Evaluating the Frontier: A Comparative Analysis of Leading LLMs in Advanced Cognitive Tasks

Anonymous EMNLP submission

Abstract

In recent advancements of natural language processing, Large Language Models (LLMs) have demonstrated unprecedented capabilities in understanding, generating, and interacting. This evaluation explores the ability of proficiency in mainstream LLMs including GPT-3.5, GPT-4.0-turbo, GPT-4.0-vision-preview,Claude-2, and Gemini Pro, extending across mathematics, implicit reasoning, long-context understanding, multi-modal reasoning , and fault-identification abilities. Current literature often underscores the qualitative triumphs of LLMs without quantifying their holistic abilities and limitations in rigorous scenarios. This study aims to fill the gap through a series of methodically crafted evaluations. Our methodology for assessing the capabilities of mainstream LLMs is grounded in a hands-on approach, leveraging the practical functionality of their respective API endpoints. Through this method, we engage in a analysis of their performance, ultimately quantifying their effectiveness using a reliable scores—metrics system that constitute a balanced representation of each model's capabilities.And we've crafted unique prompts and different experiment for different tasks to test each model's strengths and weaknesses effectively.

1 Introduction

The evolution of machine learning has ushered in an era where Large Language Models (LLMs) such as GPT-3.5, GPT-4.0-turbo, Claude-2, and Gemini Pro have emerged.In the realm of Natural Language Processing (NLP), These LLMs models have each pushed the boundaries of what artificial intelligence can achieve. GPT-3.5 laid the groundwork with robust conversational abilities and contextual understanding, while its successor, GPT-4.0-turbo, built on this with improved efficiency and an even more nuanced grasp of complex language tasks, such as summarization and language inference . In turn, Claude-2 focused on optimizing conversational interactions to maximize user engagement and computational sustainability (Lannelongue et al., 2023). Gemini Pro advanced the field into the multi-modal domain, handling not only text but also integrating visual elements to enhance tasks like visual question answering and providing a more comprehensive understanding in scenarios where visuals complement textual information(Singh et al., 2023). Across different NLP tasks, including the challenging realm of machine translation, these models have shown proficiency and adaptability(Jiao et al., 2023).

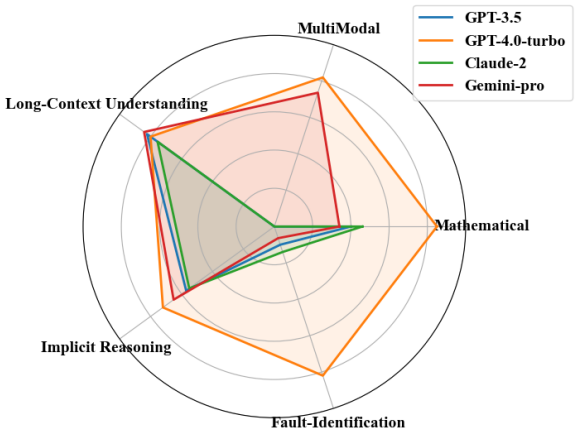

Figure 1: The performance of mainstream LLMs varies in terms of Mathematical MultiModal,Long-context understanding, Implicit Reasoning, and Fault-Consistency. (ps. The max score on Mathematical and Fault-Identification is 20 and 50. GPT-3.5,Claude-2 has no multimodal capabilities,we set it to 0,The multimodal score of GPT-4-turbo comes from GPT-4-vision-preview.)

To understand the limits and possibilities of current LLMs, evaluation necessitates empirical experiments to scrutinize the capabilities of state-of-the-art LLMs. Thus in this paper, we conduct a comprehensive evaluation into the prowess of contemporary LLMS utilizes a diverse assembly of

Dataset	Size	Input	Output	Description
GSM-Hard (Chen et al., 2023)	936	Question	Number	GSM8K with larger number
StrategyQA (Geva et al., 2021)	2780	Question	Yes/No	Multi-hop commonsense reasoning
Squad (Rajpurkar et al., 2016)	3892	Q+Context	Sentences	Long-context understanding
ScienceQA (Lu et al., 2022)	21208	Q+Image	Option	Multi-modal reasoning
FELM (Chen et al., 2023)	846	Question	Yes/No	Factuality segments

