When LLMs Meet Cunning Questions:
A Fallacy Understanding Benchmark for Large Language Models

Anonymous ACL submission

Abstract

Recently, Large Language Models (LLMs) have made remarkable evolutions in language understanding and generation. Following this, various benchmarks for measuring all kinds of capabilities of LLMs have sprung up. In this paper, we challenge the reasoning and understanding abilities of LLMs by proposing a **F**a**L**lacy **U**nderstanding **B**enchmark (**FLUB**) [1] containing cunning questions that are easy for humans to understand but difficult for models to grasp. Specifically, the cunning questions that FLUB focuses on mainly consist of the tricky, humorous, and misleading questions collected from the real internet environment. And we design three tasks with increasing difficulty in the FLUB benchmark to evaluate the fallacy understanding ability of LLMs. Based on FLUB, we investigate the performance of multiple representative and advanced LLMs, reflecting our FLUB is challenging and worthy of more future study. Interesting discoveries and valuable insights are achieved in our extensive experiments and detailed analyses. We hope that our benchmark can encourage the community to improve LLMs' ability to understand fallacies.

1 Introduction

Large Language Models (LLMs) have shown great abilities to understand human languages, including information extraction (Wei et al., 2023), text correction (Li et al., 2023), complex reasoning (Bang et al., 2023), etc. Researchers have constructed numerous question-answering benchmarks to test the capabilities of LLMs in various aspects. By using collected questions to interact with LLMs, researchers can analyze the behavior of LLMs to compare the performance of different LLMs and study how to further improve LLMs.

Although many LLM benchmarks have sprung up, we believe that existing benchmarks are not challenging enough to truly measure the human-like intelligence of LLMs. In particular, we are still wondering whether LLMs can understand cunning questions that may contain misleading, wrong premise, intentional ambiguity, and so forth, considering that almost all LLMs are trained on "cleaned" and "correct" corpora. Therefore, we build a **F**a**L**lacy **U**nderstanding **B**enchmark (**FLUB**) to challenge LLMs for solving these problems.

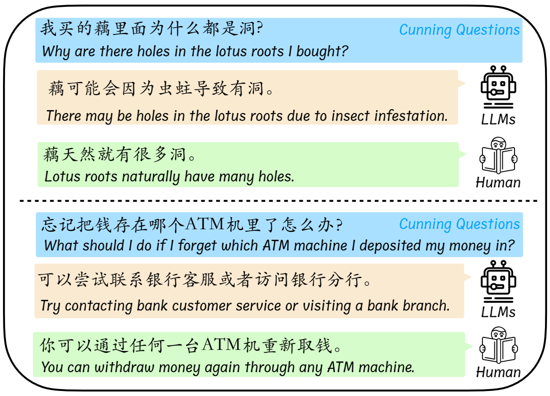

Figure 1: The examples of how LLMs and humans perform when faced with cunning questions. The LLM we use is ChatGPT-3.5 on Jan 23, 2024.

Figure 1 shows the running examples from FLUB. From these cases, we directly feel the different behaviors of LLMs and humans when facing cunning questions. In the first example, LLMs ignore the common sense that the lotus root itself has many holes in its structure and fall into the trap of the cunning question, wrongly judging that the holes in the lotus root are caused by insect infestation. In the second example, LLMs fail to see the logic that depositing money into random ATMs does not create problems and therefore give an answer that seems reasonable but is absurdly laughable. In fact, these cunning questions for LLMs are very easy to handle for human intelligence. **Therefore, it is very urgent and meaningful to construct a benchmark composed of cunning questions to**

[1] We will release our dataset after peer review.

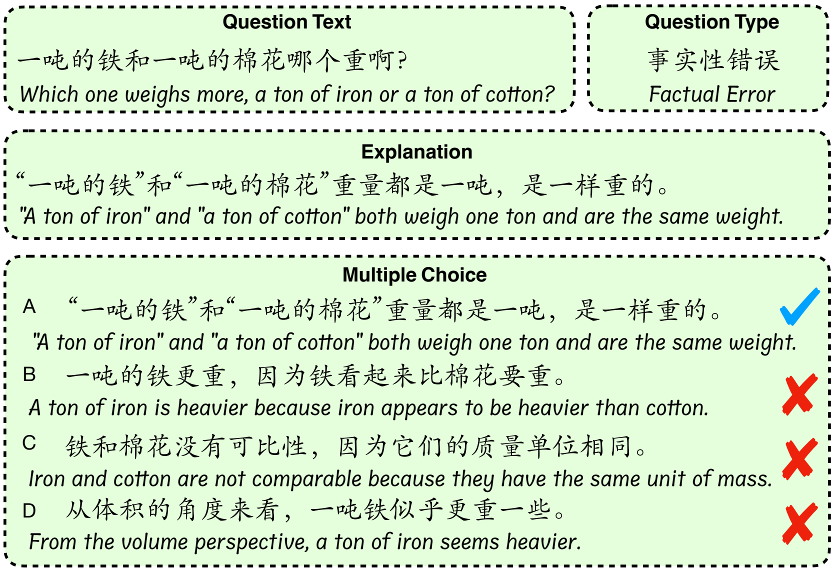

Figure 2: The data annotation example of FLUB.

evaluate and thereby promote the improvement of LLMs' fallacy understanding capabilities.

Inspired by the above motivation, we collect real cunning questions as our raw data from a famous Chinese online forum, the "Ruozhi Bar" (retard forum) [2]. This forum is popular for its cunning and unreasonable posts, which are generally easy for humans to understand but challenging for LLMs. The characteristics of the posts contained in this forum are consistent with our research motivation, so choosing it as the data source well supports FLUB's evaluation of LLMs' fallacy understanding ability. After data cleaning and annotating of question types, FLUB has 8 fine-grained types of cunning questions and most of the questions in FLUB fall into two types of fallacy, namely, faulty reasoning and word game. Moreover, we also manually annotated one correct answer (i.e., the explanation of the question) and three confusing wrong answers for each question in FLUB, as shown in Figure 2.

Based on our constructed FLUB and its annotation information, we design three tasks with increasing difficulty to test whether the LLMs can understand the fallacy text and solve the "cunning" questions. Specifically, (1) **Answer Selection**: The model is asked to select the correct one from the four answers provided by FLUB for each input question. (2) **Question Type Classification**: Given a cunning question as input, the model is expected to directly identify its fallacy type defined in our scheme. (3) **Question Explanation**: We hope the model sees a cunning question and intelligently generates a correct explanation for the question, just like humans, without falling into its trap.

In our experiments, we select representative and advanced LLMs to be evaluated on FLUB. Our empirical study reveals: (1) LLMs are very poor in their ability to perceive fallacy types in cunning questions. (2) For a specific task, LLMs with larger parameter sizes do not always perform better. (3) There is a close relationship between the Answer Selection task and the Question Explanation task, and the interaction between them is critical to promoting the understanding of fallacies in LLMs. (4) On FLUB, the widely used Chain-of-Thought and In-context Learning techniques deserve further improvement and research. We believe that our proposed FLUB and all our findings are crucial for LLMs to comprehend the fallacy and handle cunning questions in the real world.

