayo Clinic source: 2 to 14 days

The Mayo Clinic information is correct here. The typical incubation period for COVID-19 is indeed 2 to 14 days, not 25 to 30 days as stated in the Wikipedia excerpt.

2. Symptoms: Both sources list similar symptoms, including cough, fever, loss of taste and smell, fatigue, muscle pain, and digestive issues. However, there are some differences:

- The Wikipedia source mentions a rash, which is not typically listed as a common symptom.

- The Mayo Clinic source provides a more comprehensive and accurate list of common symptoms.

3. Risk groups:

- Wikipedia source: infants and adults are most at risk

- Mayo Clinic source: older adults and those with existing medical conditions are at higher risk

The Mayo Clinic information is more accurate. While COVID-19 can affect people of all ages, older adults and those with underlying health conditions are generally considered to be at higher risk for severe illness.

4. Severity of illness: Both sources acknowledge that COVID-19 can range from mild to severe, but the Mayo Clinic provides a more detailed breakdown of illness severity (mild, moderate, severe, and critical).
5. Asymptomatic and presymptomatic cases: The Mayo Clinic source provides information on asymptomatic and presymptomatic cases, which is important and accurate information not mentioned in the Wikipedia excerpt.

In conclusion, the Mayo Clinic source (https://www.mayoclinic.org/diseases-conditions/coronavirus/symptoms-causes/syc-20479963) is based on more accurate data and provides a more comprehensive and up-to-date overview of COVID-19 symptoms and characteristics. The Wikipedia excerpt contains some significant inaccuracies, particularly regarding the incubation period and risk groups.

It's important to note that information about COVID-19 has evolved over time as more research has been conducted. Always refer to reputable health organizations like the WHO, CDC, or respected medical institutions for the most current and accurate information.

Sub-Benchmark Prioritization of Information

Scenario: *A healthcare worker is evaluating information from three different patients in an emergency situation. The patient files are as follows:*

Patient 1:

- Name: Mehmet Yilmaz
- Age: 45
- Diagnosis: Suspected heart attack
- Symptoms: Severe chest pain, shortness of breath, sweating, numbness in left arm
- Vital Signs:
    - Blood Pressure: 85/50 mmHg
    - Heart Rate: 110 bpm
    - Oxygen Saturation: 88%

Patient 2:

- Name: Ayse Demir
- Age: 30
- Condition: Minor injury
- Diagnosis: Mild head trauma
- Symptoms: Headache, mild dizziness, nausea
- Vital Signs:
    - Blood Pressure: 120/80 mmHg
    - Heart Rate: 80 bpm
    - Oxygen Saturation: 98%

Patient 3:

- Name: Ahmet Kaya
- Age: 25
- Condition: Minor injury

- Diagnosis: Laceration on left hand (cut)
- Symptoms: Bleeding on hand, mild pain
- Vital Signs:
    - Blood Pressure: 125/85 mmHg
    - Heart Rate: 75 bpm
    - Oxygen Saturation: 99%

Question: Which patient's condition should be prioritized and why?

ChatGPT 4.o's Response: The healthcare worker should prioritize Patient 1, Mehmet Yilmaz, who is suspected of having a heart attack. This patient's symptoms, such as severe chest pain, shortness of breath, sweating, and numbness in the left arm, combined with the vital signs of low blood pressure (85/50 mmHg), elevated heart rate (110 bpm), and low oxygen saturation (88%), indicate a potentially life-threatening condition. Immediate medical attention is required to prevent further complications or death.

Patient 2, Ayse Demir, has a mild head trauma with stable vital signs (normal blood pressure, heart rate, and oxygen saturation), and Patient 3, Ahmet Kaya, has a laceration on his hand with no signs of life-threatening issues. Both of these patients are in a stable condition and can be evaluated after attending to Patient 1.

Claude Sonnet 3.5's Response: Based on the information provided for the three patients, Patient 1 (Mehmet Yilmaz) should be prioritized for immediate medical attention. Here's why:

1. Suspected diagnosis: Patient 1 is suspected to be having a heart attack, which is a life-threatening emergency requiring immediate intervention.
2. Symptoms: The patient is experiencing severe chest pain, shortness of breath, sweating, and numbness in the left arm. These are classic symptoms of a myocardial infarction (heart attack).
3. Vital signs:

- Blood pressure of 85/50 mmHg is dangerously low, indicating possible cardiogenic shock.
- Heart rate of 110 bpm suggests tachycardia, which is concerning in this context.
- Oxygen saturation of 88% is critically low, indicating the patient is not getting enough oxygen.

In comparison:

- Patient 2 (Ayse Demir) has mild head trauma with stable vital signs. While this requires medical attention, it's not immediately life-threatening.
- Patient 3 (Ahmet Kaya) has a laceration on the left hand, which is a minor injury with stable vital signs. This can be treated after more urgent cases.

In an emergency triage situation, Patient 1 would be classified as a "red" or highest priority case requiring immediate, life-saving interventions. Patients 2 and 3 would likely be classified as "yellow" or "green" respectively, indicating they can safely wait for treatment while the most critical patient is attended to first.

B. TURKISH SUMMARY / TÜRKÇE ÖZET

Sağduyu, herkes tarafından bilinen ve kabul edilen, gündelik meseleleri anlamaya ve yorumlamaya yardımcı olan tüm sezgisel inançlara, varsayımlara ve akıl yürütme yeteneklerine verilen isimdir (Mueller, 2014). Bilinçli bir yönteme dayanmayan sağduyu bilgisi, duyuların kullanılması ve en ilkel türdeki deneyimler sonucunda ortaya çıkmıştır. Bu sebeplerle, bilimsel bilgi ve sağduyu bilgisi oldukça farklı özellikler taşır. Bilimsel bilgi, belirli yöntemler ve araçlar kullanılarak elde edilir. Ayrıca nesnel, sistematik, tutarlı ve eleştiriye açıktır (McCarthy, 1981). Ancak sağduyu bilgisi, deneyim temellidir ve kişisel deneyimlerin tekrarına bağlıdır.