Table 1: The datasets used in the evaluation.

datasets, each meticulously curated to interrogate specific facets of these LLMS. The datasets include **GSM-Hard (Gao et al., 2022)**, challenging the mathematical logic and problem-solving abilities of the models with intricate numerical puzzles. **StrategyQA (Geva et al., 2021)** examines their aptitude for abstract reasoning and inferential logic through indirectly-framed questions. With **Squad (Rajpurkar et al., 2016)**, we evaluate the long-context understanding ability of the models, assessing their proficiency in following and contributing to complex, domain-specific conversations. In the realm of multi-modal ability, **ScienceQA (Lu et al., 2022)** serves as a gauge for the models' multimodal capabilities in parsing and conveying complex principles spanning various scientific fields. Finally, the **FELM (Chen et al., 2023)** dataset focuses on factuality across diverse domains, spanning from world knowledge to math and reasoning.

Together, these datasets create a multifaceted testing ground, providing a rigorous benchmark for LLMs' performance across a broad spectrum of mathematics, implicit reasoning, long-context understanding, multi-modal reasoning, and fault-identification abilities. Figure 1 indicates that today's mainstream large models have strong abilities in multi round dialogue and implicit reasoning. Meanwhile, it is evident that the GPT-4.0 turbo still holds a leading position in various tasks, demonstrating its error detection function that distinguishes it from other models. Google's Gemini Pro model is a strong competitor to GPT-4.0 turbo in terms of implicit reasoning and multimodal ability. In the future, mainstream large-scale language models should focus on improving graphic understanding and processing large-scale numerical calculations to further enhance their capabilities. The prospect of large-scale language models is constantly evolving, and GPT-4-Turbo is in a leading position while facing competition and improvement opportunities.

In summary, we have conducted a comprehensive assessment of the capabilities of current large language models. By using carefully crafted prompts and interacting with official API endpoints, we are able to objectively measure their performance on specific tasks and datasets. This experimental approach not only simulates real-world scenarios but also provides standardized data for us to gain insights into the abilities and limitations of these models.

2 Experimental Setup

2.1 Dataset

In our comprehensive evaluation, we rigorously assess the Large Language Models (LLMs) across various tasks. For mathmatical reasoning, We select the wildly-used GSM-Hard(Gao et al., 2022) train set to examines numerical and logical reasoning through complex math problems. For implicit reasoning, We use StrategyQA(Geva et al., 2021) train set inferential and strategic thinking required to interpret implicit queries. For long-context understanding, the squad(Rajpurkar et al., 2016) dataset evaluates the models' aptitude for long-context understanding. For multi-modal ability ,ScienceQA(Lu et al., 2022) gauges the multi-modal reasoning across scientific field , and for fault-identification ablility, the FELM(Chen et al., 2023) set assesses about math,commonsense reasoning, wikipedia and science module. Table 1 presents the statistics of the datasets we used.

2.2 Mainstream LLMs system

In the exploration of the capabilities of mainstream LLMs, several models from leading organizations stand out. OpenAI's GPT-3.5[1], also known as ChatGPT, and the subsequent advancement, GPT-4.0[1] are paradigms of AI conversational prowess, with improvements in comprehension and task-specific performance. Anthropic's Claude-2[2] emerges as a strong contender, emphasizing ethical AI development and exhibiting high adeptness in contextually nuanced interactions. Google's contribution,