2 The FLUB Benchmark

2.1 Benchmark Construction

Data Collection We collect raw cunning question data from "Ruozhi Bar" in Baidu Tieba [3]. "Ruozhi Bar" is one of the most famous online forums in the Chinese internet community, and people often post some interesting or "silly" questions on it just for fun. We find that many of the posts on this forum are tricky questions or brain-teaser-like texts, which is exactly in line with our purpose of using cunning questions to challenge

[2] https://tieba.baidu.com/f?kw=%E5%BC%B1%E6%99%BA&ie=utf-8

[3] https://tieba.baidu.com

Question Type	# of Samples	Example
错误类比 False Analogy	11	很多人出门后担心刚刚没有关门，为什么进门后不担心刚刚没有开门? Many people worry about forgetting to close the door when they leave home. Why don't they worry about whether they have opened the door when they come in?
冷笑话 Lame Jokes	44	忘记把钱存在哪个ATM机里了怎么办? 银行好几台ATM机，还长得都一样。 What should I do if I forget which ATM I deposited money into? The bank has several ATMs, and they all look the same.
字音错误 Phonetic Error	5	因为美国队长，小明每次在美国排队都要排一个多小时。 Because of Captain America (also read as "long queues in America" in Chinese), Xiao Ming has to wait over an hour whenever he queues in the U.S.
歧义 Ambiguity	35	语文老师说我写的句子是病句，我应该给这个病句吃头孢，还是打点滴呢? My teacher said the sentence is grammatically incorrect ("sick sentence" in Chinese). Should I give this sentence some antibiotics or administer an IV drip?
悖论 Paradox	29	"凡事无绝对"这句话过于绝对。 The phrase "Nothing is absolute" is too absolute.
事实性错误 Factual Error	12	一吨的铁和一吨的棉花哪个重啊? Which one weighs more, a ton of iron or a ton of cotton?
推理错误 Reasoning Error	445	根据我在养老院的调查数据，我国的人口老龄化已经相当严重了。 According to my survey data from nursing homes, the aging of the population in our country has become quite severe.
文字游戏 Word Game	239	人类70%是水，所以10个人里有7个人是水伪装成的人! 70% of the human body is water, so 7 out of 10 people are water disguised as humans!
未分类 Undefined	24	在高速路的服务区开酒吧有可行性吗? Is it feasible to open a bar at a highway service area?

Table 1: Question types of FLUB and corresponding examples.

LLMs, so we decide to utilize this forum as our data source. As a result of automatic crawling, we initially collect 9,927 candidate posts, including the title, body text, and the first comment of the post. Notably, according to the Baidu Bar agreement [4], the data on Baidu Tieba can be used for academic research free of charge and without liability.

Data Cleaning We employ annotators to manually filter out irrelevant posts that do not present tricky or cunning questions. Since the collected original posts contain irrelevant content such as links and pictures, we also require data annotators to extract the fallacious and illogical contents from the title and body text of each post and rewrite them into a complete question. Besides, it is worth noting that we carefully ensure that the questions remained in FLUB are ethical texts. This process includes user information anonymization, sensitive information removal, and filtering of impolite posts. In total, we obtain 844 data samples to form FLUB.

Data Annotation To ensure the annotation quality of FLUB and taking into account the characteristics of our study, when we select annotators, our criteria for selecting annotators is that the person must be a native Chinese speaker and have a bachelor's degree. The detailed annotation workflows for each type of information are as follows:

1. **Question Type Annotation**: To do this kind of annotation, we first define 8 question types within the collected questions along with their corresponding examples. Subsequently, each data sample is processed by three junior annotators, who are required to select an appropriate question type for the sample. We achieve the initial annotation results based on the voting results among three annotators. The initial annotation results become the final annotation information after being reviewed by the senior annotator (and modified if necessary). The schema for the types is shown in Table 1.

2. **Correct Explanation Annotation**: We assign two junior annotators to write the explanation or answer for each sample independently. We ask them to try to explain the given question in a detailed, objective, and unambiguous way. The senior annotator then selects (and modifies if necessary) the more suitable text written by the two junior annotators.

[4] https://baike.baidu.com/item/%E8%B4%B4%E5%90%A7%E5%8D%8F%E8%AE%AE/8397765

3. **Wrong Candidates Annotation**: This part annotation is to obtain the wrong candidate answer that may be likely to be answered incorrectly for each question. We assign three junior annotators for each sample and require each of them to write three different incorrect answers based on their understanding of the question. Particularly, we emphasize to each junior annotator that the three different wrong answers they write should ensure diversity and resemble as much as possible the answers that LLMs can easily produce. For each sample's nine initial incorrect answers, the senior annotator selects the three most challenging sentences as the final wrong candidates.

It is worth mentioning that we have prepared sufficient and representative samples for annotators to learn and pre-annotate to ensure that they fully understand the information we want to annotate before they officially start annotation. Our entire annotation process lasted 2 weeks. Other annotation details are presented in Appendix A

2.2 Dataset Analysis

Data Distribution As shown in Table 1, most data in FLUB belong to the types of reasoning errors and word games. This is because these two types of questions appear widely in "Ruozhi Bar" forum whose purpose is to challenge human intelligence. A large number of cunning questions involving reasoning errors and word games ensure that FLUB is challenging enough. Besides, we observe that some types of questions are relatively rare, such as phonetic errors. In fact, this is because our data come entirely from the real world and are all carefully constructed by netizens. Cases of cunning questions caused by phonetic errors are indeed rare in the real world. From another perspective, the data distribution also reflects that FLUB is real and close to human lives, so it can better measure the intelligence gap between humans and LLMs.

Annotation Quality Since question-type annotation is essentially a classification process performed by multiple annotators, we analyze the annotation quality of this information. Specifically, we calculate Fleiss' Kappa (Falotico and Quatto, 2015) to reflect the three junior annotator's Inter-Annotator Agreement (IAA). Our final obtained Fleiss Kappa result is greater than 0.767, which shows that our annotation results have excellent consistency and quality (Landis and Koch, 1977).