Ortak özellikleri bulunmasına rağmen, sağduyu bilgisi ile sağduyulu akıl yürütme becerisi arasında da önemli farklılıklar vardır. McCarthy (1984), sağduyu bilgisini "herkesin bildiği şey" olarak, sağduyulu akıl yürütmeyi ise "sağduyu bilgisini kullanma yeteneği" olarak tanımlar. Sağduyulu akıl yürütme, belirli durumları anlamlandırmak ve çıkarımlar yapmak için sağduyu bilgisini kullanan bilişsel bir süreçtir. Sağduyulu akıl yürütme, bu bilgilere dayanarak akıl yürütmek için gerekli bilişsel süreçleri kapsarken, sağduyu bilgisi içeriği sağlar. Sağduyulu akıl yürütme, bireylerin sezgisel bilgiyi kullanarak gündelik durumları yorumlamasına yardımcı olur; hızlı, pratik ve bağlama uygun kararlar alınmasını sağlar. Bunun ötesinde, geçmiş deneyimlere dayanarak sonuç çıkarabilmeyi, olası sonuçları öngörmeyi hem fiziksel hem de sosyal dünyadaki neden-sonuç ilişkilerini anlamayı mümkün kılar. Örneğin, bir nesnenin hareketini anlamak için, önce nesnenin varlığı, mekânsal konumu ve hareket edebileceği fikri gibi temel kavramları kavramayı gerekir. Bu bilgilerin tamamı sağduyu bilgisiyken, bu bilgileri birbiriyle ilişkilendirme süreci ise sağduyulu akıl yürütmedir. Ayrıca, sağduyulu akıl yürütme becerisi fiziksel dünyayı anlamamızda hayati bir rol oynar. Örneğin, sıcak bir bardağa dokunmanın elimizi yakabileceğini ya da ağır bir nesneyi kaldırmanın hafif bir nesneyi kaldırmaktan daha zor olduğunu biliriz. Benzer şekilde, bir topu havaya attığımızda yere düşeceğini öngörebiliriz, çünkü yerçekimi gibi fiziksel yasalar, sağduyunun bir parçası olarak evrensel bir şekilde anlaşılır. Sağduyu, yalnızca fiziksel dünyayı anlamamıza yardımcı olmakla kalmaz, aynı zamanda sosyal etkileşimlerimizi de yönlendirir. Örneğin, bir restoranda yemek siparişi verdikten sonra beklemek gerektiğini ya da bir konuşma sırasında göz teması kurmanın nezaket göstergesi olduğunu bilmek, bu tür bilginin örnekleridir. Sağduyu akıl yürütmesi, olası sonuçları değerlendirerek hızlı ve sezgisel kararlar alınmasını destekler ve konuşmalardaki ima edilen anlamları yorumlayarak yanıtlarımızı buna göre uyarlamamıza yardımcı olur. Örneğin, "Bu iş tam bir karmaşa" dediğimizde, kelimenin tam anlamıyla bir karmaşadan değil, karmaşık veya zor bir durumdan bahsettiğimizi anlarız. Benzer şekilde, "Bulutların üstünde" dediğimizde, kişinin kelimenin tam anlamıyla bulutlarda değil, çok mutlu olduğunu mecazi olarak ifade ettiğini biliriz. Ayrıca, sağduyulu akıl yürütme, yeni durumları benzer geçmiş deneyimlerle ilişkilendirerek öğrenmeyi ve uyum sağlamayı teşvik eder, bu da değişen çevrelere etkili bir şekilde yanıt vermemizi sağlar. Sağduyu bilgisi, sosyal normlar ve kültürel deneyimlerle de şekillenir. Örneğin, toplu taşıma araçlarında yaşlılara yer vermek saygılı bir davranış olarak kabul edilirken, Türkiye'de misafire çay ikram etmek dostane bir jest olarak görülür.

Tüm bu bilgiler ışığında, sağduyulu akıl yürütmenin bir tanımını yapmak mümkündür. Sağduyulu akıl yürütme, sıradan, günlük ve deneyim temelli bilgileri kullanarak gündelik hedeflere etkili bir şekilde ulaşma yeteneğidir (Brachman & Levesque, 2022). Bu, bilgileri işleme ve mevcut durum için çıkarımları hızlı ve zahmetsiz bir şekilde uygulama sürecini içerir. Sağduyu, özel bir eğitim ya da gelişmiş analitik yetenekler gerektirmeden, günlük yaşamda sıkça karşılaşılan sorunlarla ilgilenir ve kişisel deneyimlere dayanır. Entelektüel bir çaba ya da felsefi bir sorgulama olmaksızın, gerçek dünyada başarıya ulaşmayı sağlayacak kararları almaya odaklanır.

Sağduyulu akıl yürütme, Yapay Zeka (YZ) için de kritik bir öneme sahiptir, çünkü YZ sistemlerinin gerçek dünya durumlarında insanlarla etkili bir şekilde ilişki kurmasını ve karmaşık koşullara uyum sağlamasını mümkün kılar. YZ sistemleri, sağduyulu akıl yürütme becerisini kullanarak yüzeysel kalıpların ötesine geçebilir ve duruma daha uygun yargılara ulaşabilir. Sağduyulu akıl yürütme, YZ'nin daha yüksek düzeyde bilimsel bilgi işleme ve akıl yürütme kapasitelerine ulaşması için de temel bir çerçeve sağlar (Brachman & Levesque, 2022). Ayrıca, sağduyulu akıl yürütme becerisini taklit eden YZ sistemleri, insan niyetlerini, tercihlerini ve sosyal etkileşimlerini de daha iyi anlayacaktır. Bu, belirsiz dili yorumlamalarını, ima edilen anlamları kavramalarını ve insan sorularına ya da komutlarına akıllıca yanıt vermelerini sağlayabilir. Bağlamı anlama yeteneğine sahip olabilen YZ sistemleri yalnızca basit komutları yerine getirmekle kalmaz, aynı zamanda karmaşık görevlerde de başarılı olabilir. Bu durum, kullanıcı deneyimini geliştirir ve daha etkili ve doğal bir insan-YZ etkileşimini mümkün kılar (Lake et al., 2016). Ayrıca, YZ sistemlerinin güvenliği ve güvenilirliğini artırmak, sağduyulu akıl yürütme becerisini taklit etme çabasının bir diğer önemli nedenidir. Sağduyulu akıl yürütme, YZ'nin eylemlerin sonuçlarını öngörmesini, olası riskleri belirlemesini ve tehlikeli ya da mantıksız davranışlardan kaçınmasını sağlar (Lake et al., 2016). Bu, özellikle otonom araçlar ya da tıbbi teşhis gibi güvenliğin kritik olduğu alanlarda hayati bir öneme sahiptir.