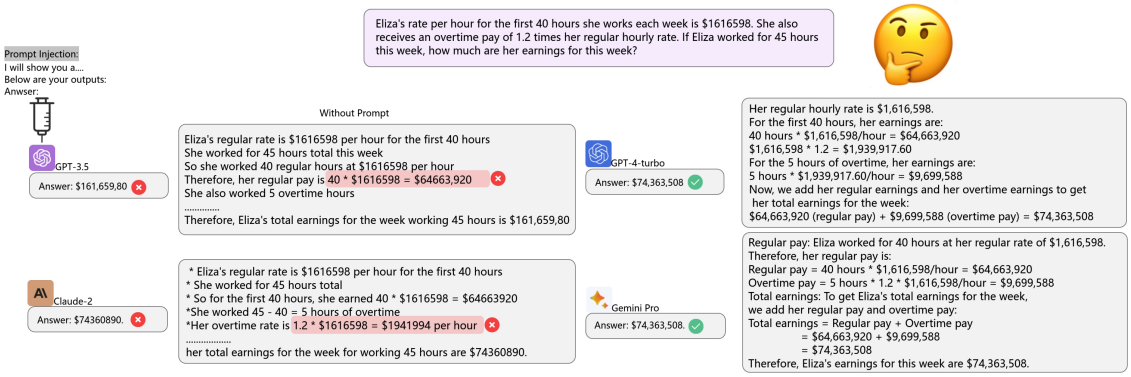

Figure 2: Use the effect of Prompt to improve the stability of model output and make it easier to judge.

Gemini Pro[4], stands as a testament to their expansive data harnessing capabilities, bringing a potent LLM into the arena noted for its versatility and accuracy.We assess these LLMs' performance using official API [1] [2] [3].

2.3 Evaluation Method

The NLP large model exhibits a high degree of diversity in generating answers, providing the model with rich expressive capabilities and flexibility. However, this diversity poses a challenge for our evaluation. To facilitate the assessment of these answers, we adopt a method based on the "prompt engineering" technology to standardize the answer generation. Through clear prompts, this method can guide the model to produce answers related to specific themes or formats, and improve the assessability and controllability of the answers. As shown in the Fig.2 , we can specify "Below are you outputs: Answer:" to make the model generate standardized answers. Subsequently, we can use string processing tools to extract the model's answers for convenient evaluation. This approach helped us create conditions that would not only streamline the processing of the data but also faithfully reflect the inherent capabilities of the LLMs.

Mathematical Evaluation In our Mathematical Evaluation, we utilized the training set of the challenging GSM-Hard dataset, renowned for its very large number and extensive computational demands. To probe the mathematical acuity of mainstream LLMs, we crafted specialized prompts introduced in Appendix B, intended to elicit model responses that could be directly compared to solutions within the GSM-Hard dataset. This comparative analysis provided us with a clear benchmark of the LLMs' performance in mathematical tasks.

Implicit Reasoning Evaluation For the implicit reasoning evaluation, we employed the StrategyQA dataset, which is challenging that demands logical reasoning abilities. This dataset includes not only questions but also relevant term and their descriptions, requiring the models to demonstrate an understanding of objective facts with description and their training sample. This comparison provides evaluation of the LLMs' performance on implicit reasoning tasks.

Long-Context Understanding Evaluation To evaluate the long-context understanding of LLMs, we utilized the Squad dataset, which is composed of long context that test a model's ability to process and respond to extended narratives (Zhou et al., 2020). We then compared these responses with the benchmark answers in the Squad dataset and employed evaluation function to analyze the semantic similarity between the sentences, .

Multi-Modal Reasoning Evaluation The multi-modal reasoning evaluation was conducted on ScienceQA dataset, which presents a comprehensive challenge requiring the integration of visual processing with textual understanding. We constructed prompts specifically designed to push the limits of mainstream LLMs, pushing them to synthesize information from both the images provided and the accompanying text, thus effectively quantifies the models' capabilities in multi-modal contexts.

Fault-Identification Evaluation Fault identification evaluation was conducted on FELM dataset due to its segments of long context. The dataset's

[1] https://platform.openai.com/docs/api-reference
[2] https://developer.poe.com/
[3] https://ai.google.dev/docs

split response into many segments which are correct or false, with this ask LLMs to judge the fault among many segments, a critical ability for effective reasoning (Hendrycks et al., 2021). This mimics real-world scenarios in which LLMs could 'cure' themselves one day.

3 Experiments

3.1 Mainstream LLMs Performance on Mathematical

In this part of experiments,we engaged GPT-3.5-turbo, GPT-4-turbo, Claude-2, and Gemini Pro in a series of one-on-one question-and-answer sessions. These sessions were designed to test each model's ability to handle mathematical problems involving large numbers, using raw prompts to standardize the responses with these LLMs.