给你输入一个句子或问题，其中存在不合理或幽默之处。另外给出四个选项，你需要选出最能准确描述给定句子或问题的不合理或幽默之处的一个选项，并说明选择该选项的理由。

你的输出必须严格遵循以下格式：
分析：<简要地分析四个选项中哪一个准确描述给定句子或问题的不合理或幽默之处，说明选择该项的理由>
答案：<只能输出“A，B，C，D”中的一个>

以下是输入：
{sentence}

选项：
{options}

Task 1

给你输入一个句子或问题，其中存在不合理或幽默之处。你需要用一句话解释其中的不合理或幽默之处，然后从“候选分类”中选出一个最适合该句子或问题的类别，并说明理由。

候选分类：{candidates}

你的输出必须严格遵循以下格式：
解释：<用一句话解释输入句子或问题的不合理或幽默之处>
理由：<解释符合“候选分类”中某一类别的理由>
分类：<从“候选分类”中选出的类别>

以下是输入：
{sentence}

Task 2

给你输入以下的句子，其中存在不合理或幽默之处。请在三句话以内简要地解释该句子的不合理或幽默之处。

{sentence}

Task 3(a)

请你在三句话以内简要地回答下面的问题：

{sentence}

Task 3(b)

Figure 3: Our designed prompts for FLUB. Task 3(a) is for the questions that are not expressed in the form of inquiries. Task 3(b) is for inquiries. Note that here we show Chain-of-Thought prompts for Task 1 and Task 2, and their prompts without Chain-of-Thought are in Appendix B. The English translations of our prompts are also in Appendix B.

2.3 Benchmark Task Setups

To evaluate the fallacy understanding ability of LLMs, we design three benchmark tasks on FLUB: Answer Selection, Question Type Classification, and Question Explanation. For each task, we design prompts to guide LLMs on the expected output. Particularly, for Task 1 and Task 2, to stimulate the reasoning ability of LLMs, we design prompts with the Chain-of-Thought idea (Wei et al., 2022) as shown in Figure 3. For Task 3, we believe that the task goal itself is straightforward enough, so it is not suitable for the Chain-of-Thought. Below we describe the details of the three benchmark tasks:

Task 1: Answer Selection In Task 1, LLMs are required to select the correct answer from four given candidate explanations for each question. The annotation of candidate explanations is illustrated in Figure 2. In general, each sample in this task is a tuple $\{p, q, O_A, O_B, O_C, O_D, l\}$, where p is our given prompt as shown in Figure 3, q is the input question, O_A, O_B, O_C, and O_D are four candidate explanations, and $l \in \{A, B, C, D\}$ is the golden label indicating O_l is the correct explanation. **The design motivation of this task is to test whether LLMs can distinguish right from wrong when seeing the correct and wrong answers in the context of a given cunning question.**

Task 2: Question Type Classification If LLMs are directly tasked with determining the corresponding question type, it will help us in conducting an initial automated assessment of the LLM's understanding ability. The question type classification task is specifically designed to evaluate whether LLMs can classify the cunning question into categories aligned with human intuition based on the hidden irrational aspects within the current question. The annotated problem types are shown in Table 1. During task evaluation, all the problem types will be combined with the prompt to allow LLMs to directly pick the correct type of cunning question. **We believe that LLMs understanding the concept or type of fallacy first is an unavoidable prerequisite for them to handle cunning questions well.**

Task 3: Question Explanation To further test whether LLMs truly understand the given question, we design the explanation task. In this task, the designed prompt and questions are directly input into LLMs, enabling them to "read" input questions and generate corresponding explanations. Note that since some of the questions are not expressed in the form of inquiries, we have additionally set a prompt to guide LLMs in identifying the question (See Figure 3). The generated explanations will be compared with the correct explanation for evaluation. **If LLMs can generate correct explanations, we believe that they have the ability to identify the traps of cunning questions and have come close to human intelligence.**

Evaluation Metrics For Task 1 and Task 2, we automatically calculate accuracy directly based on the LLMs' selection and classification.

To evaluate the quality of LLMs' generated explanations in Task 3, we employ **automated evaluation** along with **human evaluation** to score their explanations respectively. For automated evaluation, inspired by MT-Bench (Zheng et al., 2023), we construct prompts that incorporate the task instruction, questions, LLM's explanations and reference answers. These prompts are fed into GPT-4, which is tasked with assigning a score ranging from 1 to 10. The prompt for the automated evaluation is illustrated in Appendix C. For human evaluation, we hire 3 evaluation annotators to rate LLMs' explanations, with scores ranging from $\{1, 2, 3, 4, 5\}$. To ensure fair evaluation of the explanations of LLMs, we developed a set of scoring guidelines for annotators, including the definitions and relevant examples for each score. The scoring guidelines of human evaluation are presented in Appendix D.

3 Experiments

3.1 Experimental Settings

To better reflect the evaluation of `FLUB`'s fallacy understanding ability of LLMs, we select some advanced LLMs that are widely used and have great influence in the Chinese community:

- **ERNIE-Bot** (Baidu, 2023) is a series of closed-sourced commercial LLMs released by Baidu. We evaluate the three latest chat models in the ERNIE-bot series, including `ERNIE-Bot-3.5`, `ERNIE-Bot-3.5-Turbo`, and `ERNIE-Bot-4.0`.

- **ChatGPT** (OpenAI, 2023) ChatGPT is undoubtedly the hottest and most studied model developed by OpenAI. Currently, ChatGPT mainly has GPT-3.5 and GPT-4, so we evaluate `GPT-3.5-Turbo` and `GPT-4-Turbo`.

- **ChatGLM3** (Du et al., 2022) is the latest open-sourced model of the ChatGLM series, and ChatGLM is a series of bilingual LLMs based on the General Language Model (GLM) framework. We evaluate the only open-sourced parameter size of ChatGLM3, which is 6B (i.e., `ChatGLM3-6B`).

- **Qwen** (Bai et al., 2023) is a series of open-sourced LLMs that encompasses distinct models developed by the Alibaba Group. We select three chat Qwen models with various parameter sizes, including `Qwen-7B-Chat`, `Qwen-14B-Chat`, and `Qwen-72B-Chat`.

- **Yi** (01-AI, 2023) series models are open-sourced LLMs trained from scratch by 01-AI. The Yi models are trained on large-scale multilingual corpus as the bilingual models. In our experiments, we select `Yi-6B-Chat` and `Yi-34B-Chat` to be evaluated on `FLUB`.

- **Baichuan2** (Yang et al., 2023) is a series of open-sourced multilingual models that have achieved the competitive performance of its size on many Chinese benchmarks. Based on their open-sourced status, `Baichuan2-7B-Chat` and `Baichuan2-13B-Chat` are selected by us.

In the process of our running LLMs inference, for closed-sourced LLMs, we access corresponding models via the official APIs. Meanwhile, open-sourced models are deployed on 1 to 4 NVIDIA A100 GPUs depending on their parameter size.

