A Closer Look at the Self-Verification Abilities of Large Language Models in Logical Reasoning

Ruixin Hong[1,2]*, **Hongming Zhang**[2], **Xinyu Pang**[1], **Dong Yu**[2], **Changshui Zhang**[1]
[1]Institute for Artificial Intelligence, Tsinghua University (THUAI);
[1]Beijing National Research Center for Information Science and Technology (BNRist);
[1]Department of Automation, Tsinghua University, Beijing, P.R.China
[2]Tencent AI Lab, Seattle
hrx20@mails.tsinghua.edu.cn, hongmzhang@tencent.com,

Abstract

Logical reasoning has been an ongoing pursuit in the field of AI. Despite significant advancements made by large language models (LLMs), they still struggle with complex logical reasoning problems. To enhance reasoning performance, one promising direction is scalable oversight, which requires LLMs to identify their own errors and then improve by themselves. Various self-verification methods have been proposed in pursuit of this goal. Nevertheless, whether existing models understand their own errors well is still under investigation. In this paper, we take a closer look at the self-verification abilities of LLMs in the context of logical reasoning, focusing on their ability to identify logical fallacies accurately. We introduce a dataset, FALLACIES, containing 232 types of reasoning fallacies categorized in a hierarchical taxonomy. By conducting exhaustive experiments on FALLACIES, we obtain comprehensive and detailed analyses of a series of models on their verification abilities. Our main findings suggest that existing LLMs could struggle to identify fallacious reasoning steps accurately and may fall short of guaranteeing the validity of self-verification methods. Drawing from these observations, we offer suggestions for future research and practical applications of self-verification methods.[1]

1 Introduction

Logical reasoning is not only a crucial aspect of human intelligence but also one of the long-term pursuits of artificial intelligence (McCarthy, 1989). It is indispensable in intelligent systems, enabling problem-solving, decision-making, and critical thinking. Large language models (LLMs) have recently achieved remarkable advancements in a wide range of tasks (OpenAI, 2023). Being prompted appropriately, LLMs exhibit the emergent ability to "reason" step by step like humans (Wei et al., 2022a,b). Nonetheless, increasing research suggests that LLMs struggle with intricate logical reasoning problems, occasionally producing unfaithful reasoning steps fraught with logical fallacies (Arkoudas, 2023; Liu et al., 2023a; Pan et al., 2023; Hong et al., 2023).

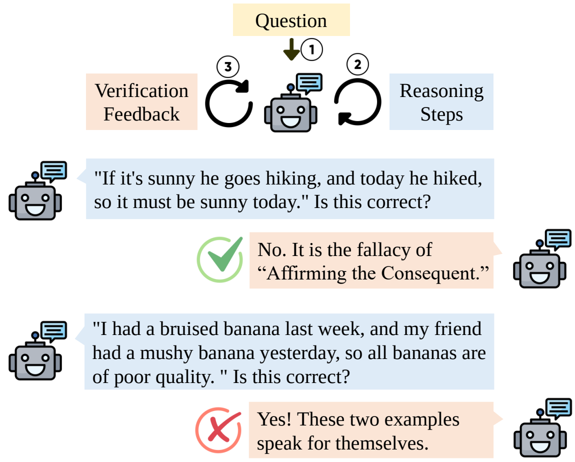

Figure 1: The self-verification approach requires LLMs to identify the fallacious steps in their own reasoning process. However, LLMs might be susceptible to certain types of fallacies and fail to identify them, leading to the potential failure of self-verification.

To tackle this issue, a prevalent and promising approach is scalable oversight (Bowman et al., 2022), where the LLMs could be boosted based on their own evaluation signals (Leike and Sutskever, 2023). In this regard, various strategies of self-verification using LLMs are proposed to enhance reasoning performance (Ling et al., 2023; Weng et al., 2022; Xie et al., 2023; Madaan et al., 2023; Miao et al., 2023). As shown in Figure 1, the LLMs first generate the reasoning process and then self-verify their own output. The verification results are then used to refine the output or further improve the models. Their *assumption* is that LLMs can reliably identify fallacious reasoning steps.