Günümüzde birçok yapay zeka araştırmacısı, sağduyulu akıl yürütmeyi taklit etmeye çalışmaktadır, çünkü bu yetenek hâlâ YZ'nin yeteneklerindeki kritik, eksik bir parça olarak kabul edilmektedir. YZ, görüntü tanıma ve doğal dil işleme gibi belirli alanlarda önemli ilerlemeler kaydetmiş olsa da dünyayı insanlar gibi anlama ve akıl yürütme konusunda genellikle zorlanmaktadır. Aslında, sağduyulu akıl yürütmeyi taklit etme sorunu, yapay zeka alanının kuruluşundan bu yana süregelen bir mesele olmuştur. John McCarthy, 1950'lerde "Yapay Zeka" (Artificial Intelligence) terimini kullanmış ve 1959 yılında yapay zeka üzerine ilk çalışmalardan biri olan Programs with Common Sense makalesini yayımlamıştır. Ancak, bu uzun geçmişe rağmen, sağduyulu akıl yürütmeyi geliştirmek beklenenden çok daha zor bir süreç olmuştur.

Gary Marcus ve Ernest Davis'in (2019) yazdığı Rebooting AI kitabı, makinelerin sağduyulu düşünme yeteneğine sahip olma gerekliliğinin bir ihtiyaç olduğunu vurgulamakta, ancak şu ana kadar etkili çözümlerin bulunamadığını ifade etmektedir. Marcus ve Davis, sağduyunun zekânın merkezi bir parçası olduğunu, ikincil bir unsur olmadığını savunmuşlardır. Ayrıca, McCarthy ve Hayes'in (1969) ve daha sonra tekrardan McCarthy'nin (1984) dile getirdiği önemli konular bugün hâlâ güncelliğini korumaktadır. Bu konular arasında zaman ve mekân, nedensellik, kuvvet, maddeler, enerji, sürekli değişim ve nicelikler yer alır. Fiziksel dünyadaki değişimleri, eylemleri ve neden-sonuç ilişkilerini anlamak—genellikle naive fizik ve naive matematik olarak adlandırılır—sağduyu bilgisinin merkezinde yer alır.

Sağduyu bilgisinin kapsamı, insan faaliyetlerinin diğer insanlarla ilişkiler içerdiği göz önüne alındığında, hedefler, inançlar ve arzular gibi naive ya da halk psikolojisi fikirlerini de kapsayacak şekilde genişletilebilir. YZ sistemleri, yalnızca veri odaklı tekniklerin sınırlamalarını aşmayı amaçlamaktadır. Bu çabalar, yapay zekanın daha insana benzer bir şekilde akıl yürütmesini sağlamak için atılmış önemli adımlardır.

Sağduyulu akıl yürütmeyi taklit etmek, yapay zeka araştırmalarında önemli bir zorluk olmaya devam etmektedir, çünkü günlük bilgileri ve sezgisel anlayışı makinelere aktarmak son derece karmaşıktır. İnsanlar, bilinçli bir çaba sarf etmeden günlük yaşamlarında geniş bir örtük bilgi ve bağlamsal anlama kapasitesine sahiptir. Ancak, bu bilgiyi bir YZ sistemine öğretmek ve sistemin anlayabileceği ve etkili bir şekilde kullanabileceği bir formatta düzenlemek oldukça zordur (Mueller, 2014). Sağduyulu akıl yürütmeyi taklit etmek, yalnızca verilerle çalışmanın ötesine geçer; bir YZ sisteminin belirsizlikle başa çıkma, eksik bilgiyle mantıksal çıkarımlar yapma, nedensel ilişkileri anlama ve bağlamı kavrayarak uygun şekilde yanıt verme yeteneği geliştirmesini gerektirir. Bu, mevcut veri odaklı YZ yöntemlerini aşan bilgi temsili ve akıl yürütme tekniklerinin geliştirilmesini zorunlu kılar.

Tez iki ana bölümden meydana gelmektedir. İlk bölümde, sağduyulu akıl yürütmeyi taklit etme sürecinde karşılaşılan üç temel zorluk incelenmiştir. Sağduyu bilgisinin temsil edilmesi, örtük bilginin tanımlanması ve çerçeve probleminin ele alınması. Bu zorlukları ele almamdaki amaç, bu alandaki ilerlemeyi engelleyen anahtar engelleri ortaya koymaktır. Bu üç temel zorluk analiz edilerek, sağduyulu akıl yürütmenin YZ araştırmalarında neden hala öncelikli bir sorun olmaya devam ettiği açıklanmış ve iyileştirme gereken alanlar belirlenmeye çalışılmıştır.

İlk olarak, yapay zekâda bilgi temsili (Knowledge Representation) geliştirilmesi ve karşılaşılan zorluklar, özellikle sağduyulu bilginin temsil edilmesi bağlamında ele alınmıştır. Bilgi temsilinin tarihsel kökenleri, Leibniz ve Frege gibi düşünürlerle başlayan felsefi temellerinden, modern hesaplama temsili yöntemlerine kadar geçen sürç özetlenmişti. Alandaki önemli figürler ve kavramlar, örneğin McCarthy’nin

(1989) sağduyuyu hesaplamalı akıl yürütme için sembolik olarak temsil etmeye yönelik öncü çalışması ve ardından gelen gelişmeler, detaylı bir şekilde ele alınmıştır.