Models	Open Source	Selection Accuracy	Classification Accuracy	Explanation GPT-4 Score
ERNIE-Bot-3.5 (Baidu, 2023)	✗	52.76 (38.37)	**21.71** (16.59)	6.349
ERNIE-Bot-3.5-Turbo (Baidu, 2023)	✗	32.97 (34.65)	1.71 (10.12)	5.783
ERNIE-Bot-4.0 (Baidu, 2023)	✗	<u>75.66</u> (71.34)	<u>20.00</u> (12.32)	<u>7.729</u>
GPT-3.5-Turbo (OpenAI, 2023)	✗	50.48 (48.08)	5.61 (7.68)	6.233
GPT-4-Turbo (OpenAI, 2023)	✗	**79.38** (82.73)	15.37 (15.00)	**8.952**
ChatGLM3-6B (Du et al., 2022)	✓	35.01 (48.44)	17.56 (18.54)	4.983
Qwen-7B-Chat (Bai et al., 2023)	✓	38.49 (34.17)	**19.27** (24.88)	5.392
Qwen-14B-Chat (Bai et al., 2023)	✓	42.57 (39.69)	<u>17.68</u> (18.78)	5.241
Qwen-72B-Chat (Bai et al., 2023)	✓	**58.63** (59.35)	15.12 (15.49)	**7.335**
Yi-6B-Chat (01-AI, 2023)	✓	32.61 (36.57)	12.80 (17.80)	5.731
Yi-34B-Chat (01-AI, 2023)	✓	<u>47.96</u> (61.15)	7.20 (20.73)	<u>6.970</u>
Baichuan2-7B-Chat (Yang et al., 2023)	✓	43.17 (36.45)	2.44 (6.34)	5.476
Baichuan2-13B-Chat (Yang et al., 2023)	✓	37.05 (40.41)	4.02 (4.15)	5.787
Random	-	25.00	11.11	-

Table 2: The main results on FLUB. The results in parentheses are the performance with Chain-of-Thought. We **bold** the optimal and <u>underline</u> the suboptimal of closed-source and open-source models for convenience.

3.2 Automatic Evaluation Results

For Task 1 and Task 2, we automatically calculate the accuracy. For Task 3, we utilize GPT-4 to automatically score the explanations generated by LLMs. The automatic results are presented in Table 2 and we have the following insights:

1. **For the difficulty of different tasks**, as we expected, the Answer Selection task is the simplest, which shows that LLMs should have a certain ability to distinguish right from wrong when seeing correct and wrong answers. However, we also see that the performance of all models on the Question Type Classification task is unsatisfactory, with accuracy rates below 25%. This deficiency may stem from the models' limited capability to comprehend the semantics of various question categories.

2. **For the model performance of different scale parameters**, overall, models of larger scale are better equipped to understand cunning questions, which aligns with intuitive expectations. Of course, there are exceptions. We find that Qwen-7B-Chat achieves better performance on the Question Type Classification task than other larger models, which indicates that Qwen-7B-Chat's fallacy type understanding ability is very competitive.

3. **For the connection between different tasks**, the comparative outcomes among different models across the three tasks are not consistent. Nevertheless, models that exhibit superior performance in the Answer Selection task tend to generate more plausible explanations. This phenomenon reminds us that there is a close relationship between the Answer Selection task and the Question Explanation task. The interaction between these two tasks is very critical for improving the fallacy understanding ability of LLMs.

4. **For the impact of Chain-of-Thought**, to our surprise, judging from the results, Chain-of-Thought does not bring qualitative improvements to LLMs' reasoning ability and fallacy understanding ability. Especially for the Answer Selection task, Chain-of-Thought even has negative impacts. This phenomenon demonstrates the challenging nature of FLUB and implies that we need to study new strategies besides Chain-of-Thought to stimulate LLMs' reasoning capabilities.

3.3 The Impact of In-context Learning

We select 5 high-performing LLMs to study the impact of in-context learning on LLMs' fallacy understanding ability. Demonstrations used for in-context learning are randomly selected. As shown in Figure 4, unlike Chain-of-Thought which has almost no positive effect, the LLM's performance with in-context learning is basically on the rise as demonstrations increase. This indicates that letting LLMs see more examples can improve their fallacy understanding ability, but the number of examples must be large enough because we have also seen that when only one shot example is added, the performance of LLMs tends to decline compared to the zero-shot cases.

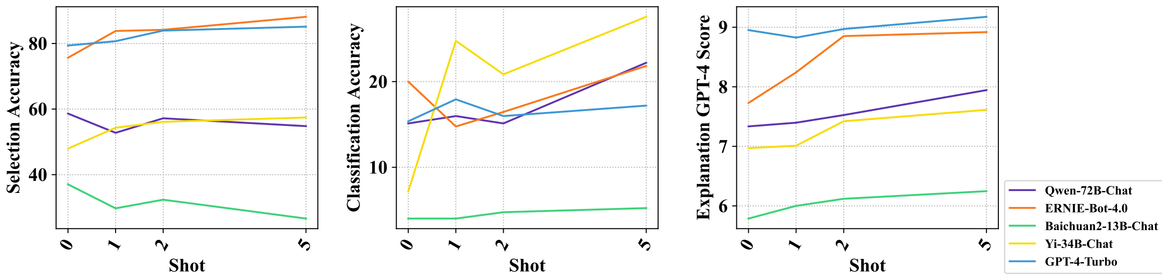

Figure 4: The results of in-context learning with 0/1/2/5-shots demonstrations.

Models	Human	GPT-4	Correlation
GPT-4-Turbo	7.12	8.60	0.57
ERNIE-Bot-4.0	5.82	7.20	0.71
Qwen-72B-Chat	5.74	7.82	0.42
Yi-34B-Chat	5.42	6.44	0.74
Baichuan2-13B-Chat	4.42	5.84	0.63
Overall	-	-	0.69

Table 3: Human evaluation and automated evaluation results on the explanation task. Note that we multiply the human results by 2 to normalize their range to be the same as the GPT-4 results' range. The reported correlations are Spearman's rank correlation coefficients. All correlations are extremely significant with $p < 0.01$.

3.4 Human Evaluation of Explanation

To verify the effectiveness of our designed automatic GPT-4's evaluation for Task 3, we randomly select 50 data samples from FLUB, along with outputs from 5 high-performing LLMs for human evaluation by our contracted annotators. From the human evaluation results in Table 3, we observe that:

1. The overall correlation coefficient between the automatic and human evaluation is 0.69, indicating a high consistency between GPT-4 scores and human preferences. Besides, the correlation results also verify the effectiveness of our designed GPT-4 score for Task 3.

2. Both automatic and human evaluations exhibit a broadly consistent ranking across the selected five models. The GPT-4-Turbo achieves superior performance over all other models. In contrast, human annotators perceive marginal performance disparities among ERNIE-Bot-4.0, Qwen-72B-Chat, and Yi-34B-Chat models. In addition, a notable discrepancy emerges in the evaluation of Qwen-72B-Chat model, where human annotators assign lower ratings than those derived from GPT-4's automatic evaluation.

3. From the results of human evaluation, except for GPT-4-Turbo, which can exceed the passing score of 6, the performance of other LLMs is still not ideal, which shows that the community still needs to further study how to improve the fallacy understanding ability of LLMs.