Though empirical evidence demonstrates their

* Work done during the internship at Tencent AI Lab.
[1]Data is available at https://github.com/Raising-hrx/FALLACIES.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 900–925
June 16-21, 2024 ©2024 Association for Computational Linguistics

preliminary success, a thorough and comprehensive evaluation of the underlying assumption remains unexplored. First, these efforts typically use the performance of the final task (e.g., answer accuracy) to illustrate their effectiveness. However, this is a proxy metric that does not directly reflect the ability of LLMs to identify logical fallacies. LLMs might possibly arrive at the correct answer despite the existence of fallacious intermediate steps (Wei et al., 2022b; Creswell et al., 2023; Lanham et al., 2023). Second, they are usually only concerned with *whether or not* the reasoning step is fallacious, rather than *what type of* fallacy exists in the step, which could be informative for the self-improvement of LLMs. Such an oversimplification precludes a definitive analysis of the LLMs' ability to identify different types of fallacies.

In this paper, we provide a comprehensive evaluation of the verification abilities of LLMs in logical reasoning. Specifically, we collect a dataset, FALLACIES, containing 4,640 reasoning steps for 232 types of fallacies. We evaluate whether LLMs can distinguish between correct and fallacious reasoning steps. Such directed evaluation can provide a more accurate reflection of the verification abilities of LLMs, as it steers away from proxy metrics and delves straight into the actual performance in identifying fallacious steps. Compared to previous datasets, our dataset features more types of fallacies, a larger scale, more fine-grained reasoning, and explicit premises and conclusions. Furthermore, we adopt a hierarchical taxonomy of fallacies, which divides fallacies into two main categories and nine subcategories, allowing for a more systematic approach toward analyzing the verification abilities of LLMs across varying aspects.

We conduct exhaustive experiments on a range of LLMs. First, experimental results show that most LLMs struggle with accurately identifying the fallacious steps. Most LLMs only achieved an overall accuracy rate of less than 80%, suggesting that LLMs could lack sufficient logical verification abilities. Thus, we should be more cautious about the self-verification methods of LLMs. Second, the performance of LLMs can be remarkably imbalanced in different types of fallacies. Most LLMs perform much worse at identifying fallacies related to logical structure than those related to content, pointing toward key directions for improving LLMs' verification and reasoning abilities. Third, we also find that LLMs encounter challenges when it comes to classifying different types of fallacies. Presenting LLMs with the definitions of fallacies does not appear to improve their ability to recognize fallacies. This raises a call for further research to delve into the underlying mechanisms through which LLMs understand reasoning and fallacies. In summary, we present a comprehensive evaluation of the verification abilities of LLMs and highlight their limitations in identifying fallacies, urging the community to apply the self-verification methods with caution.

2 Related Work

2.1 Language Models for Logical Reasoning

Compared with traditional symbol-based logic reasoning systems, using language models to directly reason over natural language is a more flexible and popular approach (Yu et al., 2023; Yang et al., 2023b). Many efforts have been devoted to improving the logical reasoning abilities of language models from different perspectives, including fine-tuning methods (Clark et al., 2020; Dalvi et al., 2021), pre-training methods (Pi et al., 2022; Jiao et al., 2022), and modular methods (Hong et al., 2022; Yang et al., 2022). As the model scale increases, specially designed prompts (e.g., Chain-of-Thought prompts (Wei et al., 2022b)) can elicit the step-by-step reasoning abilities of LLMs, achieving remarkable improvement in multiple reasoning tasks (Chu et al., 2023). However, some studies find that LLMs still struggle with complex logical reasoning problems (Arkoudas, 2023) and could be susceptible to logical fallacies (Payandeh et al., 2023). There is still a lack of comprehensive research to investigate the understanding of LLMs on different logical fallacies.

2.2 Self-Verification with Large Language Models

A prevalent method to enhance the capacity of LLMs is through learning and correction via high-quality verification feedback. This feedback proves instrumental in various aspects of model optimization. For instance, it can be used to fine-tune the models (Ouyang et al., 2022; Scheurer et al., 2023; Chen et al., 2023a; Zelikman et al., 2022; Wang et al., 2023; Bai et al., 2022; Huang et al., 2022). In addition, verification feedback can be utilized to re-rank the model outputs (Lightman et al., 2023; Weng et al., 2022; He et al., 2023; Ni et al., 2023; Ling et al., 2023). It can also be em-