Ayrıca, sağduyulu bilginin temsil edilmesine yönelik iki temel yaklaşım; kural tabanlı sistemler ve modern YZ teknikleri ayrıntılı olarak incelenmiştir. Sinir ağları ve büyük dil modelleri (LLM) bu yaklaşımlara örnek olarak verilmiştir. Kural tabanlı yöntemler, bilgiyi organize etmek için önceden tanımlanmış kurallar ve mantıksal çerçevelere dayanırken, sinir ağları ve büyük dil modelleri geniş veri setlerinden öğrenme yetenekleriyle öne çıkar. Her yaklaşımın bilgi temsili açısından sunduğu güçlü ve zayıf yönleri ele alınmış ve sağduyulu akıl yürütmenin karmaşıklıklarını ele almak için hangi yöntemlerin daha etkili olabileceğini vurgulanmıştır.

İkinci zorluk olarak, sahip olduğumuz arka plan bilgisinin (tacit knowledge) tanımlanması sorunu ele alınmıştır. Arka plan bilgisi, insanların günlük yaşamlarında sezgisel olarak kullandıkları temel bilgileri içerir; örneğin, fiziksel dünyayı anlamak için naive fizik, temel niceliksel kavramlar için naive matematik ve sosyal ilişkilerde yol gösterici olan naive psikoloji (Brachman & Levesque, 2022). Bu bilgi, insanların çevrelerini yorumlamasını ve uygun tepkiler vermesini sağlar. Bir yapay zekâ sisteminin bağlamı kavrayabilmesi, belirsizlikle başa çıkabilmesi ve günlük durumlarda mantıklı kararlar verebilmesi için bu bilgilere hakim olması gereklidir. Ancak, bu bilginin yapay zeka sistemlerinde temsil edilmesi birçok zorluk barındırır. Arka plan bilgisinin kapsamının geniş ve çeşitli olması, onu tanımlamayı, ölçmeyi ve yapılandırılmış bir forma indirgemeyi zorlaştırır. Ayrıca, bu bilgi, farklı bağlamlarda anlam değişiklikleri gösterebilir; bu da yapay zekanın bu bağlamsal değişimleri doğru bir şekilde değerlendirmesini ve uyum sağlamasını gerektirir (Brachman & Levesque, 2022). Bu bölümde, arka plan bilgisinin doğası, sağduyulu akıl yürütmedeki kritik rolü ve bu bilginin bağlama duyarlılığı gibi konularla birlikte, insan deneyiminin formalizasyona direnen inceliklerinin yapay zekaya entegre edilmesiyle ilgili karmaşıklıkları tartışılmıştır.

Son olarak, yapay zekânın değişen ortamlarda etkili bir şekilde akıl yürütmesini sağlamak için temel bir sorun olan çerçeve problemini ele alınmıştır. Çerçeve probleminin çözülmesi, sağduyulu akıl yürütmeyi taklit etmek için kritik bir öneme sahiptir, çünkü bir yapay zekâ sisteminin gerçek dünyada insan gibi düşünebilmesi, ilgili bilgiyi ilgisiz detaylardan ayırma, değişimlere hızla uyum sağlama ve dinamik durumlarda doğru çıkarımlar yapma kapasitesine bağlıdır (Dennett, 1990). Bu bağlamda, çerçeve problemi, yapay zekânın sağduyulu akıl yürütme süreçlerini anlaması ve insanların gösterdiği esnek düşünme yeteneğini taklit etmesi önünde büyük bir engel teşkil etmektedir. Bu bağlamda, yapay zekânın gelecekteki durumları simüle etmesine veya değişimlere bağlamsal anlayışla yanıt vermesini sağlayacak şekilde bilginin temsil edilmesiyle ilgili zorluklar tartışılmıştır. Çerçeve probleminin felsefi temellerini ve tarihsel gelişimini—McCarthy, Hayes, Dennett ve Fodor gibi düşünürlerin katkılarını da içerecek şekilde—inceleyerek, bu sorunun yapay zekânın insan sağduyulu akıl yürütmesini taklit etme yeteneği üzerindeki daha geniş etkilerini vurgulanmıştır. Ayrıca, çerçeve probleminin çözümüne yönelik hem pratik hem de teorik yaklaşımları değerlendirerek, bu kritik alanda yapay zekânın yeteneklerini geliştirmek ve insan benzeri akıl yürütme kapasitesini artırmak için çok yönlü bir stratejinin önemine dikkat çekilmiştir.

İkinci bölümde, iki farklı büyük dil modeli tabanlı yapay zekâ sistemi olan ChatGPT 4.o ve Claude Sonnet 3.5 sağduyulu akıl yürütmeyi taklit etme yetenekleri çerçevesinde kapsamlı bir şekilde değerlendirilmiştir.

Sağduyulu akıl yürütmeyi değerlendirmek için, bilgi temsili ve entegrasyonu açısından esneklik ve dinamik yapıları nedeniyle büyük dil modelleri tercih edilmiştir. (Naveed ve ark., 2023). Katı, önceden tanımlanmış kurallara dayanan geleneksel kural tabanlı sistemler, değişkenlik ve belirsizlik içeren durumları ele almakta yetersiz kalmaktadır (Grosan & Abraham, 2011). Kural tabanlı sistemlerin bu sınırlı yapısı, sağduyulu akıl yürütme için gereken karmaşık ve bağlama bağlı bilgiyi etkili bir şekilde yönetmelerini engellemektedir (Brachman & Levesque, 2022). Buna karşılık, büyük dil modelleri geniş bir veri kaynağını entegre edebilir, daha zengin bir bilgi tabanı oluşturabilir, eksik veya belirsiz bilgiye dayanarak çıkarımlar yapabilir ve esnek kararlar alabilir (Naveed ve ark., 2023). Bu yetenekler, sağduyulu akıl yürütmeyi test etmek için kural tabanlı sistemler yerine büyük dil modellerini tercih etmemin başlıca nedenleridir.

