Gemini in Reasoning: Unveiling Commonsense in Multimodal Large Language Models

Anonymous EMNLP submission

Abstract

The burgeoning interest in Multimodal Large Language Models (MLLMs), such as OpenAI's GPT-4V(ision), has significantly impacted both academic and industrial realms. These models enhance Large Language Models (LLMs) with advanced visual understanding capabilities, facilitating their application in a variety of multimodal tasks. Recently, Google introduced Gemini, an advanced MLLM for multimodal integration. Despite its advancements, preliminary benchmarks indicate that Gemini lags behind GPT models in commonsense reasoning tasks. However, these assessments, using a limited dataset like HellaSWAG, may not fully reflect Gemini's true potential in commonsense reasoning. To address this gap, our study undertakes a thorough evaluation of Gemini's performance in complex reasoning tasks that necessitate the integration of commonsense knowledge across modalities. We carry out a comprehensive analysis of 12 commonsense reasoning datasets, ranging from general to domain-specific tasks. This includes 11 datasets focused solely on language, as well as one that incorporates multimodal elements. Our experiments across ten LLMs and two MLLMs demonstrate Gemini's competitive commonsense reasoning capabilities. We also highlight common challenges faced by current LLMs and MLLMs in commonsense reasoning, emphasizing the need for further advancements.

1 Introduction

Commonsense reasoning, integral to human cognition, plays a crucial role in navigating the intricacies of everyday life. Consider a scenario where someone decides what to wear based on the weather. This decision extends beyond the mere selection of attire; it involves understanding weather patterns, the suitability of clothing for different temperatures, and the social context of the occasion. It's about synthesizing diverse pieces of knowledge: a forecast predicting rain, the practical necessity for a raincoat, and the societal expectation of dressing appropriately for an event. This reasoning goes beyond simply processing information; it entails integrating varied pieces of knowledge that humans often take for granted. A major challenge in Natural Language Processing (NLP) research is the ambiguity and under-specification of human language. Individuals rely heavily on their commonsense knowledge and reasoning abilities to decipher these ambiguities and infer missing information. Commonsense reasoning has consistently posed unique challenges in NLP research (Li et al., 2021; Bian et al., 2023), encompassing spatial, physical, social, temporal, and psychological aspects, along with an understanding of social norms, beliefs, values, and the nuances of predicting and interpreting human behavior (Liu and Singh, 2004). Models often lack this innate commonsense, hindering their ability to contextualize data coherently, in stark contrast to the human capacity for effortlessly understanding everyday situations (Shwartz and Choi, 2020; Bhargava and Ng, 2022).

Recent advances in Large Language Models (LLMs) have sparked unprecedented enthusiasm in the NLP community and beyond, significantly enhancing a wide array of applications (Min et al., 2021; Zhao et al., 2023; Wang et al., 2023; Kasneci et al., 2023; He et al., 2023). Building on these achievements, Multimodal Large Language Models (MLLMs) have emerged as a pivotal focus in the next wave of AI (Wu et al., 2023b), speculated to advance towards Artificial General Intelligence (AGI), which aims to develop AI systems smarter than humans and beneficial for all of humanity (Rayhan et al., 2023). The rise of MLLMs, particularly OpenAI's GPT-4V(ision) (Yang et al., 2023) and Google's Gemini (Team et al., 2023), marks significant progress in this area. Among these developments, Gemini emerges as a formidable challenger to the state-of-the-art MLLM, GPT-4V, specially engineered for multimodal integration. Its release has ignited constructive discussions about the current level of AGI achievement. In widely used academic benchmarks, Gemini has attained new state-of-the-art status in the majority of tasks. However, preliminary evaluations of Gemini, especially when compared to models like the GPT series, have indicated potential shortcomings in its commonsense reasoning capabilities, a fundamental aspect of human cognition. Yet, it is important to consider that basing the assessment of Gemini's commonsense reasoning abilities solely on the HellaSWAG dataset (Zellers et al., 2019b) may not comprehensively reflect Gemini's full scope in this domain.

To address the gap in the comprehensive evaluation of Gemini's real-world performance in commonsense reasoning tasks, our study conducts extensive experiments across 12 commonsense reasoning datasets, covering a broad spectrum of domains such as general, physical, and temporal reasoning. The definitions of all tasks can be found in Appendix A. We experiment with ten popular LLMs for the language dataset evaluation, such as Gemma (Team et al., 2024), Llama3 (AI@Meta, 2024), Gemini Pro (Team et al., 2023), and GPT-4 Turbo (OpenAI, 2023). For the multimodal dataset, we assess both Gemini Pro Vision and GPT-4V. Our key findings are summarized as follows: (1) Overall, Gemini Pro's performance is comparable to that of GPT-3.5 Turbo, demonstrating marginally better average results across 11 language datasets (1.4% higher accuracy), though it lags behind GPT-4 Turbo by an average of 8.2% in accuracy. Moreover, Gemini Pro Vision exhibits lower performance than GPT-4V on the multimodal dataset, except for temporal-related questions. (2) Approximately 65.8% of Gemini Pro's reasoning processes are evaluated as logically sound and contextually relevant, indicating its potential for effective application in various domains. (3) Gemini Pro encounters significant challenges in temporal and social commonsense reasoning, indicating key areas for further development. (4) Our manual error analysis reveals that both open-source and closed-source LLMs struggle most with misunderstanding contextual information, accounting for 27.4% and 25.7% of total errors, respectively. Furthermore, Gemini Pro Vision struggles particularly with identifying emotional stimuli in images, especially those involving human entities, which constitutes 32.6% of its total errors.

In summary, our contributions are threefold:

(1) We provide the first thorough evaluation of Gemini Pro's efficacy in commonsense reasoning tasks, employing 12 diverse datasets that span both language-based and multimodal scenarios.

(2) Our study reveals that Gemini Pro exhibits performance comparable to GPT-3.5 Turbo and Llama3-70B in language-only tasks, demonstrating logical and contextual reasoning processes. However, it lags behind GPT-4 Turbo in accuracy and encounters challenges in temporal and social reasoning, as well as in emotion recognition in images.

(3) Our findings lay a valuable foundation for future research in the field of commonsense reasoning within LLMs and MLLMs, highlighting the necessity to enhance specialized domains in these models and the nuanced recognition of mental states and emotions in multimodal contexts.

