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

Yinghui Li[1]*, Qingyu Zhou[2,*,†], Yuanzhen Luo*, Shirong Ma[1], Yangning Li[1], Hai-Tao Zheng[1,†], Xuming Hu[3,†], Philip S. Yu[4]

[1]Tsinghua University, [2] Bytedance Inc.

[3]The Hong Kong University of Science and Technology (Guangzhou)

[4]University of Illinois Chicago

liyinghu20@mails.tsinghua.edu.cn

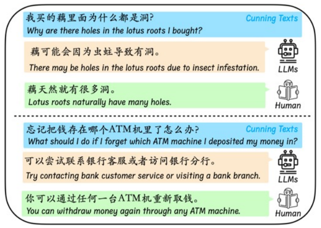
NEURAL INFORMATION PROCESSING SYSTEMS

Abstract

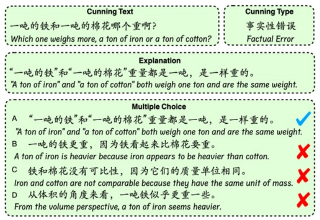
Recently, Large Language Models (LLMs) make remarkable evolutions in language understanding and generation. Following this, various benchmarks for measuring all kinds of capabilities of LLMs have sprung up. In this paper, we challenge the reasoning and understanding abilities of LLMs by proposing a **F**a**L**lacy **U**nderstanding **B**enchmark **(FLUB)** containing cunning texts that are easy for humans to understand but difficult for models to grasp. Specifically, the cunning texts that FLUB focuses on mainly consist of the tricky, humorous, and misleading texts 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. Our data and codes are available at https://github.com/THUKElab/FLUB.

The FLUB Benchmark

we collect real cunning texts as our raw data from a famous Chinese online forum, **the "Ruozhiba". 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 cunning types, **FLUB has 8 fine-grained types of cunning texts** and most of the texts 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 cunning text) and three confusing wrong answers for each input text in FLUB.



(a) The examples of how LLMs and humans perform when faced with cunning texts. The LLM we use is ChatGPT-3.5 on Jan 23, 2024.



(b) We design three tasks, namely Cunning Type Classification, Fallacy Explanation, and Answer Selection (i.e., Multiple Choice).

Figure 1: The running examples and annotation examples of FLUB.

The Cunning Types of FLUB

Cunning Type	Definition	Example
错误类比 False Analogy	由于事件A的发生具有或伴随有某种属性，从而错误地类比出与事件A相似的事件B也应该具有该属性，或者错误地类比出与事件A相反的事件B应该具有相反的属性。 Due to the occurrence of event A having or being accompanied by a certain attribute, it is erroneously analogized that event B, which is similar to event A, should also have that attribute, or that event B, which is opposite to event A, should have the opposite attribute.	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	由于缺乏对某个常识或事实的认知，从而得出某个不符合逻辑的问题或结论。注意，该句子往往因为其不寻常的认知缺失从而导致该句子可能令人发笑。 Due to a lack of understanding of a common sense or fact, a illogical question or conclusion can be drawn. Note that this sentence may be funny due to its unusual cognitive impairment.	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	通过改变固定词汇中多音字的发音，从而利用新的发音得到的句子。注意，如果读者没有领会到句子中发音的改变，会导致读者认为该句子不符合逻辑。 Sentences obtained by changing the pronunciation of polyphonic words in fixed vocabulary. Note that if the reader does not appreciate the change in pronunciation in the sentence, it will lead the reader to think that the sentence is illogical.	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), XiaoMing has to wait over an hour whenever he queue in the U.S.
歧义 Ambiguity	通过改变句子中某个多义词的词义，从而得出不符合逻辑的问题或结论。 By changing the meaning of a polysemy word in a sentence, illogical questions or conclusions can be drawn.	语文老师说我的句子是病句，我应该给这个病句吃头孢，还是打点滴呢? My teacher said the sentence is grammatically wrong. Should I give this sentence some antibiotics or administer an IV drip?
悖论 Paradox	句子或者问题的表述前后矛盾。 The expression of a sentence or question is contradictory.	"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	由于缺乏对某个事实的认知，或者对事实进行扭曲，从而提出无意义的问题或结论。 Due to a lack of understanding or distortion of a fact, meaningless questions or conclusions are raised.	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	从一个事件中推断出一个错误或者无意义的结论，或者颠倒了事件的因果关系。 Inferring an incorrect or meaningless conclusion from an event, or reversing the causal relationship of the event.	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	错误地改变句子中文字的意思或含义，在此基础上提出问题或者得出结论。 Mistakenly changing the meaning of words in a sentence, and based on this, raising questions or drawing conclusions.	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	句子本身具有错误，或者句子的表述不符合正常逻辑，但是不属于上述任何一个类别。 The sentence itself has errors, or the expression of the sentence does not conform to normal logic, but does not belong to any of the above categories.	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?