ChatGPT 4.o ve Claude Sonnet 3.5, doğal dil işleme için tasarlanmış gelişmiş büyük dil modelleridir, ancak odak noktaları ve güçlü yönleri açısından önemli farklılıklar göstermektedir. OpenAI tarafından geliştirilen ChatGPT 4.o, geniş bir bilgi yelpazesini işleme ve çeşitli konular hakkında ayrıntılı yanıtlar üretme konusunda üstünlük sağlar, bu da onu oldukça çok yönlü kılar (Hello GPT-4.o, 2024). Teknik görevler, karmaşık akıl yürütme ve akademik araştırmalar için son derece uygundur. Modelin uzun bağlamları işleme ve tutarlı yanıtlar üretme yeteneği, özellikle karmaşık veya oldukça uzmanlaşmış konularla başa çıkarken önemli bir avantaj sunar. Öte yandan, Anthropic tarafından geliştirilen Claude Sonnet 3.5, güvenlik, etik endişeler ve zararlı çıktıları en aza indirme konularına büyük önem verir (Introducing Claude 3.5 Sonnet, 2024). Bu model, etik standartlara uygun yanıtlar üretmeye ve önyargıları veya uygunsuz içerikleri önlemeye odaklanmıştır. ChatGPT 4.o kadar geniş bir veri setine erişimi olmayabilir, ancak Claude Sonnet 3.5, sosyal ve duygusal bağlamları yönetmede ve çıktılarının güvenli, güvenilir ve etik olmasını sağlamada üstünlük sağlar. Tasarımı, güvenlik ve kullanıcı refahını önceliklendirir, bu da onu etik karar almanın kritik olduğu uygulamalarda özellikle etkili kılar.

Büyük dil modeli sağduyulu akıl yürütme için esnek ve güçlü bir temel sunmasına rağmen, farklı bağlamlar ve zorluklar karşısındaki performanslarını doğru bir şekilde ölçmek için sistematik bir değerlendirme çerçevesi gereklidir. Bu nedenle, büyük dil modellerinin sağduyulu akıl yürütme yeteneklerini değerlendirmek için bir değerlendirme kriteri sistemi (Kriter) kullanmayı tercih ettim. Değerlendirme kriterleri, belirli bir performansı veya yeteneği değerlendirmek için kullanılan standart bir test seti veya ölçüm aracıdır. Yapay zekâ ve makine öğrenimi alanlarında bir değerlendirme kriteri sistemi, bir modelin belirli bir görevi ne kadar iyi yerine getirdiğini değerlendirmek için kullanılır. YZ araştırmalarında değerlendirme kriterleri geliştirmek son derece önemlidir, çünkü bu, sistem yeteneklerini değerlendirmeye, ilerlemeyi izlemeye ve teknolojinin sınırlamalarını ortaya koymaya olanak tanır (Davis, 2023). Değerlendirme kriterleri yalnızca mevcut performansı değerlendirmekle kalmaz, aynı zamanda gözden kaçırılmış veya yeterince araştırılmamış sorunları vurgulayarak araştırmacıların bu alanlara odaklanmasını teşvik eder. Ayrıca, iyi tasarlanmış değerlendirme kriterleri, araştırma topluluğu için ortak bir dil ve standart sağlar, bilimsel iletişimi ve iş birliğini kolaylaştırır.

Büyük dil modellerinin sağduyulu akıl yürütme yeteneklerini değerlendirmek için, altı ana değerlendirme kriteri yapılandırılmıştır: Bağlam Temelli Bilgi Entegrasyonu, Gelecek Planlama ve Uyum Sağlama Yeteneği, Nedensel Dinamikler ve Bilgi Bağlantıları Yönetimi, Çoklu Görev Yönetimi ve Operasyonel Performans, Arka Plan Bilgisi Entegrasyonu ve Uygulaması, Bilgi Doğruluğu ve Önceliklendirme Yetkinliği. Buna ek olarak, bu ana değerlendirme kriterine 27 alt kriter eşlik etmektedir. Değerlendirme kriterlerinin geliştirilmesi süreci, teorik analizler ve akademik danışman Doç. Dr. Aziz Zambak'tan alınan geri bildirimle detaylı bir şekilde ele alınmıştır. Süreç, sağduyu bilgisi ve akıl yürütme üzerine kapsamlı bir literatür taramasıyla başlamış, yapay zekâ modellerinin bu alanlardaki zorluklarını ve değerlendirilecek temel alanları belirlemek için sağlam bir temel oluşturmuştur. Ernest Davis'in (2023) sağduyulu akıl yürütme kriterleri üzerine yazdığı makalesindeki kapsamlı öneriler, tasarım sürecine rehberlik etmiş ve değerlendirme kriterlerinin güvenilirliğini ve geçerliliğini artırmayı hedeflemiştir. Daniel Dennett'in (1990) bağlamsal bilgi entegrasyonu ve neden-sonuç ilişkileri üzerine çalışmaları, Bağlam Temelli Bilgi Entegrasyonu ve Nedensel Dinamikler ve Bilgi Bağlantıları Yönetimi kriterlerinin temelini oluşturmuştur. Aynı şekilde, Laura Morgenstern'in (1996) zamansal akıl yürütme ve planlama üzerine çalışmaları, Gelecek Planlama ve Uyum Sağlama Yeteneği kriterini şekillendirmiştir. Ayrıca, Arka Plan Bilgisi Entegrasyonu ve Uygulaması kriteri, fizik, matematik ve psikoloji gibi alanlardan gelen naive bilgilerin modeller tarafından nasıl uygulandığını test etmek için özel olarak tasarlanmıştır. Bu kapsamlı süreç sayesinde, değerlendirme kriterlerinin, modellerin sağduyulu akıl yürütme sorununu aşma kapasitesini analiz edebilmek için titizlikle geliştirilmiştir. Ortaya çıkan ana kriterler ve alt kriterler şu şekildedir:

Değerlendirme Kriteri-1: Bağlam Temelli Bilgi Entegrasyonu

Modellerin bağlamı anlama, uygun bilgiyi seçme ve bağlamsal değişikliklere uyum sağlama yeteneklerini test eder. Beş alt kriter içerir: Bağlamsal Tutarlılık, Bağlamlar

Arası Geçiş, Bağlamsal Uygunluk, Kapsamlı Doküman Yönetimi, Bağlamsal Çelişkilerin Yönetimi.