2 Experimental Setup

2.1 Datasets

We experiment with 12 datasets related to different types of commonsense reasoning, which include 11 language-based datasets and one multimodal dataset. The language-based datasets encompass three main categories of commonsense reasoning problems. **General and Contextual Reasoning:** (1) CommonsenseQA (Talmor et al., 2019), focusing on general commonsense knowledge; (2) Cosmos QA (Huang et al., 2019), emphasizing contextual understanding narratives, (3) αNLI (Bhagavatula et al., 2019), introducing abductive reasoning, which involves inferring the most plausible explanation; and (4) HellaSWAG, centering around reasoning with contextual event sequences. **Specialized and Knowledge Reasoning:** (1) TRAM (Wang and Zhao, 2023b), testing reasoning about time; (2) NumerSense (Lin et al., 2020), focusing on numerical understanding; (3) PIQA (Bisk et al., 2020), assessing physical interaction knowledge; (4) QASC (Khot et al., 2020), dealing with science-related reasoning; and (5) RiddleSense (Lin et al., 2021), challenging creative thinking through riddles. **Social and Ethical Reasoning:** (1) Social IQa (Sap et al., 2019), testing the understanding of social interactions; and (2)

ETHICS (Hendrycks et al., 2020), evaluating moral and ethical reasoning. For the multimodal dataset (vision and language), we select VCR (Zellers et al., 2019a), a large-scale dataset for cognition-level visual understanding. For datasets like TRAM and ETHICS, which include multiple tasks, we extract the commonsense reasoning part for experiments. We employ accuracy as the performance metric for all datasets. More details about each dataset, as well as example questions, are in Appendix B.

2.2 Models

We consider ten popular LLMs for language-based dataset evaluation, including the open-source models Llama-2-70b-chat (Touvron et al., 2023), Phi-3-mini-128k-Instruct, Phi-3-medium-128k-Instruct (Abdin et al., 2024), Gemma-2B-Instruct, Gemma-7B-Instruct (Team et al., 2024), Llama3-8B-Instruct, and Llama3-70B-Instruct (AI@Meta, 2024), as well as the closed-source models Gemini 1.0 Pro (Team et al., 2023), GPT-3.5 Turbo, and GPT-4 Turbo (OpenAI, 2023). Each of the closed-source models is accessed using its corresponding API key, while open-source models are loaded from Hugging Face. Given the constraints of API costs and rate limitations, we randomly select 200 examples from the validation set for each language-based dataset following (Wang and Zhao, 2023b) and 50 examples from the validation set for the VCR dataset following (Liu and Chen, 2023). For all evaluations, we employ greedy decoding (i.e., temperature = 0) during model response generation. Notably, there are instances where the models decline to respond to certain queries, particularly those involving potentially illegal or unethical content. Sometimes, models provide answers that are outside the scope of the options. In these cases, we categorize these unanswered questions as incorrect.

2.3 Prompts

In the evaluation of language-based datasets, we employ 5-shot chain-of-thought (CoT) prompting (Wei et al., 2022), which generally enhances the reasoning capabilities of models by helping them better understand and solve tasks through a logical sequence of thought processes. For the multimodal dataset, we utilize zero-shot standard prompting (SP) (Kojima et al., 2022) to assess the authentic end-to-end visual commonsense reasoning abilities of MLLMs, ensuring that their performance is measured without prior exposure to similar tasks.

3 Results

3.1 Overall Performance Comparison

Table 1 demonstrates the accuracy comparison of ten LLMs under the 5-shot CoT setting on 11 language-based commonsense reasoning datasets. There are several key takeaways. First, GPT-4 Turbo achieves the highest average performance across the datasets. Llama3-70B follows closely with a strong performance of 83.7%, significantly improving upon its predecessor, Llama2-70B, due to a substantial increase in pretraining data. Gemini Pro also demonstrates competitive performance with an average score of 82.1%. Smaller-scale models, such as Phi-3-mini and Gemma-2B, while not leading in performance, still demonstrate the potential for achieving reasonable accuracy with fewer resources. From a dataset standpoint, it is evident that while these models exhibit commendable performance across a broad spectrum of commonsense domains, they encounter challenges in specific areas, particularly those involving temporal reasoning (TRAM) and social commonsense (Social IQa). These insights highlight the progress and remaining challenges in the development of LLMs for commonsense reasoning.

For the multimodal VCR dataset, we report the performance of GPT-4V and Gemini Pro Vision in Table 2. The VCR consists of three subtasks: (1) Q $\rightarrow$ A, which involves generating an answer to a question based on the visual context; (2) QA $\rightarrow$ R, which requires the model to produce a rationale for a given answer; and (3) Q $\rightarrow$ AR, which challenges the model to both answer the question and justify the response with appropriate rationales. In all subtasks, GPT-4V demonstrates superior performance compared to Gemini Pro Vision, indicating a more robust capacity for integrating visual and textual information to provide coherent responses. In Q $\rightarrow$ AR, the relatively lower performance of both models, compared to the other two subtasks, suggests that there is considerable room for improvement in understanding the interplay between visual cues and commonsense reasoning.

3.2 Effects of Commonsense Domain

Referring to Section 2.1, we have categorized 11 language-based datasets into three groups and presented the performance for each setting within each group in Figure 1. Our findings indicate that GPT-4 Turbo consistently leads with particularly high scores, notably achieving 91.2% in Social and Eth-