3.5 Case Study

To reflect FLUB's challenge to LLMs, we conduct a case study on the two advanced models with better overall performance in the question explanation task in Table 4. From the first case, we see that GPT-4-Turbo gives a relatively perfect explanation, while ERNIE-Bot-4.0's answer does not explain the causal relationship clearly although its final conclusion is correct. According to ERNIE-Bot-4.0's explanation, if the egg is added with water, it can be restored. This is obviously wrong. In the second case which is more difficult, both ERNIE-Bot-4.0 and GPT-4-Turbo easily fail when facing this cunning question. Specifically, ERNIE-Bot-4.0 follows the trap of the question, not clearly stating that "putting heads on the shore" is an impossible operation, but giving a dumbfounding explanation. In comparison, GPT-4-Turbo's performance is slightly better, but it does not perceive the trap in the question at all, resulting in an answer that is not what is questioned. It can be seen from these two cases that LLMs' ability to handle cunning questions is still insufficient.

4 Related Work

4.1 Reasoning Evaluation of LLMs

Our proposed FLUB is for evaluating the fallacy understanding ability of LLMs, which is closely related to the reasoning of LLMs (Chang et al., 2023; Guo et al., 2023). Therefore, this section first reviews related works on the evaluation of commonsense and logical reasoning of LLMs:

Input Question 1	有个问题，鸡蛋干加水会变成鸡蛋吗? I have a question, will dried eggs turn into eggs if water is added to them?
Correct Answer	鸡蛋干是将鸡蛋浓缩加工，而不是表面上“鸡蛋缺水变干”的意思，所以鸡蛋干加水不会变成鸡蛋。 Dried eggs are concentrated and processed eggs, rather than the superficial meaning of “eggs become dry due to lack of water”, so adding water to dried eggs will not turn them into eggs.
`ERNIE-Bot-4.0`	...鸡蛋干是鸡蛋的加工制品，不是鸡蛋。加水不能使鸡蛋干还原成鸡蛋... Dried eggs are processed products of eggs, not eggs. Adding water will not turn the dried eggs back into eggs.
`GPT-4-Turbo`	...鸡蛋干是经过脱水处理的鸡蛋产品，加水只能使其回软，但不能恢复到原始的鸡蛋状态... Dried eggs are dehydrated egg products. Adding water can only soften them, but cannot return them to their original egg state.
Input Question 2	游泳初学者怕呛死，下水前能先把头放岸边吗? Beginner swimmers are afraid of choking. Can they put their heads on the shore before entering the water?
Correct Answer	头是人身体的一部分，如果头离开了身体，不用呛死人就已经死亡了。 The head is part of the human body. If the head leaves the body, the person will die without choking him.
`ERNIE-Bot-4.0`	...将头部放在岸边并不是一种标准的游泳训练做法。游泳初学者下水前不应将头放在岸边... Keeping your head on the bank is not a standard swim training practice. Beginner swimmers should not put their heads on the shore before entering the water.
`GPT-4-Turbo`	...游泳初学者在下水前可以先在岸边练习将头部放入水中并慢慢学会控制呼吸... Beginner swimmers can practice putting their heads in the water on the shore and slowly learn to control their breathing before entering the water.

Table 4: Representative case study. We mark correct/partially correct/wrong explanations in green/orange/red.

Commonsense Reasoning Existing commonsense reasoning benchmarks include CommonsenseQA (Talmor et al., 2019), PIQA (Bisk et al., 2020), Social IQA (Sap et al., 2019), HellaSWAG (Zellers et al., 2019), and MC-TACO (Zhou et al., 2019). Their task is presented in the form of multiple-choice questions. The recent LLMs reasoning evaluation works (Bang et al., 2023; Bian et al., 2023) have demonstrated that LLMs represented by ChatGPT often cannot accurately utilize commonsense knowledge for the reasoning process, resulting in the commonsense reasoning task still being challenging for LLMs.

Logical Reasoning For logical reasoning data resources, they can be mainly divided into two categories: Natural Language Inference (Saha et al., 2020; Tian et al., 2021; Liu et al., 2021) and Multiple-Choice Reading Comprehension (Liu et al., 2020; Wang et al., 2022; Liu et al., 2023a). So far, there have been studies that have conducted in-depth analyses of the performance of LLMs on these two types of tasks. Liu et al. (2023b) show that logical reasoning is very challenging for LLMs, especially for out-of-distribution data samples.

In summary, research on the reasoning ability of LLMs is the focus of current and future LLMs-centric research. The fallacy understanding ability and cunning questions we are concerned about are actually comprehensive challenges of LLMs.

4.2 Humor in NLP

We have noticed that some samples in `FLUB` contain humorous expressions. Therefore, NLP research on humor (Anjum and Lieberum, 2023) is instructive for future exploration on `FLUB`. Particularly, as a representative task in humor research, word game tasks with puns as the core have been continuously paid attention to by researchers (Hempelmann, 2008; Chen and Soo, 2018; Popova and Dadić, 2023). According to our statistics, a large proportion of `FLUB` are cunning questions belonging to word games. Therefore, we believe that how to improve the humor recognition and processing capabilities of LLMs is also the key to improving the performance of LLMs on `FLUB`.

5 Conclusion

In this work, we construct `FLUB`, a high-quality benchmark consisting of cunning questions designed to evaluate the fallacy understanding ability of LLMs. Furthermore, we evaluate advanced LLMs on `FLUB`. Detailed analyses indicate `FLUB` is very challenging and of great research value. To date, most existing LLMs still can not understand the fallacy well, which results in them being far from dealing with complex problems in the real world as easily as humans. We believe that the benchmark and the research direction we provide are valuable for the LLMs community.

Limitations

One limitation of `FLUB` may be that it consists of Chinese data. In particular, many of the cunning questions in `FLUB` have certain Chinese cultural and language characteristics as backgrounds, which places extremely high demands on LLMs' knowledge storage. However, as a community that cannot be ignored in the field of NLP, the development of Chinese NLP and Chinese LLMs has been devoted by generations of researchers.

In addition, we are also actively looking for resources to build the English version of `FLUB`, namely FLUB2.0. But our own resources are limited after all, so we hope that the introduction of `FLUB` can attract more researchers in the community to pay attention to the importance of fallacy understanding in LLMs, and join the research dedicated to improving the fallacy understanding ability of LLMs.

Ethics Statement

In this paper, we present a new benchmark, `FLUB`. We have described the details of the collection, preprocessing, and annotation of `FLUB`. And we ensure that no infringement or unethical behavior occurred during the dataset construction. In terms of the data itself, to ensure that the dataset we need to release in the future meets ethical requirements, we spend lots of energy on data anonymization, data desensitization, improper data cleaning, etc. Besides, the cunning questions we are concerned about come from daily life and are very common. Therefore, the new research direction and tasks we propose will not cause harm to human society.

References

01-AI. 2023. Yi, `https://github.com/01-ai/Yi`.

Aftab Anjum and Nikolaus Lieberum. 2023. Exploring humor in natural language processing: A comprehensive review of JOKER tasks at CLEF symposium 2023. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 1828–1837. CEUR-WS.org.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Baidu. 2023. Ernie-bot, `https://cloud.baidu.com/product/wenxinworkshop`.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

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. *CoRR*, abs/2303.16421.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 113–117. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Rosa Falotico and Piero Quatto. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49:463–470.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

Christian F Hempelmann. 2008. Computational humor: Beyond the pun? *The Primer of Humor Research. Humor Research*, 8:333–360.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Yangning Li, Feng Zhou, Hai-Tao Zheng, and Qingyu Zhou. 2023. On the (in)effectiveness of large language models for chinese text correction. *CoRR*, abs/2307.09007.

Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. 2021. Natural language inference in context - investigating contextual reasoning over long texts. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13388–13396. AAAI Press.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023a. Logiqa 2.0 - an improved dataset for logical reasoning in natural language understanding. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:2947–2962.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023b. Evaluating the logical reasoning ability of chatgpt and GPT-4. *CoRR*, abs/2304.03439.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3622–3628. ijcai.org.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Olga Popova and Petra Dadić. 2023. Does ai have a sense of humor? clef 2023 joker tasks 1, 2 and 3: using bloom, gpt, simplet5, and more for pun detection, location, interpretation and translation. *Proceedings of the Working Notes of CLEF*, 3.

Swarnadeep Saha, Yixin Nie, and Mohit Bansal. 2020. Conjnli: Natural language inference over conjunctive sentences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8240–8252. Association for Computational Linguistics.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 4462–4472. Association for Computational Linguistics.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through logicnli. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3738–3747. Association for Computational Linguistics.

Siyuan Wang, Zhongkun Liu, Wanjun Zhong, Ming Zhou, Zhongyu Wei, Zhumin Chen, and Nan Duan. 2022. From LSAT: the progress and challenges of complex reasoning. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30:2201–2216.

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.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3361–3367. Association for Computational Linguistics.

给你输入一个句子或问题，其中存在不合理或幽默之处。另外给出四个选项，你需要选出最能准确描述给定句子或问题的不合理或幽默之处的一个选项。

注意，你必须直接输出你的答案，不能包含任何解释，答案必须属于"A, B, C, D"中的一个。

以下是输入：
{sentence}

选项：
{options}

Task 1

给你输入一个句子或问题，其中存在不合理或幽默之处。你需要从“候选分类”中选出一个最适合该句子或问题的类别。

候选分类：{candidates}

注意，你必须直接输出你的答案，不能包含任何解释，答案必须属于候选分类中的一个。

以下是输入：
{sentence}

Task 2

Figure 5: Our designed prompts without Chain-of-Thought for Task 1 and Task 2.

Give you a sentence or question that contains some irrationality or humor. Then give you four options, you need to choose the one that best describes the irrationality or humor of the given sentence or question, and give your reasons for choosing that option.

Your output must strictly follow the following format:
Analyze：<Briefly analyze which of the four options accurately describes the irrationality or humor of the given sentence or question, and explain why you chose that option>
Answer：<Only one of “A, B, C, D” can be output>

The following is the input:
{sentence}

Options：
{options}

Task 1

Give you a sentence or question that contains some irrationality or humor. You need to explain what is unreasonable or humorous in one sentence, and then choose a type from the “candidate types” that best fits the sentence or question and explain why.

Candidate types：{candidates}

Your output must strictly follow the following format:
Explain：<Explain what is unreasonable or humorous in one sentence>
Reason：<Explain the reason for choosing the type from “candidate types”>
Type：<The type selected from “candidate types”>

The following is the input:
{sentence}

Task 2

Input the following sentence to you, which contains some irrationality or humor. Please briefly explain what makes this sentence unreasonable or humorous in three sentences or less.

{sentence}

Task 3(a)

Please briefly answer the following questions in three sentences or less:

{sentence}

Task 3(b)

Figure 6: The English translations of our designed prompts for FLUB in Figure 3.

Give you a sentence or question that contains some irrationality or humor. Give you four options, you need to choose the one that best describes the irrationality or humor of the given sentence or question.

Your Note that you must output your answer directly without any explanation. The answer must belong to one of "A, B, C, D".

The following is the input:
{sentence}

Options：
{options}

Task 1

Give you a sentence or question that contains some irrationality or humor. You need to choose a type from the “candidate types” that best fits the sentence or question.

Candidate types：{candidates}

Your Note that you must output your answer directly, without any explanation, and the answer must belong to one of the candidate types.

The following is the input:
{sentence}

Task 2

Figure 7: The English translations of our designed prompts for FLUB in Figure 5.

A Details of FLUB Annotation

Since the annotation difficulty of different information is different, the salary we pay to the annotators we employ is also different. Specifically, we pay each person who annotates the question type \$0.5 per sample, each person who writes the correct explanation \$1 per sample, and each person who writes the wrong candidates \$2 per sample. It is worth noting that in addition to the junior annotators providing the initial annotation results, we also set three senior annotators with a salary of \$2 per sample, who are responsible for carefully checking the correctness of the annotation results provided by the junior annotators.

B Translations of Prompts set in FLUB

Our designed prompts without Chain-of-Thought for Task 1 and Task 2 are shown in Figure 5. The English translations of all our designed prompts for each task of FLUB are shown in Figure 6 and Figure 7.

C Automatic Evaluation Prompt in Task 3

To enable GPT-4 can automatically score explanations generated by other LLMs according to the settings of Task 3, we carefully design the corresponding prompt, as shown in Figure 8.

D Scoring Guidelines in Task 3

Scoring Objective For the LLMs' output response to each question, please refer to the given annotated answer to score the output response. This score will be used to measure the quality of the model response and the degree of matching with the answer.

Scoring Rules The scoring values are defined as $\{1, 2, 3, 4, 5\}$. The higher the score, the higher the quality of the model output. **The specific definitions and examples of each score are as follows:**

D.1 Definition and Examples of Score 1

Definition The LLM's response doesn't point out the irrationality or humor in the given question, which is completely unrelated to the correct answer.

D.1.1 The First Example of Score 1

Question: Since instant noodles can be eaten dry, then dry crispy noodles can also be boiled? (既然方便面可以干吃，那干脆面也可以水煮？)

Model Response: Yes, dry crispy noodles can also be boiled. Boil in boiling water until softened and serve. (是的，干脆面也可以水煮。将其放入沸水中煮至软化即可食用。)

Correct Answer: Dry crispy noodles are crispy and delicious ready-to-eat snack food. You can eat them as soon as you open the package. What people need is its crispness. So its taste will be lost after being boiled. Therefore, crispy noodles cannot be boiled. (干脆面是一种香脆可口的即

```
[Instruction]
Please act as an impartial judge and evaluate the quality of the response provided by an
AI assistant to the user question displayed below. Your evaluation should consider
correctness and helpfulness. You will be given a reference answer and the assistant's
answer. Begin your evaluation by comparing the assistant's answer with the reference
answer. Identify and correct any mistakes. Be as objective as possible. After providing
your explanation, you must rate the response on a scale of 1 to 10 by strictly following
this format: \"[[rating]]\", for example: \"Rating: [[5]]\".

[Question]
{question}

[The Start of Reference Answer]
{answer}
[The End of Reference Answer]

[The Start of Assistant's Answer]
{response}
[The End of Assistant's Answer]
```

Figure 8: The designed prompt of automatic evaluation for Task 3.