The raw prompt methodology was selected to reflect a realistic use case, where users may pose questions directly and expect accurate numerical computations in return, without providing additional context or computational aids. This approach is crucial for evaluating the practical utility of LLMs in real-world scenarios where users rely on conversational AI for immediate and precise answers to complex queries.

GSM-Hard	
	Acc.
GPT-3.5-turbo	8.6
GPT-4-turbo	**16.9**
Claude-2	9.3
Gemini Pro	6.8

Table 2: Mainstream LLMS performance on mathematica,**Bold** values indicate the best scores across different LLMs.

In experiments on GSM hard datasets, we found that the GPT-4-turbo model demonstrated the strongest mathematical ability in the tested LLM. However, these experiments also revealed a common weakness of mainstream LLMs: they are prone to hallucinations when dealing with mathematical problems with large values, leading to incorrect answers. For example, consider the following mathematical problem: calculation 2^{1000}.The correct answer is a very large number, approximately $1.0715086071862673 \times 10^{301}$. However, due to the magnitude of the values, mainstream LLMs (including GPT-4 turbo) may experience hallucinations, leading to incorrect answers. Most models may give an incorrect answer, such as $2^{1000} = 0$.

This example demonstrates the problem of hallucinations that mainstream LLMs are prone to when dealing with mathematical problems with large numerical values. Although GPT-4 turbo performs well in mathematical abilities, there are still challenges in dealing with this specific type of problem. This discovery provides valuable guidance for further improving the mathematical reasoning ability of the model.

3.2 Mainstream LLMs Performance on Implicit Reasoning

In this section,we evaluated the performance of GPT-3.5-turbo, GPT-4-turbo, Claude-2, and Gemini Pro. The task for the LLMs was to select the most reasonable objective fact option from a set of multiple choices, a test designed to assess their grasp of commonsense accuracy. This evaluation was conducted using the StrategyQA dataset, which is specifically tailored to measure how well models can handle questions that require implicit reasoning with known facts.

StrategyQA			
	Pre.	Rec.	F_1
GPT-3.5-turbo	41.1	**93.8**	57.2
GPT-4-turbo	**52.1**	90.0	**66**
Claude-2	39.7	93.6	55.7
Gemini Pro	**52.1**	86.7	65.1

Table 3: Mainstream LLMS performance on implicit reasoning ,**Bold** values indicate the best scores across different LLMs.

In the domain of implicit reasoning, mainstream LLMs have demonstrated a notable proficiency, adeptly navigating the subtleties of implied meaning and inference. GPT-4-turbo and Gemini Pro, in particular, stand out for their advanced ability to parse and reason through the nuanced undercurrents of language that go beyond explicit factual information. This suggests a sophisticated level of understanding that is essential for complex cognitive tasks.

3.3 Mainstream LLMs Performance on Long-Context Understanding

In assessing mainstream LLMs' aptitude for processing long-context information, we applied a evaluation metric. This method was specifically

selected to compute values for Average (Avg), Exact Match (EM), and the F_1 scores, which together offer a multifaceted view of an LLM's performance in the context of Long-Context Understanding. Responses from the LLMs, prompted by the intricate and lengthy texts in the Squad dataset, were evaluated against these metrics. Through the evaluation method, we were able to determine how accurately and completely the models could comprehend and recall details from extended text passages.

	Squad		
	Avg.	Em.[A.2]	F_1[A.1]
GPT-3.5-turbo	69.7	57.0	82.3
GPT-4-turbo	66.1	52.3	80.0
Claude-2	63.8	50.6	75.5
Gemini Pro	**75.2**	**66.1**	**84.2**

Table 4: Mainstream LLMS performance on Long-Context Understanding ,**Bold** values indicate the best scores across different LLMs.