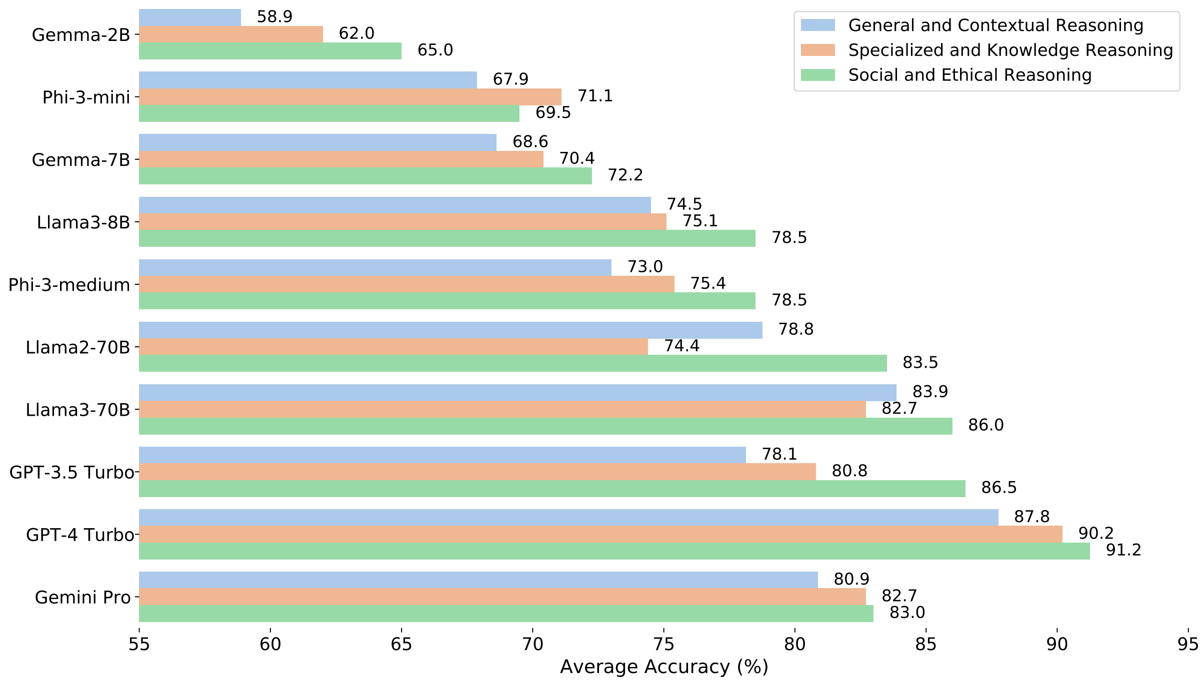


Figure 1: Average model performance across three major commonsense reasoning categories over 11 language-based datasets, including General and Contextual Reasoning (CommonsenseQA, Cosmos QA, αNLI, HellaSWAG), Specialized and Knowledge Reasoning (TRAM, NumerSense, PIQA, QASC, RiddleSense), and Social and Ethical Reasoning (Social IQa, ETHICS).

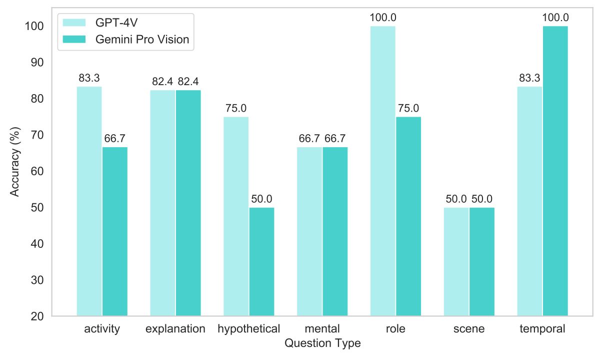


Figure 2: Performance comparison between GPT-4V and Gemini Pro Vision on the VCR dataset, categorized by question type, with a focus on the "Q $\rightarrow$ A" subtask. Within our sample of 50 questions, the distribution across each type is as follows: activity (12), explanation (16), hypothetical (3), mental (4), role (5), scene (4), and temporal (6). GPT-4V matches or surpasses Gemini Pro Vision in performance across these question types, with the exception of the temporal category.

ical Reasoning. Llama3-70B shows significant improvement over Llama2-70B, benefiting from increased pretraining data and enhanced architecture, and performs well in Social and Ethical Reasoning with a score of 86.0%. Gemini Pro also demonstrates strong performance, slightly outperforming GPT-3.5 Turbo in two of the three categories. Notably, its performance dip in the Social and Ethical Reasoning group may stem from its tendency to refuse to answer questions that could potentially involve unethical content, which we have counted as incorrect in our evaluation. Based on our experiments, among the 200 samples, Gemini Pro refuses to answer 3.0% of the problems (6 in total) in the Social IQa dataset and 6.5% of the problems (13 in total) in the ETHICS dataset. Smaller-scale models like Phi-3-medium and Phi-3-mini exhibit respectable accuracies, with Phi-3-medium achieving solid scores across the board. For Gemma models, the 7B variant shows a significant jump in performance compared to the 2B variant, up to 16.6%, as demonstrated in Social and Ethical Reasoning. Overall, most models exhibit robust capabilities in handling Social and Ethical Reasoning datasets, suggesting a relatively advanced grasp of moral and social norms. However, there is a notable disparity in their performance on General and Contextual Reasoning tasks, indicating a potential gap in their understanding of broader commonsense principles and their application in varied contexts. The Specialized and Knowledge Reasoning category, particularly in the realms of temporal and riddle-based

Table 1: Performance comparison (%) of ten LLMs across 11 language-based commonsense reasoning datasets. The best results for the 5-shot setting are boldfaced.

Dataset	Gemma *2B*	Phi-3-mini *3.8B*	Gemma *7B*	Llama3 *8B*	Phi-3-medium *14B*	Llama2 *70B*	Llama3 *70B*	GPT-3.5 *175B*	GPT-4 *>175B*	Gemini Pro *>175B*
CommonsenseQA	55.0	68.5	66.0	70.0	71.5	76.5	79.0	76.0	**80.0**	79.0
Cosmos QA	61.0	70.0	72.5	74.5	75.0	81.0	85.5	78.5	**88.0**	84.5
αNLI	58.0	67.0	66.5	77.5	72.0	80.5	84.5	78.0	**88.0**	81.5
HellaSWAG	61.5	66.0	69.5	76.0	73.5	77.0	86.5	80.0	**95.0**	78.5
TRAM	63.0	69.5	66.5	70.0	72.5	70.0	78.5	72.0	**82.0**	76.0
NumerSense	56.0	67.0	63.5	71.5	72.0	75.5	81.5	82.5	**86.0**	82.0
PIQA	68.5	76.0	79.0	81.5	78.0	78.5	88.0	89.5	**95.5**	90.5
QASC	60.5	74.5	76.0	79.0	79.0	82.0	86.0	85.0	**92.5**	82.5
RiddleSense	62.0	68.5	67.0	73.5	75.5	66.0	79.5	75.0	**95.0**	82.5
Social IQa	57.5	62.5	65.0	72.5	71.0	77.5	80.5	78.0	**84.5**	78.5
ETHICS	72.5	76.5	79.5	84.5	86.0	89.5	91.5	95.0	**98.0**	87.5
Average	59.0	69.6	70.1	75.5	75.1	77.6	83.7	80.9	**89.5**	82.1

Table 2: Performance comparison between GPT-4V and Gemini Pro Vision on the VCR dataset. "Q $\rightarrow$ A" evaluates question-answering accuracy, "QA $\rightarrow$ R" assesses answer justification, and "Q $\rightarrow$ AR" measures the performance of both correctly answering questions and selecting rationales. GPT-4V outperforms Gemini Pro Vision across all subtasks.