The Benchmark Task Setups of FLUB

- **Task 1: Answer Selection.** In Task 1, LLMs are required to select the correct answer from four given candidate explanations for each input text. . 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 text.
- **Task 2: Cunning Type Classification.** If LLMs are directly tasked with determining the corresponding cunning type, it will help us in conducting an initial automated assessment of the LLM's understanding ability. The cunning type classification task is specifically designed to evaluate whether LLMs can classify the cunning text into categories aligned with human intuition based on the hidden irrational aspects within the current text.
- **Task 3: Fallacy Explanation.** To further test whether LLMs truly understand the given cunning text, we design the explanation task. In this task, the designed prompt and input texts are directly input into LLMs, enabling them to "read" input texts and generate corresponding explanations.
- **Automatic Evaluation Metrics:** For Task 1, we calculate **Accuracy** directly based on the LLMs' selection results. For Task 2, considering that there are a few cunning types in FLUB with small sample size, we choose the **F-1 Score** to measure the performance of LLMs. For Task 3, we assign a **GPT-4 Score** ranging from 1 to 10.
- **Human Evaluation Settings:** For Task 1 and Task 2, **we conduct human evaluations to explore how well human-level intelligence could perform these two tasks.** For the human evaluation of Task 3, we **mainly want to verify the effectiveness of the automatic GPT-4 score we use,** therefore, we hire 3 evaluation annotators to rate LLMs' explanations, with scores ranging from {1, 2, 3, 4, 5}.

Automatic Evaluation Results

Table 1: We **bold** the optimal and underline the suboptimal of closed/open-source models. We report the overall performance by calculating the **geometric mean** of the three tasks. We color the result that Chain-of-Thought (CoT) brings positive / negative gain as green$^{\uparrow}$ / red$^{\downarrow}$.

Models	Open Source	Selection Accuracy w/o CoT	Selection Accuracy CoT	Classification F-1 Score w/o CoT	Classification F-1 Score CoT	Explanation GPT-4 Score w/o CoT	Explanation GPT-4 Score CoT	Overall Performance w/o CoT	Overall Performance CoT
ERNIE-Bot-3.5-Turbo [15]	✗	32.97	34.65$^{\uparrow}$	1.99	6.09$^{\uparrow}$	5.78	5.83$^{\uparrow}$	7.24	10.72$^{\uparrow}$
ERNIE-Bot-3.5 [15]	✗	52.76	38.37$^{\downarrow}$	10.33	11.15$^{\uparrow}$	6.35	6.22$^{\downarrow}$	15.13	13.86$^{\downarrow}$
ERNIE-Bot-4.0 [15]	✗	75.66	71.34$^{\downarrow}$	11.84	**14.42**$^{\uparrow}$	7.73	8.11$^{\uparrow}$	19.06	20.28$^{\uparrow}$
GPT-3.5-Turbo [16]	✗	50.48	48.08$^{\downarrow}$	3.09	6.15$^{\uparrow}$	6.23	7.00$^{\uparrow}$	9.91	12.74$^{\uparrow}$
GPT-4-Turbo [16]	✗	**79.38**	**82.73**$^{\uparrow}$	**12.31**	13.97$^{\uparrow}$	**8.95**	**9.21**$^{\uparrow}$	**20.60**	**22.00**$^{\uparrow}$
ChatGLM3-6B [17]	✓	35.85	35.01$^{\downarrow}$	7.48	9.34$^{\uparrow}$	4.98	4.82$^{\downarrow}$	11.01	11.64$^{\uparrow}$
Qwen-7B-Chat [18]	✓	38.49	33.69$^{\downarrow}$	8.00	10.97$^{\uparrow}$	5.39	5.65$^{\uparrow}$	11.84	11.98$^{\uparrow}$
Qwen-14B-Chat [18]	✓	42.57	43.05$^{\uparrow}$	**10.34**	10.44$^{\uparrow}$	5.24	6.24$^{\uparrow}$	13.21	14.10$^{\uparrow}$
Qwen-72B-Chat [18]	✓	**58.63**	**61.51**$^{\uparrow}$	9.32	**12.26**$^{\uparrow}$	**7.34**	**7.90**$^{\uparrow}$	**15.89**	**18.13**$^{\uparrow}$
Yi-6B-Chat [19]	✓	32.37	29.26$^{\downarrow}$	8.87	9.84$^{\uparrow}$	5.73	5.39$^{\downarrow}$	11.81	11.58$^{\downarrow}$
Yi-34B-Chat [19]	✓	47.96	48.80$^{\uparrow}$	4.74	11.70$^{\uparrow}$	6.97	7.52$^{\uparrow}$	11.66	16.17$^{\uparrow}$
Baichuan2-7B-Chat [20]	✓	43.17	37.17$^{\downarrow}$	1.02	4.45$^{\uparrow}$	5.48	4.85$^{\downarrow}$	6.23	9.29$^{\uparrow}$
Baichuan2-13B-Chat [20]	✓	37.05	38.01$^{\uparrow}$	3.52	4.58$^{\uparrow}$	5.79	5.84$^{\uparrow}$	9.11	10.06$^{\uparrow}$
Random	-	25.00		7.90		-		-	
Human	-	93.35		63.69		-		-	

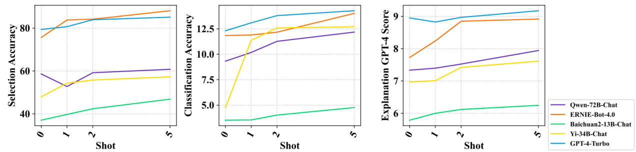


Figure 3: The results of in-context learning with 0/1/2/5-shots demonstrations.

Human Evaluation of Explanation

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