During our experiments focused on long-context understanding, notable models such as GPT-3.5-turbo, GPT-4-turbo, Claude-2, and Gemini Pro were evaluated. Notably, Gemini Pro emerged as the outlier, achieving a breakthrough victory with a slight advantage over other mainstream LLMs. On the whole, the current suite of mainstream LLMs displayed commendable performance in this area. With all models scoring an F_1 above 80, the results underscore a significant level of proficiency in comprehending lengthy texts. This overarching success indicates that these advanced LLMs are increasingly adept at navigating complex narratives and maintaining consistency over longer passages, a testament to the rapid evolution of language models in understanding and processing extended content.

3.4 Mainstream LLMs Performance on Multi-Modal Reasoning

We explored the capabilities of LLMs in processing and integrating information from both text and images. The models involved in this evaluation were GPT-4-vision-preview and Gemini Pro, both of which are equipped with multi-modal capabilities that allow them to interpret and analyze visual content in conjunction with textual data.

Utilizing the ScienceQA dataset, which includes questions that require an understanding of both the provided text and associated images, we tasked the LLMs with selecting the correct answer from a set of multiple-choice options.

	ScienceQA
	Acc.
GPT-4-vision-preview	**82.4**
Gemini Pro	73.6

Table 5: Mainstream LLMS performance on Multi-Modal Reasoning ,**Bold** values indicate the best scores across different LLMs.

The raw $prompt^B$ approach was employed to simulate a straightforward interaction with the models, without any pre-processing or additional hints that could influence their performance.GPT-4-vision-preview, in particular, stands out, illustrating the evolving competence of mainstream LLMs in synthesizing insights from images alongside text, which is a notable advancement in the realm of multi-modal artificial intelligence.

3.5 Mainstream LLMS Performance on Fault-Identification

Our research has found that the FELM dataset reveals the excellent ability of GPT-4 turbo in inference error detection. For example, when faced with a large number of text paragraphs, GPT-4 turbo can accurately identify logical fallacies, which makes it in sharp contrast to other mainstream models such as Gemini Pro in this field. For example, consider the following sentence: "All cats can fly, and Garfield is a cat, so Garfield can fly." Logical fallacy: There is an obvious logical error in this sentence because not all cats can fly.

	FELM		
	Pre.	Rec.	F_1
GPT-3.5-turbo	31.8	7.1	11.6
GPT-4-turbo	**55.0**	**33.0**	**41.2**
Claude-2	37.9	7.5	12.5
Gemini Pro	30.2	2.0	3.2

Table 6: Mainstream LLMS performance on Fault-Identification ,**Bold** values indicate the best scores across different LLMs.

GPT-4 turbo can accurately capture this logical error and point out the correctness of the statement. This powerful reasoning ability enables GPT-4 turbo to perform excellently in ensuring factual consistency. This discovery provides strong support for further improving the quality and accuracy of natural language processing models.

4 Conclusion

In conclusion, the performance of mainstream Large Language Models (LLMs) has shown that while they excel in certain areas of natural language understanding, they still face significant challenges in tasks that require advanced numerical reasoning, error detection, and multi-modal data integration.

The current state of LLMs indicates that they are capable of impressive feats in implicit reasoning and language-based tasks. However, their ability to process large numbers accurately, implicit reasoning, and update their knowledge base with new information is limited.

Despite the introduction of Google's Gemini Pro, GPT-4-turbo remains the reigning champion in the realm of Large Language Models (LLMs). GPT-4-turbo's performance in areas such as mathematical reasoning, multi-modal abilities, and fault identification has solidified its position as the leading model. While Gemini Pro represents a significant advancement in AI, current assessments indicate that it has not surpassed the performance threshold set by GPT-4-turbo in these key areas. Thus, GPT-4-turbo continues to maintain its dominance in the LLMs' landscape.

Limitations

we must acknowledge a critical issue that pervades current Large Language Models (LLMs): output instability. The responses generated by LLMs can vary with each invocation, even when presented with identical prompts. This inconsistency poses a challenge for researchers attempting to evaluate the models' performance, as it introduces a level of variance that can compromise the reliability of results. Repeatability is the bedrock of empirical analysis; thus, the fluctuating nature of LLM outputs can undermine the accuracy of systematic testing.