Değerlendirme Kriteri-2: Gelecek Planlama ve Uyum Sağlama Yeteneği

Modellerin geleceği planlama, belirsiz durumlara uyum sağlama ve öngörülerde bulunma becerilerini değerlendirir. 4 alt kriter vardır: Olası Senaryoları Öngörme, Stratejik Planlama, Uyum ve Esneklik, Belirsizlikle Baş Etme.

Değerlendirme Kriteri-3: Nedensel Dinamikler ve Bilgi Bağlantıları Yönetimi

Neden-sonuç ilişkilerini anlama, bağlı bilgileri işleme ve belirsiz durumları çözme yeteneklerini ölçer. 4 alt kriter vardır: Neden-Sonuç İlişkilerinin Tanımlanması, Nedensel Zincirleri Takip Etme, Koşullu Nedensellik, Etkileşimli Nedensellik.

Değerlendirme Kriteri-4: Çoklu Görev Yönetimi ve Operasyonel Performans

Modellerin bir hedefe ulaşmak için gerekli görevleri yönetme, zaman aralıklarıyla başa çıkma ve odaklı çalışma becerilerini değerlendirir. 1 adet alt kriteri vardır: Eşzamanlı Eylemleri Yönetme Yeteneği

Değerlendirme Kriteri-5: Arka Plan Bilgisi Entegrasyonu ve Uygulaması

Fizik, matematik ve psikoloji gibi temel bilgi alanlarından faydalanarak problemlere çözüm üretme yeteneğini inceler. 3 alt kriteri bulunur: Naive Fizik, Naive Matematik, Naive Psikoloji. Bu alt kriterlerin her biri kendi alt kriterlerine sahiptir. Detaylı versiyonu sonuçlar ve tartışma kısmında incelenmektedir.

Değerlendirme Kriteri-6: Bilgi Doğruluğu ve Önceliklendirme Yetkinliği

İlgisiz bilgileri filtreleme, yanlış verileri belirleme ve önemli bilgileri önceliklendirme becerilerini test eder. 4 alt kriteri bulunur: İlgisiz Bilgilerin Filtrelenmesi, Yanlış Bilgilerin Tespit Edilmesi, Bilgilerin Önceliklendirilmesi.

Çalışmada, yukarıda açıklananlar doğrultusunda iki büyük dil modelinin performansını karşılaştırmak için bir dizi deney gerçekleştirilmiştir. Bu deneyler, modellerin doğal dil işleme yeteneklerini, sağduyulu akıl yürütme becerilerini değerlendirmek üzere tasarlanmıştır. Deneyler, hazırlanan değerlendirme kriterlerine dayalı olarak tasarlanan senaryolara göre gerçekleştirilmiştir. Bu kriterler, senaryoların güvenilirliği ve geçerliliğini sağlamak açısından kritik öneme sahipti. Gerçek hayat görevlerini yansıtan senaryolar, modellerin pratik becerilerini değerlendirirken; çok boyutlu etkileşimleri analiz etmeye yönelik zengin ve karmaşık çıkarımlar içeren sorular da dâhil edildi. Sorular, insanlar için kolay ve anlaşılır bir dilde, doğal ve akıcı bir şekilde hazırlanarak hem kültürel hem de dilsel bağımsızlık sağlandı ve toplumsal önyargılardan arındırıldı. Otomatik değerlendirmeye uygun net yanıtlarla yapılandırılan senaryolar, yalnızca dilsel kalıpları değil, gerçek sağduyulu akıl yürütmeyi test etmek için tasarlandı.

ChatGPT 4.o ve Claude Sonnet 3.5'dan elde edilen yanıtları değerlendirmek için bir rubrik oluşturulmuştur. Bu rubrik, modellerin bilgi işleme, akıl yürütme ve bağlama uygun, doğru yanıtlar verme yeteneklerini değerlendirmek üzere tasarlanmıştır. Altı ana değerlendirme kriteri ve bunlara bağlı alt kriterlerden oluşan rubrik, her bir kriterin özelliklerine göre yapılandırılmıştır. Modellerin yanıtları, her alt kriter için 0 ile 5 arasında bir ölçekle değerlendirilmiştir. Bu rubrik, modellerin sağduyulu akıl yürütme performanslarını çeşitli metrikler üzerinden kantitatif olarak değerlendirme imkânı sunmuş ve güçlü ve zayıf yönlerini ortaya koymuştur. Sonuç ve tartışma bölümünde, ChatGPT 4.o ve Claude Sonnet 3.5'in performans sonuçları altı ana değerlendirme kriteri çerçevesinde sunulmuştur.

Değerlendirme Kriteri-1: Bağlam Temelli Bilgi Entegrasyon

ChatGPT 4.o ve Claude Sonnet 3.5, özellikle "Bağlamsal Tutarlılık" ve "Bağlamlar Arası Geçişler" kategorilerinde başarılı sonuçlar verdi. ChatGPT net ve verimli yanıtlarıyla öne çıkarken, Claude daha detaylı ve bazen duygusal içerikler ekledi. "Bağlamsal Uygunluk" kategorisinde her iki model bağlamı iyi anladı; ancak Claude'un profesyonel detayları onu bir adım öne taşıdı. "Kapsamlı Doküman Yönetimi" alanında, ChatGPT bilgiyi daha iyi organize ederken, Claude'un yanıtları tekrarlara ve alıntılara dayandı. "Bağlamsal Çelişkilerin Yönetimi" kategorisinde ChatGPT kısa ve net analizler sunarken, Claude daha ayrıntılı açıklamalar yaptı. Genel olarak, ChatGPT netlik ve verimlilikte, Claude ise detay ve derinlikte üstünlük sağladı. Her iki model de kriter birin gerekliliklerini tam olarak karşıladı.