Method	Q $\rightarrow$ A	QA $\rightarrow$ R	Q $\rightarrow$ AR
GPT-4V	80.0	72.0	56.0
Gemini Pro Vision	74.0	70.0	48.0

challenges, highlights specific deficiencies in the models' abilities to process complex temporal sequences and to engage in the abstract and creative thought required to decipher riddles.

Regarding the multimodal dataset, Figure 2 details the comparative performance between GPT-4V and Gemini Pro Vision across different question types, in alignment with the guidelines of the VCR dataset (Zellers et al., 2019a). We concentrate on the "Q $\rightarrow$ A" subtask as it most directly assesses the models' visual commonsense capabilities. Considering the data sample for each type, Gemini Pro Vision's performance either matches or is slightly lower than GPT-4V's, except in temporal-type questions, where it surpasses GPT-4V. This suggests its enhanced capability not only in recognizing but also in contextualizing time-related elements within visual scenarios.

3.3 Reasoning Justification within MLLMs

To assess the reasoning capabilities of MLLMs, particularly their ability to provide not only correct answers but also sound and contextually grounded reasoning in matters of commonsense, we adopted a systematic sampling approach. We selected a subset of four models for this process, including Llama2-70B, GPT-3.5 Turbo, GPT-4 Turbo, and Gemini Pro. For each of the language datasets evaluated with these models, we randomly selected 30 questions that were correctly answered and 30 questions that were incorrectly answered by each LLM, following (Bian et al., 2023). In cases where a dataset presented fewer than 30 incorrect answers, we included all available incorrect responses to ensure comprehensive analysis. After selecting these questions, we prompted each model to explain "*What is the rationale behind the answer to the question?*" The reasoning processes provided by the models were then manually reviewed and classified as either True or False, based on their logical soundness and relevance to the question. Figure 3 illustrates a comprehensive view of the average reasoning correctness across the 11 datasets, in terms of the sampled correct and incorrect questions. In fact, not every model had 30 incorrect questions for each dataset. In such scenarios, we scaled the available data up to 30 questions to ensure standardized computation. Figure 3 shows that GPT-4 Turbo's leading performance in both correct and incorrect answers highlights its advanced reasoning mechanisms and its ability to maintain coherent logic, even when the final answers are not accurate. Additionally, Gemini Pro has emerged as a notably proficient model, generally demonstrating commendable reasoning abilities and offering a well-rounded approach to commonsense reasoning. GPT-3.5, while trailing slightly behind Gemini Pro, still demonstrates competitive reasoning abilities.

Appendix C presents two real examples from Gemini Pro and GPT-3.5, illustrating the cases of a correct answer with a correct rationale and an incorrect answer with an incorrect rationale, respectively.

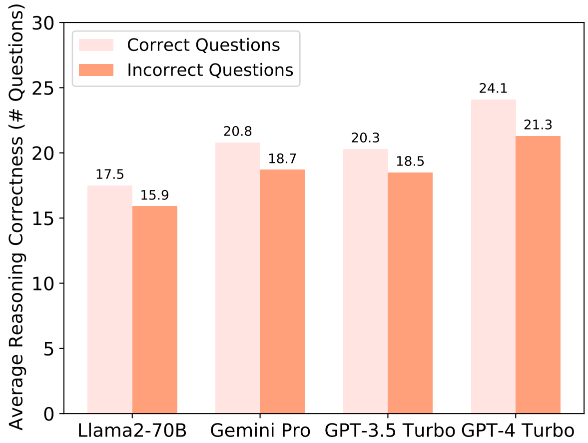


Figure 3: Average reasoning correctness across 11 language datasets. The comparison among four LLMs is based on a random sample of 30 correct and 30 incorrect questions per dataset. In cases where a dataset contained fewer than 30 incorrect questions, the data were scaled up to maintain consistency in the sample size.

Moving to the multimodal perspective, our analysis of GPT-4V and Gemini Pro Vision on the VCR dataset reveals notable patterns in reasoning correctness. With GPT-4V at 24% and Gemini Pro Vision at 26%, approximately one-quarter of the cases showed both models correctly identifying the answers but failing to provide appropriate rationale. This discrepancy suggests that while the models can often determine the correct outcomes, their ability to understand or explain the underlying reasoning behind these answers is not consistently aligned. Furthermore, in the instances of incorrect answers, GPT-4V and Gemini Pro Vision showed correct rationales 16% and 22% of the time, respectively. This indicates that, despite arriving at incorrect conclusions, the models demonstrate a capacity for effective reasoning or logical processing. However, this does not consistently translate into accurate outcomes, implying that while some aspects of the required knowledge are captured, other crucial elements are likely missed.

3.4 Case Study: Gemini Pro in Commonsense Reasoning

Given our focus on evaluating the commonsense reasoning capabilities of the Gemini Pro model, we conduct a qualitative analysis to assess its performance across representative examples in four major categories (three language-based and one multimodal), as described in Section 2.1. To ensure an authentic end-to-end capability evaluation, we present examples under the zero-shot learning setting, employing standard prompting techniques. Due to space constraints, we present two examples here; additional examples are in Appendix D.

General (CommonsenseQA). In the general commonsense evaluation (General and Contextual Reasoning category) using the CommonsenseQA dataset, consider the example question: “People are what when you’re a stranger? (A) train (B) strange (C) human (D) stupid (E) dangerous.” Gemini Pro correctly chose (B) “strange,” and its reasoning process is notable. It recognized that while all options relate to the concept of a “stranger”, only “strange” accurately encapsulates the neutral and open-ended nature of the question. The model effectively ruled out other options: (A) “train”, for being too specific and unrelated; (C) “human”, as accurate but not capturing the question’s essence; (D) “stupid”, for being judgmental and offensive; and (E) “dangerous”, due to its negative connotation. This selection of “strange” indicates an understanding of the unfamiliar nature associated with strangers, highlighting Gemini Pro’s capability in interpreting and applying general commonsense knowledge appropriately.