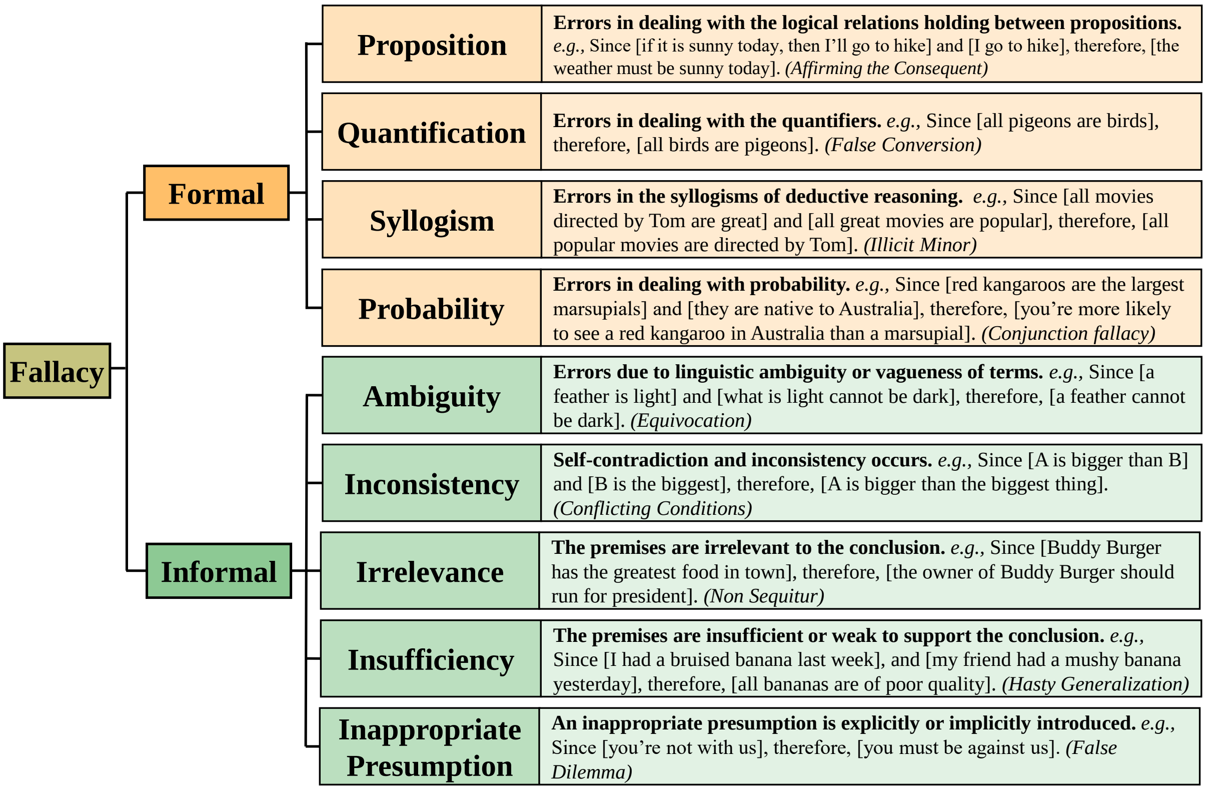

Figure 2: The hierarchical taxonomy of fallacies. For each sub-category, we present its definition and an example of a fallacy within the sub-category. We use square brackets to indicate the premises and conclusions.

ployed to refine the outputs (Madaan et al., 2023; Shinn et al., 2023; Chen et al., 2023b) and guide the generation process of the models (Yao et al., 2023; Xie et al., 2023). A straightforward approach to collecting feedback is to collect directly from humans (Ouyang et al., 2022; Fernandes et al., 2023), but this could prove to be costly and unable to provide instant feedback. Alternatively, collecting feedback from external tools (Gou et al., 2023; Chern et al., 2023) or metrics (Jung et al., 2022) could be more feasible but limited to specific tasks. Thus, some researchers turn to use large language models themselves to provide verification feedback (Ling et al., 2023; Weng et al., 2022; Xie et al., 2023; Madaan et al., 2023; Miao et al., 2023), which is more scalable. Nonetheless, recent papers (Huang et al., 2023; Valmeekam et al., 2023) have raised doubts about the self-verification abilities of the LLMs. For instance, Huang et al. (2023) find that LLMs struggle to self-correct their responses without external feedback. However, they still leave some open questions for subsequent research, such as what exactly the performance of LLMs is to verify a single reasoning step. In this paper, we delve deeply into this subject, critically examining the verification abilities of LLMs from the perspective of logical reasoning.