食型休闲食品，打开包装就可以吃，吃的就是它的香脆，水煮之后会失去它的口感。所以干脆面不可以水煮。)

D.1.2 The Second Example of Score 1

Question: To say that Ruozhi Bar is all a bunch of retards is that considered a compliment or an insult? (说弱智吧都是一帮弱智，算夸奖还是辱骂？)

Model Response: This sentence is insulting because it uses the derogatory term “retarded” to describe a group of people and expresses contempt and insult to them. (这句话是辱骂，因为它使用贬义词汇“弱智”来形容一群人，表达了对他们的轻视和侮辱。)

Correct Answer: “Ruozhi Bar” is the name of the post bar, and the IQ of the members inside has nothing to do with what the post bar name is. “Retarded” means intelligence below the normal level, and to say ‘Ruozhi Bar” is a bunch of retarded is an insult. (“弱智吧”是贴吧名，里面成员的智商与贴吧名是什么无关。“弱智”的意思是智力低于正常水平，说弱智吧都是一帮弱智是一种辱骂。)

D.2 Definition and Examples of Score 2

Definition Although LLM’s response points out the unreasonable/humorous points in the question, the types and key points pointed out are inconsistent with the opinions of explanations.

D.2.1 The First Example of Score 2

Question: Dreams are all reversed. I dream of myself looking in the mirror, so I am the one in the mirror. (梦都是反的，我梦到自己照镜子，所以我才是镜子里的人。)

Model Response: The humor of this sentence lies in the literal interpretation of the saying “Dreams are all reversed”. Usually, people think that the scene in a dream is the opposite of reality, but in this sentence, the author dreams of himself looking in the mirror, if we understand that “Dreams are all reversed”, then the author is actually a reflection in the mirror, which is obviously an absurd interpretation of reality, and therefore has a sense of humor. (这个句子的幽默之处在于对“梦都是反的”这一说法的字面解释。通常，人们认为梦中的情景与现实相反，但在这个句子中，作者梦到自己照镜子，如果按照“梦都是反的”来理解，那么作者实际上是镜子里的倒影，这显然是对现实的荒诞解释，因此具有幽默感。)

Correct Answer: “Dreams are all reversed” means that the dream is opposite to reality. If “I dream of myself looking in the mirror” is contrary to reality, the reality should be “I do not look in the mirror in reality”, rather than “I dream of myself looking in the mirror, and I am the person in the mirror”. (“梦都是反的”是指梦境与现实是相反的，“我梦到自己照镜子”与现实相反，应该是“现实中没有照镜子”，而不是“我梦到自己照镜子，反过来我是镜中人”。)

D.2.2 The Second Example of Score 2

Question: If all three of Nezha’s heads suffered from cerebral thrombosis for ten years, did he have cerebral thrombosis for ten years or thirty years? (哪吒三头都得了十年脑血栓那是算十年脑血栓还是三十年脑血栓？)

Model Response: Nezha is a fictional character in Chinese mythology that does not exist. Cerebral thrombosis is a cerebrovascular disease, usually caused by blood clotting in the blood vessels of the brain resulting in brain hypoxia and ischemia. Therefore, the claim that Nezha's three heads have cerebral thrombosis is not realistic, let alone calculate whether it is ten years or thirty years. (哪吒是中国神话中的虚构人物，没有真实存在。脑血栓是一种脑血管疾病，通常是由于脑部血管内的血液凝固形成的血栓导致脑部缺氧缺血。因此，哪吒三头得脑血栓的说法是不符合实际的，更不用计算是十年还是三十年。)

Correct Answer: Nezha is an individual, assuming Nezha's three heads all have cerebral thrombosis, it should be something that happens at the same time, and will not produce accumulation, so it should not be 30 years. (哪吒是一个个体，假设哪吒的三个头都得了脑血栓，那应该是同时发生的事情，并不会产生累加，所以不应该是三十年。)

D.3 Definition and Examples of Score 3

Definition The LLM's response correctly points out the type of irrationality/humor, but the key points of irrationality/humor pointed out by the correct answer are inconsistent. Alternatively, the response is consistent with the key points of irrationality/humor pointed out by the correct answer, but the judgment on the type of question is inconsistent.

D.3.1 The First Example of Score 3

Question: The drug is partially poisonous with side effects, so we should not take the drug in parts. (是药三分毒，所以药不能分三次喝。)

Model Response: The humor of this sentence lies in the misunderstanding of the literal and actual meaning of the idiom "The drug is partially poisonous with side effects". This idiom is meant to say that any drug has potential side effects, while the sentence misunderstands "partially" to mean that the drug should not be taken in parts, and thus draws the absurd conclusion that the drug cannot be taken in three separate doses, which is obviously a comical distortion of the original meaning. (这句子的幽默之处在于对成语"是药三分毒"的字面意思和实际含义的误解。成语本意是说任何药物都有潜在的副作用，而句子却把"三分"误解为药物应该分成三个部分服用，因此得出了一个荒谬的结论，即药物不能按照三次分开服用，这显然是对原意的一种滑稽曲解。)

Correct Answer: "The drug is partially poisonous with side effects" refers to any drug will have certain side effects if used improperly. "Partially" indicates that to a certain extent, there is no meaning of "in parts". Therefore, the conclusion "we should not take the drug in parts" cannot be drawn from this sentence. ("是药三分毒"指的是任何的药品如果使用不当的话都会存在一定的副作用，"三分"表示的是在一定程度上，并没有分三次的意思，所以"是药三分毒"这句话并不能得出"药不能分三次喝"这一结论。)

D.3.2 The Second Example of Score 3

Question: Did a knee-jerk reflex test on a friend, and the result was that his leg never lifted again after the hammer was knocked down. This reflex arc is too long! (给朋友做膝跳测试，结果锤下去之后他的腿再也没抬起过，这反射弧也太长了吧！)

Model Response: The knee-jerk reflex test is a simple test to check the function of the nervous system. Normally, a light tap on the tendon below the knee will cause the leg to make a quick leg lift motion. This sentence humorously exaggerates the amount of time the friend does not respond, indicating that the reflex arc is unusually long, when in fact, if there is really no response, it may indicate a neurological problem rather than the reflex arc length. In addition, using "never lifted again" has a more humorous effect because it indicates a permanent state, which is impossible in reality. (膝跳反射测试是一种检查神经系统功能的简单测试，正常情况，轻轻敲击膝盖下方的腱会导致腿部做出快速的抬腿动作。这句话幽默地夸张了朋友没有反应的时间，暗示反射弧异常地长，实际上如果真的没有反应，那可能表明存在神经系统问题，而不是反射弧长。此外，用"再也没抬起过"增加了幽默效果，因为它暗示了一种永久性的状态，这在现实中是不可能的。)

Correct Answer: "His leg never lifted again after the hammer" indicates that the friend's leg was broken, and is not related to the length of the reflex arc, here is actually a humorous way to express the matter of "breaking the friend's leg". ("锤下去之后他的腿再也没抬起过"说明把朋友的腿打断了，与反射弧长短没有关系，这里实际上是用幽默的方式将"把朋友腿打断"表达出来。)

D.4 Definition and Examples of Score 4

Definition The response correctly points out the type of irrationality/humor in the question, and the key points pointed out are correct, but the interpretation is not consistent with the correct answer.