Visual (VCR). In the visual commonsense evaluation using the VCR dataset, we analyzed Gemini Pro Vision’s response to a scenario involving physical safety and potential danger, as shown in Figure 4. Presented with an image of individuals on the edge of a cliff, the model was questioned: “What would happen if person 4 pushed person 3 at this moment?” In this context, Gemini Pro Vision’s response mirrored the logical inference that if the second person from the left (person 4) pushed the third person from the left (person 3), the result would be person 3 falling off the cliff, leading to a fatal outcome. This example from the VCR dataset underscores Gemini Pro Vision’s ability to analyze visual scenes and make predictions about the potential consequences of actions within those scenes, a crucial aspect of visual commonsense reasoning. It demonstrates the model’s grasp of spatial relations and physical consequences, providing evidence of its capacity to process and reason about complex visual information akin to human cognition.

Overall, the cases presented underscore the ad-

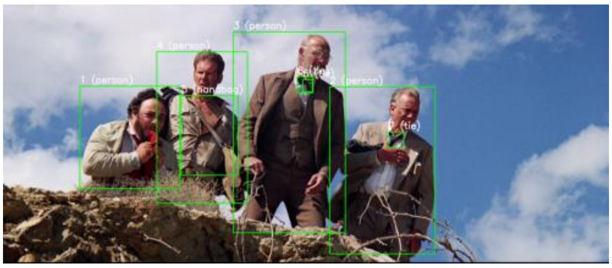


Figure 4: Example image from the VCR dataset.

vanced reasoning capabilities of Gemini Pro and Gemini Pro Vision, while also identifying challenges in achieving human-like inference.

3.5 Error Analysis

To gain a deeper understanding of model mistakes, we manually analyzed instances where a model made incorrect choices or provided inappropriate answers. We conducted a thorough examination of common error types encountered in commonsense reasoning tasks. Our focus was to separately assess open-source and closed-source LLMs to gain specific insights into their performance. Table 3 shows the proportions of five common error types, averaged across the LLMs within each category.

Context misinterpretation emerged as the most frequent error, occurring more often in open-source LLMs (27.4%) compared to closed-source LLMs (25.7%). This trend suggests that open-source models may struggle more with understanding scenarios, leading to more contextual misunderstandings. Notably, knowledge errors, where models misapplied commonsense knowledge, were higher in closed-source LLMs (25.3%) compared to open-source LLMs (23.3%). This suggests that while closed-source LLMs generally perform better in logical consistency and resolving ambiguities, they may face challenges in applying complex or nuanced knowledge accurately, possibly due to overfitting to some training data patterns. Logical errors were also common among LLMs, accounting for 22.4% in open-source LLMs and slightly less in closed-source LLMs (21.8%), indicating that closed-source models maintain more consistent logical reasoning. Ambiguity errors, at 16.2% in open-source LLMs, were reduced to 10.9% in closed-source LLMs, demonstrating the latter's effectiveness in resolving language ambiguities. In contrast, overgeneralization errors showed an increase in closed-source LLMs (16.3%) compared to open-source LLMs (10.7%), possibly due to overextending learned patterns.

Table 3: Proportion of common error types in commonsense reasoning in LLM evaluation. Misinterpret. represents misinterpretation. Closed-source LLMs refer to Gemini Pro, GPT-3.5 Turbo, and GPT-4 Turbo, while the rest belong to open-source LLMs.

Error Type	Open-Source LLMs	Closed-Source LLMs
Context Misinterpret.	27.4%	25.7%
Knowledge Errors	23.3%	25.3%
Logical Errors	22.4%	21.8%
Text Ambiguity	16.2%	10.9%
Overgeneralization	10.7%	16.3%

In our analysis of the VCR dataset, we focused on instances where either GPT-4V or Gemini Pro Vision chose incorrect answers in the Q $\rightarrow$ A subtask. The four common error types for each model are summarized in Table 4. Emotion recognition errors were the most common, with GPT-4V encountering these errors in 30.1% of cases and Gemini Pro Vision slightly more at 31.3%. This high incidence suggests that both models find interpreting emotional cues in visual content particularly challenging, underscoring the complexity of deciphering human emotions from visual stimuli. Spatial perception errors were also significant, constituting 22.5% of errors for GPT-4V and 25.2% for Gemini Pro Vision. These figures indicate the models' difficulties in accurately understanding spatial relationships and the arrangement of elements in images. Logical errors were another major error type, more pronounced in GPT-4V (27.7%) than in Gemini Pro Vision (24.9%), pointing to challenges in logical reasoning within visual contexts. Context misinterpretation, although less frequent, is still a notable issue, with GPT-4V at 19.7% and Gemini Pro Vision at 18.6%. These errors demonstrate the models' struggles with grasping the overarching context or narrative depicted in the visual content.

Overall, error analysis sheds light on the specific challenges LLMs and MLLMs face in commonsense reasoning, providing valuable insights for future improvements for future model refinement.

Table 4: Proportion of commmon error types in visual commonsense reasoning in MLLM evaluation (GPT-4V and Gemini Pro Vision). Misinterpret.: and E. represent misinterpretation and errors, respectively.

Error Type	GPT-4V	Gemini Pro Vision
Context Misinterpret.	19.7%	18.6%
Spatial Perception E.	22.5%	25.2%
Emotion Recognition E.	30.1%	31.3%
Logical Errors	27.7%	24.9%

4 Related Work

Commonsense Reasoning in NLP. Commonsense reasoning has gained renewed attention in recent years, especially in the context of advancements in LLMs that have significantly influenced numerous applications in NLP. However, there is a growing concern about their ability to understand and reason about commonsense knowledge (Storks et al., 2019; Tamborrino et al., 2020; Bhargava and Ng, 2022). This concern is echoed in various studies that focus on evaluating the capabilities of LLMs in this area (Bian et al., 2023; Weng et al., 2023; Shen and Kejriwal, 2023). Concurrently, researchers have been exploring diverse strategies to enhance the commonsense reasoning of NLP systems, from leveraging knowledge graphs to commonsense knowledge transfer (Huang et al., 2023; Ye et al., 2023; Zhou et al., 2023). Prior to delving into methodological refinements, a comprehensive evaluation is essential to understand the authentic commonsense reasoning capabilities of LLMs. In our study, we endeavor to advance this line of inquiry by examining how LLMs, particularly focusing on the Gemini model, navigate and implement commonsense reasoning in various NLP contexts.