Category	# Fallacy
Formal	**24**
proposition (prop.)	6
quantification (quant.)	6
syllogism (syl.)	8
probability (prob.)	4
Informal	**208**
ambiguity (amb.)	15
inconsistency (incon.)	3
irrelevance (irrel.)	78
insufficiency (insuf.)	58
inappropriate presumption (inappr.)	54

Table 1: Distribution of 232 fallacies in FALLACIES.

3 FALLACIES

This section outlines the design principles and process of constructing our dataset FALLACIES.

3.1 Design Principles

Covering more comprehensive error types: We propose to evaluate the verification abilities of LLMs on a wider range of types of errors. Relying solely on a single existing dataset of logical reasoning might pose a challenge in drawing comprehensive conclusions, as it may struggle to cover the types of errors that can occur in realistic scenarios. For example, synthetic datasets (e.g., ProofWriter (Tafjord et al., 2021)) are generated from fixed logic templates and vocabularies, which could leave out the reasoning errors caused by verbal ambiguity.

Hierarchical fallacy taxonomy: We propose to categorize reasoning errors at a fine-grained, hierarchical level. This approach enables a more thorough assessment of the performance of LLMs across various types of fallacies and offers a more comprehensive perspective on their effectiveness. Meanwhile, we ensure that each reasoning step exclusively pertains to a single type of fallacy to prevent different types of fallacies from potentially confounding one another.

Clarifying the premises and conclusions: The third principle is to indicate the premises and conclusions of the reasoning step explicitly. Reasoning is the process of concluding based on known information, and the validity of reasoning cannot be properly judged without providing premise information. Factors like linguistic omissions and implied background knowledge can complicate reasoning in natural language. Consequently, in our dataset, we strive to ensure that each step is a complete unit of reasoning that contains enough information to verify its correctness.

3.2 Taxonomy of Fallacy

A fallacy is an error in reasoning (Bennett, 2012; Jevons, 1872). It focuses more on whether the conclusion of an argument can be logically deduced from the premises rather than on factual errors or other aspects (Lau, 2011). Identifying logical fallacies is essential for a judicious reasoning system. Without the ability to identify logical fallacies, a reasoning system may lack essential critical thinking skills, leaving it susceptible to illogical arguments and deliberate manipulation.

Fallacies can be classified into two primary categories: **Formal Fallacies**, errors due to invalid logical structures or inference patterns, and **Informal Fallacies**, errors due to the content of premises and conclusions. Based on specific causes of error, we divide each main category into several subcategories. Each subcategory contains several finest-grained fallacies, divided based on more detailed causes of errors. Figure 2 demonstrates our hierarchical taxonomy of fallacies. We derive this taxonomy by integrating multiple scholarly resources (Fieser and Dowden, 2011; Magnus, 2005; Rescher and Schagrin, 2023).

3.3 Data Collection

We first collect 232 types of fallacies from classic academic books (Bennett, 2012; Fieser and Dowden, 2011). The authors of these books carefully collect these fallacies from many available academic resources (including peer-reviewed journals, encyclopedias, and books), covering a substantial portion of common errors in logical reasoning. We collect the definition of each fallacy, i.e., an article containing the description and examples of that fallacy. These definitions are used in academic books to help humans understand the meaning of fallacies. Figure 3 in the Appendix shows an example of the definition. We then assign these 232 fallacies to appropriate categories according to the taxonomy in Sec. 3.2. Table 1 shows the distribution of fallacies. Detailed categorizations and descriptions of these fallacies are in Appendix Table 11.

Then, we collect fallacious and correct reasoning steps. Considering that directly creating fallacious steps from scratch can be challenging, we adopt a strategy in which powerful LLMs first generate candidates, and then we let human experts revise them. Specifically, we ask GPT-4 (OpenAI, 2023) to generate fallacious steps based on the collected definitions of fallacies. To generate diverse steps covering a wider range of domains, we explicitly require the model to generate around a theme, a randomly sampled noun from ConceptNet (Speer et al., 2017). Detailed prompt can be found in Appendix A. Subsequently, human experts carefully proofread and refine these candidates to ensure they fall within the corresponding fallacies. Meanwhile, we require experts to make each step a single inference (Dagan et al., 2013), rather than complex reasoning that involves multiple intermediate steps.