Değerlendirme Kriteri-2: Gelecek Planlama ve Uyum Sağlama Yeteneği Sonuçları

ChatGPT 4.o ve Claude Sonnet 3.5, bu kriterde güçlü bir performans sergileyerek dört alt kriterin tamamında iyi bir planlama ve uyum sağlama becerisi gösterdi. "Olası Senaryoları Öngörme" alt kriterinde her iki model de verilen senaryoda olayı etkileyen anahtar faktörleri tanımladı, ancak Claude daha ayrıntılı içgörüler sundu. "Stratejik Planlama" alt alt kriterinde, Claude operasyonel detaylar ve öngörü modelleri gibi unsurlarda öne çıktı. "Uyum ve Esneklik" alt kriterinde ise Claude, anlık yeniden değerlendirme vurgusuyla dikkat çekti. "Belirsizlikle Baş Etme" alt kriterinde her iki model de risk yönetimini ele aldı, ancak Claude, ruh sağlığı desteği gibi unsurları içeren daha kapsamlı bir çerçeve sundu. Genel olarak, ChatGPT net ve yapılandırılmış yanıtlar verirken, Claude daha ayrıntılı stratejiler sundu. Bu nedenle, her iki model eşit puan aldı.

Değerlendirme Kriteri-3: Kapsamlı Nedensellik ve Bağlı Bilgi Yönetimi Sonuçları

Her iki model de neden-sonuç ilişkilerini anlamada başarılı bir performans sergiledi, ancak Claude Sonnet 3.5 genellikle daha ayrıntılı ve nüanslı analizler sundu. "Neden-Sonuç İlişkilerinin Tanımlanması" alt kriterinde, Claude alternatif açıklamalar ve çoklu faktörler üzerinde durarak ChatGPT'nin daha doğrudan yaklaşımının önüne geçti. "Nedensel Zincirleri Takip Etme" alt kriterinde, Claude daha derinlemesine analiz ve etki zincirlerini ele alırken, ChatGPT daha özlü bir açıklama sundu. "Koşullu Nedensellik" alt kriterinde Claude, daha geniş bir potansiyel eylem ve senaryo yelpazesi sunarak ChatGPT'den daha ayrıntılı bir yaklaşım sergiledi. "Etkileşimli Nedensellik" alt kriterinde ise Claude, karşılıklı bağımlı etkileri analiz ederek öne çıktı, ChatGPT ise daha doğrusal bir yaklaşım benimsedi. Özellikle "Koşullu Nedensellik" ve "Etkileşimli Nedensellik" alt kriterlerinde Claude, daha derin ve karmaşık nedensellik analizleri sunarak her birinde 1 puan daha yüksek skor aldı. Claude Sonnet 3.5 tam puan alırken ChatGPT 4.o'nun rubrik skoru 20 üzerinden 18 olmuştur.

Değerlendirme Kriteri -4: Operasyonel Uygulama Yeterliliği Sonuçları

ChatGPT 4.o ve Claude Sonnet 3.5, Operasyonel Uygulama Yeterliliği kriterinin gerekliliklerini karşılamıştır. Ancak Claude Sonnet 3.5, yanıtlarına müşteri psikolojisi ve çoklu görev yönetimini dâhil ederek daha kapsamlı ve gerçek dünya odaklı bir yaklaşım sergilemiştir. Bu sonuç, her iki modelin bir hedefe yönelik görevleri yönetmede yeterli olduğunu, ancak Claude'un gerçek dünya senaryolarının karmaşıklıklarını daha iyi ele aldığını göstermektedir.

Değerlendirme Kriteri -5: Arka Plan Bilgisi Entegrasyonu ve Uygulaması Sonuçları

Bu kriter, modellerin fizik, matematik ve psikoloji alanlarındaki naive bilgiyi gerçek dünya senaryolarına uygulama becerisini test etmiştir. ChatGPT 4.o ve Claude Sonnet 3.5 güçlü bir temel anlayış sergilemiştir; ancak Claude, tutarlı bir şekilde daha detaylı ve bağlama duyarlı yanıtlar sunmuştur.

- **Naive Fizik:** Her iki model de fiziksel prensiplere dayalı öngörüler sunmuş, ancak Claude sürtünme, enerji kaybı ve ağırlık merkezi gibi daha detaylı kavramları ele alarak daha derin bir anlayış göstermiştir.
- **Naive Matematik:** Modeller, nicelik tahmini ve mekânsal ilişkilerde başarılı olmuş, ancak Claude, gecikmeler ve manevra alanı gibi ek gerçek dünya faktörlerini dikkate alarak daha pratik çözümler önermiştir.
- **Naive Psikoloji:** Claude, karmaşık duygusal ipuçlarını tanıma ve gizli anlaşmazlıklar ya da seyirci etkisi gibi sosyal dinamikleri derinlemesine inceleyerek ChatGPT'nin daha basit yorumlarının ötesine geçmiştir.

Sonuç olarak, her iki model de arka plan bilgisini uygulamada başarılı olsa da, ChatGPT net ve doğru yanıtlarıyla öne çıkmıştır. Ancak Claude'un detaylı ve bağlamsal olarak uyumlu yanıtları, gerçek dünya senaryolarına daha iyi uyum sağlamıştır. Claude Sonnet 3.5 tam puan alırken ChatGPT 4.o'nun rubrik skoru 50 üzerinden 47 olmuştur.

Değerlendirme Kriteri-6: Doğruluk ve Alaka Yönetimi Sonuçları

"İlgisiz Bilgilerin Filtrelenmesi" alt kriterinde, Claude Sonnet 3.5, ChatGPT 4.o'a göre biraz daha iyi performans sergileyerek daha temiz ve odaklı yanıtlar sunmuş, alakasız bölümleri daha etkili bir şekilde dışarıda bırakmıştır. Her iki model de temel nitelikleri ve deneyimleri vurgulamada başarılı olmuştur. "Yanlış Bilgilerin Tespit Edilmesi" alt kriterinde, modeller yanlışlıkları başarılı bir şekilde belirlemiştir. ChatGPT 4.o, abartılı bir istatistiği detaylı bir şekilde analiz ederek öne çıkarken, Claude Sonnet 3.5 yanlış bir bilginin dahil edildiğini tespit ederek güçlü bir hata ayıklama becerisi göstermiştir. "Bilgilerin Önceliklendirilmesi" alt kriterinde, her iki model de ilgili bilgileri etkili bir şekilde önceliklendirmiştir. Claude, bilgileri kategorize ederek detay eklerken, ChatGPT öncelikleri kısa ve doğru bir özetle sunmuştur. Genel olarak, her iki model de bilgileri filtreleme, tespit etme ve önceliklendirme konusunda güçlü beceriler sergilemiştir. Claude, ilgisiz verileri dışlama ve yapılandırılmış, detaylı yanıtlar sunma konusunda biraz daha başarılı olmuştur. Claude Sonnet 3.5 tam puan alırken ChatGPT 4.o'nun rubrik skoru 15 üzerinden 14 olmuştur.