Training Paradigms in LLMs. In NLP research, pretraining language models on large-scale varied textual datasets has become essential. BERT-based models like BERT (Kenton and Toutanova, 2019) and RoBERTa (Liu et al., 2019) exemplify this, being applied to tasks ranging from disease prediction (Zhao et al., 2021) to text classification (Wang et al., 2022b) and time series analysis (Wang et al., 2022c). The debut of GPT-3 shifted this focus towards more flexible learning methods like zero-shot and few-shot learning, showcasing models' adaptability to new tasks with minimal data (Brown et al., 2020). This shift has spurred the development of novel prompting techniques to enhance LLMs' reasoning and understanding capabilities, including CoT prompting (Wei et al., 2022), self-consistency with CoT (Wang et al., 2022a), tree-of-thought prompting (Yao et al., 2023), and metacognitive prompting (Wang and Zhao, 2023a). In this work, we evaluate ten LLMs for language tasks and two MLLMs for multimodal tasks under few-shot settings to provide an in-depth understanding of their strengths and limitations in diverse commonsense reasoning tasks.

Evaluations on MLLMs. Since the release of the state-of-the-art MLLM, GPT-4V, several evaluations have been conducted across diverse tasks, including medical imaging (Wu et al., 2023a), visual question answering (Li et al., 2023; Yang et al., 2023), and video understanding (Lin et al., 2023), focusing on either on case-by-case qualitative analyses or on quantitative assessments across diverse tasks. The recent release of Google's Gemini has garnered considerable attention, and early experiments have been conducted to evaluate its capabilities in both language understanding (Akter et al., 2023) and the multimodal domain (Liu and Chen, 2023; Fu et al., 2023). However, a significant gap remains in fully comprehending the commonsense reasoning capabilities of Gemini, a known potential shortcoming since its introduction. Our work comprehensively analyzes Gemini's capabilities in this area, comparing it with other MLLMs to highlight its potential and areas for improvement.

5 Discussion

In this study, we conducted a comprehensive evaluation of ten state-of-the-art LLMs and two MLLMs, focusing particularly on Gemini Pro and Gemini Pro Vision, across 12 diverse commonsense reasoning datasets. Our findings indicate that while these models mark a significant advancement in various domains, demonstrating impressive performance in commonsense reasoning tasks, they still exhibit limitations, particularly in tasks requiring deep contextual understanding or abstract reasoning, such as those involving temporal dynamics, riddles, or intricate social scenarios. Although significant progress has been made, achieving AGI still represents a substantial goal on the horizon. Our work sets the stage for future research in this field, highlighting both the achievements and areas needing improvement in commonsense reasoning.

Looking ahead, addressing these challenges is crucial to enhance the overall effectiveness of LLMs and MLLMs in commonsense reasoning. Future research should aim to refine the models' capabilities in interpreting and reasoning within complex contexts and abstract scenarios to enhance their adaptability to real-world applications. Additionally, there is an emerging need for more holistic evaluation metrics and methodologies capable of accurately assessing the nuances of commonsense reasoning in AI systems. These metrics should evaluate not only the correctness of responses but also their logical coherence and context relevance.

6 Limitations

While this study offers valuable insights into the role of LLMs and MLLMs in commonsense reasoning, there are some limitations. Firstly, our evaluation is heavily dependent on the selected questions and datasets used for analysis. Despite their diversity, these datasets may not cover all facets of this domain. As a result, the performance and capabilities of Gemini Pro and other models can vary in real-world scenarios or with alternative datasets. Additionally, our analysis is confined to English language datasets, limiting the generalizability of our findings to multilingual contexts, where cultural nuances and linguistic differences are crucial in commonsense reasoning. Finally, our study represents a specific moment in the rapidly evolving landscape of AI, focusing on API-based systems that are subject to change. The introduction of newer models or updates to existing ones might lead to different performance outcomes, highlighting the need for ongoing evaluation and analysis.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

AI@Meta. 2024. Llama 3 model card.

Syeda Nahida Akter, Zichun Yu, Aashiq Muhamed, Tianyue Ou, Alex Bäuerle, Ángel Alexander Cabrera, Krish Dholakia, Chenyan Xiong, and Graham Neubig. 2023. An in-depth look at gemini's language abilities. *arXiv preprint arXiv:2312.11444*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Prajjwal Bhargava and Vincent Ng. 2022. Commonsense knowledge reasoning and generation with pretrained language models: A survey. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12317–12325.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chaoyou Fu, Renrui Zhang, Haojia Lin, Zihan Wang, Timin Gao, Yongdong Luo, Yubo Huang, Zhengye Zhang, Longtian Qiu, Gaoxiang Ye, et al. 2023. A challenger to gpt-4v? early explorations of gemini in visual expertise. *arXiv preprint arXiv:2312.12436*.

Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. 2023. Large language models as zero-shot conversational recommenders. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 720–730.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. In *International Conference on Learning Representations*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401.

Yongfeng Huang, Yanyang Li, Yichong Xu, Lin Zhang, Ruyi Gan, Jiaxing Zhang, and Liwei Wang. 2023. Mvp-tuning: Multi-view knowledge retrieval with prompt tuning for commonsense reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13417–13432.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Xiang Lorraine Li, Adhiguna Kuncoro, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. 2021. Do language models learn commonsense knowledge? *arXiv preprint arXiv:2111.00607*.

Yunxin Li, Longyue Wang, Baotian Hu, Xinyu Chen, Wanqi Zhong, Chenyang Lyu, and Min Zhang. 2023. A comprehensive evaluation of gpt-4v on knowledge-intensive visual question answering. *arXiv preprint arXiv:2311.07536*.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1504–1515.

Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, et al. 2023. Mm-vid: Advancing video understanding with gpt-4v (ision). *arXiv preprint arXiv:2310.19773*.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Mengchen Liu and Chongyan Chen. 2023. An evaluation of gpt-4v and gemini in online vqa. *arXiv preprint arXiv:2312.10637*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*.

OpenAI. 2023. Gpt-4 technical report.

Abu Rayhan, Rajan Rayhan, and Swajan Rayhan. 2023. Artificial general intelligence: Roadmap to achieving human-level capabilities.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473.