We require the experts to fix the fallacious steps to collect the correct contrastive steps. The experts make as few modifications as possible to turn the fallacious steps into correct ones, which do not contain any reasoning errors. For instance, the fallacious step of the fallacy of *Affirming the Consequent* in Figure 2 can be fixed into *"Since [if it is sunny today, then I'll go to hike] and [It is sunny today], therefore, [I'll go to hike today]."* More contrastive samples can be found in Table 4.

Dataset	Number of Fallacies	Number of Steps	Taxonomy of Fallacy	Granularity of Reasoning	Explicit Premises and Conclusions	Identifying Fallacy from Correct Reasoning
Stab and Gurevych (2017)	1	1,029	No	Coarse	No	**Yes**
Habernal et al. (2018)	1	2,085	No	Coarse	No	**Yes**
Jin et al. (2022)	13	2,449	Coarse	Coarse	No	No
FALLACIES (Ours)	**232**	**4,640**	**Fine & Hierarchical**	**Fine**	**Yes**	**Yes**

Table 2: Comparison of FALLACIES with existing fallacy-related datasets.

In this way, we obtain ten fallacious and ten correct steps for each of the 232 types of fallacies. To check data quality, we ask three additional experts to re-annotate 50 randomly sampled steps. They annotate each step as a correct step, a fallacious step that belongs to the corresponding fallacy, or a fallacious step that does not belong to the corresponding fallacy. Their average agreement with the labels achieves 0.856 (Cohen's Kappa), indicating the high quality of our data.

3.4 Comparison with Existing Dataset

We compare our dataset and existing fallacy-related datasets in Table 2. FALLACIES holds significant advantages across multiple dimensions. It encompasses a broader spectrum of fallacy types, presenting a fine-grained and hierarchical taxonomy of fallacies. Additionally, FALLACIES stands out for its clarity and subtlety, avoiding ambiguous judgments about the correctness of reasoning. We make explicit the premises and conclusions of each reasoning step and the granularity of reasoning in FALLACIES is finer. In contrast to the most recent dataset (Jin et al., 2022), which contained only fallacious reasoning, our dataset includes both correct and fallacious reasoning steps.

4 Experiments

4.1 Models and Prompts

We test a range of common LLMs, including GPT-4 (OpenAI, 2023), GPT-3.5 (Peng et al., 2023), Llama2 (Touvron et al., 2023), Vicuna (Zheng et al., 2023), WizardLM (Xu et al., 2023), Flan-T5 (Chung et al., 2022), Falcon (Almazrouei et al., 2023), Baichuan2 (Yang et al., 2023a), ChatGLM (Du et al., 2022), and InternLM (Team, 2023). We use the default generation parameters in the models' configuration files.[2] Details about the models (e.g., the version numbers) can be found in Table 7 in the Appendix. We evaluate the models

Prompt for Identifying Fallacious Steps

Is the following reasoning step correct? You can only answer "Yes" or "No."
{reasoning step}

using the same simple prompt in the zero-shot setting. For the prompt selection, we followed the experience of previous work (Liu et al., 2023b) and carefully designed and experimented with several different styles of prompts. In the end, we chose a prompt that worked well for all models as our prompt. Appendix B presents a detailed ablation analysis of prompt selection, as well as experiments under few-shot setting.

4.2 Metrics

We evaluate LLMs on 4,640 steps in FALLACIES. We calculate the accuracy for each type of fallacy separately. The accuracy of a higher-level category is the (macro) average of the accuracies of its subcategories. Ultimately, we take the average accuracy on formal and informal categories as the overall accuracy.

4.3 Can LLMs accurately identify fallacious steps?

Table 3 shows the accuracies of the different models for identifying fallacious steps on FALLACIES. Based on the experimental results, we have the following observations.