D.4.1 The First Example of Score 4

Question: In a vegetable market, one finds vegetative persons instead. So it must have been a job market before. (去素材库发现几个植物人躺在里面，原来之前这里是人材市场。)

Model Response: The humor of this sentence lies in the pun on the word "vegetative". Usually, "vegetative" refers to patients who have fallen into a persistent vegetative state due to a major illness or injury, but here it is used to refer to "vegetable" in the literal sense, i.e., real plants. At the same time, "job market" usually refers to a market for hiring "people" with the appropriate skill, but here it is used teasingly to refer to a market with vegetative "people". This linguistic coincidence creates an absurd humorous effect. (这个句子的幽默之处在于使用了"植物人"这个词语的双关意味。通常，"植物人"指的是那些因为重大疾病或伤害而陷入持续植物状态的病人，但在这里它被用来指代字面意义上的"植物"，即真正的植物。同时，"人材市场"通常是指招聘人才的市场，而在这里它被戏谑地用来指一个植物市场，这种语言上的巧合创造了一种荒谬的幽默效果。)

Correct Answer: The vegetable market refers to the market where vegetables are sold. "Vegetative" is a special human state similar to the living state of plants, and the human body cannot lie in the vegetable market. (素材库是指销售蔬菜的市场。"植物人"是与植物生存状态相似的特殊的人体状态，人体是无法躺进素材库中的。)

D.4.2 The Second Example of Score 4

Question: If you drink 82 Lafite, then you'll live at least 82 years. (喝了82年拉菲至少能活82岁。)

Model Response: The humor or irrationality of the sentence is that it establishes a false cause-and-effect relationship, implying that drinking a certain vintage of wine guarantees living to the same age as the vintage, which obviously has no scientific basis. In fact, a person's lifespan is affected by many factors, including genetics, lifestyle, medical conditions, and so on, rather than a single diet. In addition, this statement also carries a kind of humorous exaggeration, because it uses an absurd logic to attract attention and increase the interest of the words. (该句子的幽默或不合理之处在于它建立了一种错误的因果关系，暗示喝了某年份的葡萄酒能保证活到与年份相同的年龄，这显然没有科学依据。实际上，人的寿命受多种因素影响，包括遗传、生活方式、医疗条件等，而非单一的饮食习惯。此外，这种说法也带有一种幽默的夸张，因为它用一种荒谬的逻辑来吸引注意，增加话语的趣味性。)

Correct Answer: "82 Lafite" is a network buzzword that originated from scenes of opening Lafite wine in movies and TV series. Since the Lafite red wine produced in 1982 is of good quality and expensive, "82 Lafite" is synonymous with high standards to set off the character's extravagant spending and taste, rather than "Lafite has been drunk for 82 years". (82年的拉菲，网络流行语，源自影视剧中开拉菲的桥段，由于82年的拉菲红酒品质好、价格昂贵，所以"82年的拉菲"是高规格的代名词，来衬托人物的挥金如土和品味，而不是"喝了八十二年的拉菲"的意思。)

D.5 Definition and Examples of Score 5

Definition The LLM's response and the correct explanation are basically matched, the response correctly points out the type of unreasonable/humorous points in the question, the key points are correct, and the explanation is basically consistent with the explanation.

D.5.1 The First Example of Score 5

Question: If a police officer commits a crime and turns himself in, is it considered to have made meritorious contributions? (警察犯了事去自首算不算戴罪立功啊？)

Model Response: "Meritorious service" usually refers to the behavior of criminals who actively cooperate with the police to detect other cases or fight criminal activities after committing a crime, so as to reduce their own punishment. If the police officer has committed a crime and turned himself in, and there is no further meritorious service, then this can usually only be counted as surrender, and cannot be directly referred to as meritorious service. However, surrendering may lead to reduced punishment according to relevant legal provisions. ("戴罪立功"通常指的是犯罪分子在犯罪后，积极配合警方侦破其他案件或对抗犯罪活动，从而争取减轻自己刑罚的行为。如果警察犯了罪后去自首，并且没有进一步的立功表现，那么这通常只能算作自首，并不能直接称为戴罪立功。不过，自首可能会根据相关法律规定减轻处罚。)

Correct Answer: According to the provisions of Article 67 of the Criminal Law, “Surrender” refers to the act of automatically surrendering to the police after committing a crime and truthfully confessing one’s crime. If a criminal suspect or defendant under compulsory measures or a criminal serving a sentence truthfully confesses other crimes that are not yet known to the judicial organs, he shall be regarded as surrendering. “Meritorious service” means that the meritorious service performance in criminal law refers to the performance of criminal suspects providing important clues, exposing others’ crimes, etc., and can be mitigated after verification. Surrender confession by oneself is one’s own crime, while meritorious service is to expose the crime of others, so the police committing the crime to surrender is not regarded as meritorious service. (根据刑法第67条的规定，“自首”是指犯罪以后自动投案，如实供述自己的罪行的行为。被采取强制措施的犯罪嫌疑人、被告人和正在服刑的罪犯，如实供述司法机关还未掌握的本人其他罪行的，以自首论。“戴罪立功”就是指刑法中的立功表现是指犯罪嫌疑人提供重要线索、揭发他人犯罪行等的表现，经查实的可以获得减轻处罚。自首供述的是本人的罪行，戴罪立功是揭发他人罪行，所以条子犯了事去自首不算戴罪立功。)

D.5.2 The Second Example of Score 5

Question: Officer Wang shot the hostage and saved the robber who could have been killed at any time. (王警官击毙人质，解救了随时可能被狙杀的劫匪。)

Model Response: The unreasonableness of this sentence lies in the fact that Officer Wang’s actions are contrary to the duties of ordinary police officers. Usually, the task of the police is to protect the hostages and try to arrest the criminals, not to kill the hostages to protect the criminals. Thus, the sentence reverses the duties of the police with an absurd logic that produces a humorous effect. (这句子的不合理之处在于王警官的行动与一般警察的职责相反。通常情况下，警察的任务是保护人质并尽量逮捕犯罪分子，而不是击毙人质来保护犯罪分子。因此，这个句子以一种荒谬的逻辑颠倒了警察的职责，产生了幽默效果。)

Correct Answer: As a police officer, what Wang should do is shoot the robber to save the hostage, not shoot the hostage to save the robber. The fact that Wang killed the hostage by mistake is described humorously. (王警官作为警察，应该做的是击毙劫匪解救人质，而不是击毙人质解救劫匪，这里用幽默的方式阐述了王警官误杀人质这一事实。)