Ke Shen and Mayank Kejriwal. 2023. An experimental study measuring the generalization of fine-tuned language representation models across commonsense reasoning benchmarks. *Expert Systems*, page e13243.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. 2019. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*, pages 1–60.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Alexandre Tamborrino, Nicola Pellicanò, Baptiste Pannier, Pascal Voitot, and Louise Naudin. 2020. Pre-training is (almost) all you need: An application to commonsense reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3878–3887.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022a. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yuqing Wang and Yun Zhao. 2023a. Metacognitive prompting improves understanding in large language models. *arXiv preprint arXiv:2308.05342*.

Yuqing Wang and Yun Zhao. 2023b. Tram: Benchmarking temporal reasoning for large language models. *arXiv preprint arXiv:2310.00835*.

Yuqing Wang, Yun Zhao, Rachael Callcut, and Linda Petzold. 2022b. Integrating physiological time series and clinical notes with transformer for early prediction of sepsis. *arXiv preprint arXiv:2203.14469*.

Yuqing Wang, Yun Zhao, and Linda Petzold. 2022c. Enhancing transformer efficiency for multivariate time series classification. *arXiv preprint arXiv:2203.14472*.

Yuqing Wang, Yun Zhao, and Linda Petzold. 2023. Are large language models ready for healthcare? a comparative study on clinical language understanding. *arXiv preprint arXiv:2304.05368*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2550–2575.

Chaoyi Wu, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou, Ziheng Zhao, Ya Zhang, Yanfeng Wang, et al. 2023a. Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. *arXiv preprint arXiv:2310.09909*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. 2023b. Multimodal large language models: A survey. *arXiv preprint arXiv:2311.13165*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1).

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Shuquan Ye, Yujia Xie, Dongdong Chen, Yichong Xu, Lu Yuan, Chenguang Zhu, and Jing Liao. 2023. Improving commonsense in vision-language models via knowledge graph riddles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2634–2645.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019a. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019b. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Yun Zhao, Yuqing Wang, Junfeng Liu, Haotian Xia, Zhenni Xu, Qinghang Hong, Zhiyang Zhou, and Linda Petzold. 2021. Empirical quantitative analysis of covid-19 forecasting models. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 517–526. IEEE.

Wangchunshu Zhou, Ronan Le Bras, and Yejin Choi. 2023. Commonsense knowledge transfer for pre-trained language models. *arXiv preprint arXiv:2306.02388*.

A Commonsense Overview

Commonsense reasoning, a fundamental aspect of human intelligence, facilitates an intuitive understanding and interpretation of the world through basic and often implicit knowledge and beliefs. For instance, it involves understanding that a person carrying an umbrella on a cloudy day likely anticipates rain, or inferring that a closed door in a library signifies a need for quiet. In MLLMs, commonsense reasoning plays a vital role, enabling these models to interact with and interpret human language and visual cues in a manner that mirrors human understanding. In our study, we explore a variety of commonsense reasoning tasks. Definitions for each domain are provided as follows.

General Commonsense. This domain entails an understanding of basic, everyday knowledge about the world, such as recognizing that birds typically fly and fish live in water.

Contextual Commonsense. This domain involves interpreting information within specific contexts, such as understanding that a person wearing a coat and shivering is likely cold.

Abductive Commonsense. This domain is about formulating the most plausible explanations for observations, such as inferring that wet streets are likely due to recent rain.

Event Commonsense. This domain focuses on understanding sequences of events and the causal relationships between them, such as predicting that eating spoiled food can lead to feeling sick.

Temporal Commonsense. This domain involves understanding time-related concepts, such as the fact that breakfast is typically eaten in the morning.

Numerical Commonsense. This domain is about understanding numbers in everyday contexts, such as knowing that a cube has six faces.

Physical Commonsense. This domain concerns understanding the physical world, such as knowing that a glass will break if dropped on a hard floor.

Science Commonsense. This domain involves the application of scientific principles in daily life, such as understanding that water boils at a higher temperature at sea level than in the mountains.

Riddle Commonsense. This domain challenges creative thinking through riddles, such as deciphering a riddle where the answer is "a shadow", requiring lateral thinking to associate intangible concepts with physical entities.

Social Commonsense. This domain involves understanding social interactions, such as recognizing that a person is likely upset if he/she is crying.

Moral Commonsense. This domain deals with evaluating actions based on moral and ethical standards, such as understanding that stealing is generally considered wrong.

Visual Commonsense. This domain involves interpreting and understanding visual information in the context of the physical and social world, such as deducing that a person in a photo is likely running a race if they are wearing a number bib and surrounded by other runners.

B Datasets

We provide an overview of each dataset as well as example questions in Table 5.

C Reasoning Justification Examples

Figure 5 illustrates examples of correct and incorrect answers with their corresponding rationales for Gemini Pro and GPT-3.5.

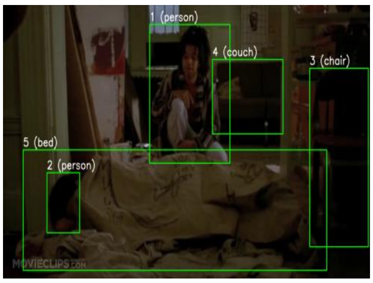
D More Examples of Gemini Pro in Commonsense Reasoning

We provide two more case examples for Gemini Pro to get a deeper understanding of its commonsense reasoning capabilities.

Temporal (TRAM). In the temporal commonsense evaluation (Specialized and Knowledge Reasoning category) using the TRAM dataset, consider the example question: "He also promises to 'come to' him. How long does it take for him to 'come to' him? (A) 100 years (B) in a minute's time (C) a few hours." Lacking sufficient context, especially regarding the identities involved and the meaning of 'come to', Gemini Pro was unable to provide a definitive answer. Gemini Pro's response highlights a significant aspect of its temporal reasoning capabilities. It illustrates the model's reliance on specific contextual information to make accurate temporal judgments. While this cautious approach is prudent to avoid incorrect assumptions, it also signifies a limitation in addressing ambiguous or incomplete information – a frequent challenge in real-world communications. This example underlines the difficulties LLMs encounter in temporal reasoning tasks, especially when faced with scenarios that offer limited or unclear context.