Further compounding this issue is the fact that LLMs, as of the current state of technology, are static in their knowledge base. They are trained on datasets that are a snapshot in time and thus lack the ability to acquire information post-training. Consequently, these models do not have the facility to incorporate or reflect the most current events, discoveries, or consensus changes that occur after their last update. This limitation restricts the models' ability to provide insights into recent developments and reduces their relevance, particularly in fields where current information is critical.

In light of these limitations, the evaluation of LLMs should be viewed with careful consideration of these constraints. The instability in output and the static knowledge base highlight the need for continued development in model architecture, training techniques, and data integration methods to enhance the robustness, repeatability, and currency of the information LLMs offer.

References

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. Felm: Benchmarking factuality evaluation of large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2022. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics (TACL)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. ParroT: Translating during chat using large language models tuned with human translation and feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15009–15020, Singapore. Association for Computational Linguistics.

Loïc Lannelongue, Hans-Erik G Aronson, Alex Bateman, Ewan Birney, Talia Caplan, Martin Juckes, Johanna McEntyre, Andrew D Morris, Gerry Reilly, and Michael Inouye. 2023. Greener principles for environmentally sustainable computational science. *Nature Computational Science*, 3(6):514–521.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, page arXiv:1606.05250.

Janvijay Singh, Vilém Zouhar, and Mrinmaya Sachan. 2023. Enhancing textbooks with visuals from the web for improved learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11931–11944, Singapore. Association for Computational Linguistics.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. KdConv: A Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7098–7108, Online. Association for Computational Linguistics.

Appendix

A Details of Evaluation Function

A.1 F_1

The F_1 evaluation metric is the harmonic mean of Precision and Recall, used to evaluate the performance of classification models. It combines the advantages of precision and recall, which can better reflect the correctness and accuracy of the model. Among them, accuracy refers to the proportion of samples that are predicted to be positive but are actually positive; Recall rate refers to the proportion of actual positive samples that are predicted to be positive.

For quantitative prediction tasks, for example, assuming a classification model has 80 positive samples predicted out of 100, and 70 actual positive samples; There are 20 predicted negative samples and 30 actual negative samples. So, the accuracy of the model is $\frac{70}{80} = 0.875$, and the recall rate of the model is $\frac{70}{100} = 0.7$. So, the F_1 evaluation index of this model is: $2 \times \frac{(0.875 \times 0.7)}{(0.875+0.7)} = 0.793$.

For natural language processing tasks such as machine reading comprehension, the calculation of F_1 Index is similar to quantitative prediction tasks. $F_1 = 2 \times \frac{P \times R}{P+R}$. Where the calculation of P and R is based on the model's prediction and the number of tokens shared by the ground truth. Here's an example: for a prediction, ground truth is"Today I ate rice" and the prediction is"Today I ate noodles with her". Obviously, the number of shared tokens N_{share} is 3. So $P = \frac{N_{share}}{N_{label}} = 0.75$, $R = \frac{N_{share}}{Npred} = 0.5$. And the final calculation for F_1 is 0.6.

```python
def mixed_segmentation(in_str, rm_punc=False):
    in_str = str(in_str).lower().strip()
    segs_out = []
    temp_str = ""
    sp_char = ['-',':','_','*','^','/','\\','~','`','+','=',
               '，','。','：','？','！','“','”','；','’','《','》','……','·','、',
               '「','」','（','）','－','～','『','』']
    for char in in_str:
        if rm_punc and char in sp_char:
            continue
        if re.search(r'[\u4e00-\u9fa5]', char) or char in sp_char:
            if temp_str != "":
                ss = nltk.word_tokenize(temp_str)
                segs_out.extend(ss)
                temp_str = ""
            segs_out.append(char)
        else:
            temp_str += char
def calc_f1_score(answer, prediction):
    f1_scores = []

    ans_segs = mixed_segmentation(answer, rm_punc=True)
    prediction_segs = mixed_segmentation(prediction, rm_punc=True)
    lcs, lcs_len = find_lcs(ans_segs, prediction_segs)
    if lcs_len == 0:
        return 0
    precision     = 1.0*lcs_len/len(prediction_segs)
    recall        = 1.0*lcs_len/len(ans_segs)
    f1            = (2*precision*recall)/(precision+recall)
    f1_scores.append(f1)
    return max(f1_scores)
```