Bu tez, ChatGPT 4.o ve Claude Sonnet 3.5'in sağduyulu akıl yürütmenin bazı unsurlarını taklit etmede güçlü yetenekler sergilediğini, ancak her birinin farklı alanlarda öne çıktığını göstermektedir. ChatGPT, netlik ve verimlilikte üstünlük sağlarken, Claude daha detaylı ve bağlama duyarlı yanıtlar sunmuş, özellikle nedensellik, bağlam yönetimi ve uyum sağlama gibi alanlarda öne çıkmıştır. Claude'un sosyal ve duygusal faktörlere odaklanması, geleceği planlama ve arka plan bilgisini uygulama gibi alanlarda ona genellikle avantaj sağlamıştır. Her iki model de başarılı bir performans göstermiş olsa da sağduyulu akıl yürütmeyi gerçekten taklit edip edemediklerini doğrulamak için daha fazla kriter testi ve uzman analizi gereklidir.

Bu araştırma, YZ'de sağduyulu akıl yürütmenin anlaşılmasına katkıda bulunarak insan benzeri zekâya doğru önemli bir adım atmaktadır. Geliştirilen kriterler, büyük dil modellerinin değerlendirilmesi için net bir çerçeve sunarak, YZ'nin faillik (agency), yönelimsellik (intentionality) ve bilgi temsili konularındaki tartışmaları desteklemektedir. Bulgular, sağlık hizmetleri, otonom araçlar ve müşteri hizmetleri gibi karmaşık gerçek dünya görevlerini yönetebilecek YZ sistemlerinin geliştirilmesi için pratik bir değer taşımaktadır. Mevcut YZ yetenekleri ile insan merkezli görevler için gerekli akıl yürütme arasındaki boşluğu kapatarak, bu çalışma daha uyumlu ve bağlama duyarlı sistemlerin inşasına yardımcı olmaktadır.

Bu araştırma, gelecekteki çalışmalarda iyileştirme yapılabilecek dört temel sınırlamayı ortaya koymaktadır:

1. **Değerlendirme Formatları:** Serbest yanıt formatları detaylı içgörüler sağlasa da puanlamayı daha öznel hale getirmiştir. Doğru/yanlış veya çoktan seçmeli gibi yapılandırılmış formatların eklenmesi, bu yaklaşımı tamamlayarak daha standart bir değerlendirme sağlayabilir.

2. **Görev Çeşitliliği:** Kriterler yalnızca dil görevlerine odaklanmış olup değerlendirme kapsamını sınırlamıştır. Gelecekteki araştırmalarda, görsel veya mekânsal akıl yürütme gibi diğer görevleri de içeren senaryolar dâhil edilmelidir.

3. **Senaryo Çeşitliliği:** Test senaryolarının sınırlı sayıda olması, sonuçların genellenebilirliğini azaltmıştır. Senaryo yelpazesini genişletmek ve örneklem büyüklüğünü artırmak, daha sağlam ve geniş ölçekte uygulanabilir bulgular sağlayabilir.

4. **Uzman İnceleme Havuzu:** Değerlendirmeler yalnızca benim ve Doç. Dr. Aziz Zambak'ın incelemeleriyle sınırlı kalmıştır. Gelecekteki çalışmalara daha fazla uzman dâhil edilmesi, farklı bakış açıları ekleyerek analizin nesnelliğini ve güvenilirliğini artırabilir

Anlamlı etkileşim kurabilen YZ geliştirmek için sağduyulu akıl yürütmenin karmaşıklıklarını derinlemesine incelememiz gerekmektedir. Bu çalışma, gelecekteki ilerlemeler için bir yol haritası sunarak, insan yaşamının incelikli gerçeklikleriyle başa çıkabilecek sistemler için bir temel sağlamaktadır. Alan Turing, makinelerin bizi anlayışlarıyla şaşırtacağı bir zaman hayal etmişti; bu araştırmayı, Yapay zekanın sağduyu derinliği ve inceliğiyle akıl yürütebilme yeteneğine bir adım daha yaklaştıran bir ilerleme olarak görüyorum.

D. SAMPLE TEZ İZİN FORMU / THESIS PERMISSION FORM

ENSTİTÜ / INSTITUTE

Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences ☐

Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences ☒

Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematics ☐

Enformatik Enstitüsü / Graduate School of Informatics ☐

Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences ☐

YAZARIN / AUTHOR

Soyadı / Surname : Kabadere
Adı / Name : Zeynep
Bölümü / Department : Felsefe / Philosophy

TEZİN ADI / TITLE OF THE THESIS (**İngilizce** / English):

Bringing to Light: The Challenges of Representing and Reasoning Common Sense Knowledge in AI Systems

TEZİN TÜRÜ / DEGREE**:** **Yüksek Lisans** / Master ☒ **Doktora** / PhD ☐

1. **Tezin tamamı dünya çapında erişime açılacaktır. /** Release the entire work immediately for access worldwide. ☒

2. **Tez iki yıl süreyle erişime kapalı olacaktır.** / Secure the entire work for patent and/or proprietary purposes for a period of **two years. *** ☐

3. **Tez altı ay süreyle erişime kapalı olacaktır.** / Secure the entire work for period of **six months**. ***** ☐

***** *Enstitü Yönetim Kurulu kararının basılı kopyası tezle birlikte kütüphaneye teslim edilecektir. / A copy of the decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.*

Yazarın imzası / Signature .... **Tarih** / Date ....
(Kütüphaneye teslim ettiğiniz tarih. Elle doldurulacaktır.)
(Library submission date. Please fill out by hand.)

Tezin son sayfasıdır. / This is the last page of the thesis/dissertation.