Social (Social IQa). In assessing Gemini Pro's performance in social commonsense reasoning using the Social IQa dataset (Social and Ethical Reasoning category), an interesting scenario was presented: "The people bullied Sasha all her life. But Sasha got revenge on the people. What will the people want to do next? (A) Do whatever Sasha says (B) Get even (C) Flee from Sasha." The correct answer is (C), but Gemini Pro's response is insightful. It chose (B) "Get even" as the most likely option, reasoning that the desire for revenge is a strong motivator and Sasha's actions likely ignited a similar desire in them. Gemini Pro considered (A) as a less likely option, depending on whether Sasha's revenge instilled deep fear and assumed complete submission. The least likely option, according to Gemini Pro, was (C), which it associated with physical harm or an ongoing threat. This response demonstrates Gemini Pro's nuanced understanding of social dynamics and emotional motivations. However, it also highlights a limitation in accurately predicting human reactions in complex social scenarios, where emotional responses might not always follow a logical pattern. This instance reveals the challenges LLMs face in accurately interpreting and responding to intricate social situations, an area that remains challenging for AI systems.

Table 5: Overview of commonsense datasets used in our experiments. "K-Way MC" signifies a multiple-choice response format with K options. Bold text in the "Example Questions" column represents the correct answers.

Dataset	Domain	Answer Type	Example Questions
General and Contextual Reasoning			
CommonsenseQA	general	5-Way MC	Where is a doormat likely to be in front of? (A). facade; **(B). front door**; (C). doorway; (D). entrance porch; (E). hallway.
Cosmos QA	contextual	4-Way MC	Given the context "It wasn't time for my book to be released... I have received about five rejection letters." What may be the reason for your book getting rejected? (A). None of the above choices; (B). I never...; (C). I felt...; **(D). It wasn't finished.**
αNLI	abductive	2-Way MC	Given the beginning of the story: Four Outlaws camped in Blood Gulch, and the end of the story: He arrested them, what is the more plausible hypothesis: (A). They found where the sheriff was; **(B). The sheriff found where they were.**
HellaSWAG	event	4-Way MC	Given the context "A boy in an orange shirt is playing a video game. the scene" and the activity label "Washing face", which of the following endings is the most appropriate continuation of the scenario? (A). changes to safety features; **(B). changes to the game itself**; (C). switches to show...; (D). cuts to the boys...
Specialized and Knowledge Reasoning			
TRAM	temporal	3-Way MC	Then the green ball told the orange ball that blue ball was stupid. How long was the green ball talking to the orange ball? (A). 5 weeks; (B). 24 hours; **(C). 15 seconds**.
NumerSense	numerical	Number	Complete the sentence by filling in <mask> with the most appropriate number. A classical guitar has <mask> strings. → **six**
PIQA	physical	2-Way MC	To reach the physical goal: trees, choose the more sensible solution: (A). provide homes for people; **(B). provide homes for animals.**
QASC	science	8-Way MC	Crabs scavanage and uses dead snail shells for what? (A). RNA; **(B). homes**; (C). making holes; (D). damage; (E). a hosta; (F). Protein; (G). matter; (H). building a nest.
RiddleSense	riddle	5-Way MC	Something very helpful if you want to go gently down a stream. (A). raft; (B). roll down hill; **(C). rowboat**; (D). water; (E). roll over.
Social and Ethical Reasoning			
Social IQa	social	3-Way MC	Kai observed and improved Jan's work. What will Kai want to do next? (A). be effective; **(B). get praise**; (C). get incredible.
ETHICS	moral	2-way MC	After I got the flu virus I called my friends to tell them that I had the virus. Is this acceptable? **(A). acceptable**; (B). unacceptable.
Visual Understanding and Reasoning			
VCR	visual	4-Way MC	 1. What is wrong with Person 2? (A). He is not happy with what is being said to him over the telephone; **(B). He is feeling depressed**; (C). He is high on pot; (D). Someone has pushed him and he's falling. 2. Given the question: What is wrong with Person 2?, and the answer to the question: He is feeling depressed, what is the rationale behind this answer? **(A). Person 1 is talking to him probably trying to cheer him up**; (B). He looks sad and is drinking; (C). He is walking with his head down; (D). He is slumped down on bed and his eyes are closed.

Question: What can be used to find your way if you're lost in the dark?
(A) lamphreys (B) Ultraviolet shielding (C) Light shining into it (D) the speed of hearing **(E) The north star** (F) How to check the temperature (G) Observing tidal waves (H) osculum

Gemini Pro Response:
(E) The north star

Rationale Question: What is the rationale behind the answer to the question?

Gemini Pro Response:
(E) The north star is the most suitable answer based on its unique position in the Northern Hemisphere and its ability to serve as a reliable reference point for determining north, which can be helpful in navigating back on track. Here's why the other options are not suitable:
• (A) Lampreys: These are eel-like fish and have no navigational value.
• (B) Ultraviolet shielding: While it can protect you from UV rays, it won't help you find your way.
•…
• (H) Osculum: This refers to a kiss and wouldn't be relevant to finding your way in the dark.

(a) Example of a correct response and rationale explanation from Gemini Pro.

Question: What is a compass used to navigate?
(A) density (B) lamphreys (C) marine (D) outer space (E) their speed (F) Antarctica (G) non-magnetic planets **(H) the Atlantic**.

GPT-3.5 Turbo Response:
(C) marine

Rationale Question: What is the rationale behind the answer to the question?

GPT-3.5 Turbo Response:
The rationale behind the answer (C) marine for the question about what a compass is used to navigate is based on **the primary and traditional use of compasses**.
Compasses work by aligning with the Earth's magnetic field, specifically pointing towards the Earth's magnetic north pole. This property makes them an invaluable tool for navigation at sea. **When you're out in the open water, it's easy to lose sight of landmarks, and without the help of a compass, it can be challenging to maintain a sense of direction.** By using a compass, mariners can determine their heading and ensure they are on the right course.
Note: The bolded sections in black highlight the generalization of compass use and the lack of specificity in relation to the question's requirement for a particular navigational context, leading to an incorrect interpretation.

(b) Example of an incorrect response and rationale explanation from GPT-3.5 Pro.

Figure 5: Model reasoning correctness justification examples. The sample questions are from the QASC dataset, with the correct answers highlighted in bold red. In example (a), Gemini Pro exhibits methodical reasoning by exclusion, carefully considering all options to reach the most logical conclusion. Conversely, example (b) illustrates GPT-3.5 Turbo's tendency towards unconventional logic, which can result in imaginative yet atypical answers. These instances emphasize the diverse strategies different models apply to commonsense reasoning tasks, revealing their distinct capabilities and limitations in such contexts.