Listing 1: Python code for calculating F_1 score between two sentences

A.2 Em (Exact Match)

The calculation of the Exact Match indicator is: if the predicted answer is completely consistent with the ground truth, then EM is 1; otherwise, it is 0.

```python
def remove_punctuation(in_str):
    in_str = str(in_str).lower().strip()
    sp_char = ['-',':','_','*','^','/','\\','~','`','+','=',
               '，','。','：','？','！','“','”','；','’','《','》','……','·','、',
               '「','」','（','）','－','～','『','』']
    out_segs = []
    for char in in_str:
        if char in sp_char:
            continue
        else:
            out_segs.append(char)
    return ''.join(out_segs)
def calc_em_score(answer, prediction):
    em = 0
    ans_ = remove_punctuation(answer)
```

```
    prediction_ = remove_punctuation(
    prediction)
    if ans_ == prediction_:
        em = 1
    return em
```

Listing 2: Python code for calculating exact match score between two sentences

B Experiments Example

For the robust evaluation of mainstream LLMs, we developed a suite of specialized prompts tailored to the unique characteristics and presumed capabilities of each model under consideration.

Mathematical

I will show you a math question. Your task is to answer the math question.
Please generate using the following format:
Answer: Your answer to the question.Please only output the digital answer, no more details.

Here is one example:
Question:Kylar went to the store to buy glasses for his new apartment.
One glass costs $5, but every second glass costs only 60% of the price.
Kylar wants to buy 5364765 glasses. How much does he need to pay for them?

Below are your outputs:
Answer: 21459061

Below are my inputs:
Question:A robe takes 2287720 bolts of blue fiber and half that much white fiber.
How many bolts in total does it take?

Answer: 3431580

Implicit Reasoning

I will show you a term ,a description and a question.
Your task is to judge the question is true or false.
The term is the key word of the question.
The description is the description of the term which is ground-truth.
The question is the question you need to judge.

If the question is true, output true, otherwise output false.
Please generate using the following format:
Answer: Your answer to the question.Please only output true or false, no more details.

Here is one example:
Term:Swastika
Description:a geometrical figure and an ancient religious icon in the cultures of Eurasia and 20th-century symbol of Nazism.

Question:Did the Hopi Indians use a symbol that was similar to the swastika ?

Answer: true

Long-Context Understanding

I will show you a context and a question.Your task is according to the context to generate the answer to the question.
Please generate using the following format:
Answer: Your answer to the question. no more details.

Here is one example:
context:Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary.
Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend "Venite Ad Me Omnes".
Next to the Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection.
It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858.
At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.....

Question:To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?

Answer: Saint Bernadette Soubirous

MultiModal

I will show you a image(could be none),a question,
a choices(a array contain different answer) and
some of these questions have a hint.

Your task is choose a choice in the choices and output the index of the choice in the choices array(start by 0).
The question will be about the image and the choices will contain only one correct answer.
The hint will help you to choose the correct answer.
Please generate using the following format:
Answer: Your answer to the question.
Please only output the correct answer index of the choices, no more details.

Here is one example:
question:Which of these states is farthest north?

choices:"West Virginia", "Louisiana", "Arizona", "Oklahoma"

Answer: 0

(PS. Only output the ids of the choices.)

Fault-Identification

I will show you a question and a list of text segments.
All the segments can be concatenated to form a complete answer to the question.
Your task is to determine whether each text segment contains factual errors or not.

Please generate using the following format:
Answer: List the ids of the segments with errors (separated by commas).
Please only output the ids, no more details.
If all the segments are correct, output "ALL_CORRECT".

Here is one example:

Question: What is the diffusion model in computer science?

Segments:
1. In computer science, the diffusion model is a mathematical model used to simulate the spread of information or data through a network or system.
2. It is often used to study phenomena such as the spread of viruses, the adoption of new technologies, or the dissemination of information in social networks.

Answer: 2,3,4,5,6

(PS. Only output the ids of the segments which have error in it)

Figure 3: The Example of Prompts