Identifying fallacious steps is still challenging for LLMs. Most LLMs struggle with accurately identifying the fallacious steps. As shown in Table 3, the performance of most LLMs in this binary classification task ranges from 60% to 80%, indicating the complexity of this task. The best result is achieved by GPT-4, which reaches an overall average accuracy of 87.7%. However, this may still fall short of guaranteeing the validity of the self-verification approach since it is only the performance of identifying single-step fallacies. For a long argument comprising multiple reasoning steps, the overall verification performance for the

[2]For GPT-4 and GPT-3.5, we set the temperature parameter to 0. Experiments were conducted mainly in Nov. 2023.

Model	Formal					Informal						Avg.
	prop.	quant.	syl.	prob.	Avg.	amb.	incon.	irrel.	insuf.	inappr.	Avg.	
Random	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Flan-T5-Large	49.2	59.2	78.1	57.5	61.0	64.3	86.7	72.8	74.8	69.5	73.6	67.3
Flan-T5-xl	48.3	78.3	78.8	66.3	67.9	67.7	83.3	72.8	74.5	72.1	74.1	71.0
Flan-T5-xxl	46.7	70.8	80.6	57.5	63.9	67.0	78.3	72.6	75.7	71.2	73.0	68.4
Llama2-7B	59.2	63.3	58.8	63.7	61.3	67.7	70.0	72.7	75.3	73.4	71.8	66.5
Llama2-13B	55.8	62.5	56.2	65.0	59.9	58.3	58.3	66.9	65.1	67.4	63.2	61.5
Llama2-70B	58.3	79.2	79.4	76.2	73.3	82.0	90.0	90.3	90.4	**88.2**	**88.2**	80.7
Baichuan2-7B	55.8	52.5	52.5	50.0	52.7	53.3	51.7	51.7	51.9	52.9	52.3	52.5
Baichuan2-13B	59.2	75.8	79.4	57.5	68.0	77.0	78.3	83.1	85.8	80.2	80.9	74.4
ChatGLM-6B	60.0	52.5	60.6	50.0	55.8	52.3	56.7	54.7	53.0	55.5	54.4	55.1
ChatGLM2-6B	60.8	61.7	69.4	55.0	61.7	65.3	83.3	76.7	77.9	72.5	75.2	68.4
InternLM-7B	51.7	70.0	71.9	48.8	60.6	70.7	83.3	74.9	75.9	72.2	75.4	68.0
InternLM-20B	57.5	67.5	74.4	55.0	63.6	73.0	85.0	77.0	77.8	74.9	77.5	70.6
Falcon-7B	31.7	45.8	41.9	52.5	43.0	64.7	66.7	72.9	75.3	71.2	70.2	56.6
WizardLM-13B	63.3	66.7	70.0	70.0	67.5	82.0	86.7	87.6	91.8	84.4	86.5	77.0
Vicuna-7B	65.0	75.0	73.8	60.0	68.4	77.0	83.3	84.0	85.3	81.7	82.2	75.3
Vicuna-13B	75.0	69.2	72.5	62.5	69.8	81.0	83.3	88.9	91.5	87.3	86.4	78.1
Qwen-14B	70.0	78.3	83.1	67.5	74.7	83.0	**91.7**	88.4	91.4	86.6	**88.2**	81.5
GPT-3.5	73.3	72.5	74.4	76.2	74.1	**84.3**	90.0	**90.6**	90.0	84.5	87.9	81.0
GPT-4	**92.5**	**84.2**	**87.5**	**88.8**	**88.2**	83.0	86.7	88.8	**92.1**	85.2	87.2	**87.7**

Table 3: Accuracy results (%) of identifying fallacious steps on FALLACIES.

Formal → Proposition → Denying the Antecedent
- Since [if you have a pinna, then you can hear] and [you do not have a pinna], therefore, [you cannot hear]. **Prediction:** Yes. ✗
- Since [if you do not have a pinna, then you cannot hear] and [you do not have a pinna], therefore, [you cannot hear]. **Prediction:** Yes. ✓

Formal → Syllogism → Exclusive Premises
- Since [no psychologists are proponents of shock therapy] and [some proponents of shock therapy are not doctors], therefore, [some doctors are not psychologists]. **Prediction:** Yes. ✗
- Since [no psychologists are proponents of shock therapy] and [some doctors are proponents of shock therapy], therefore, [some doctors are not psychologists]. **Prediction:** Yes. ✓

Informal → Irrelevance → Appeal to Pity
- Since [my horse's stirrups are broken] and [I thus had to make a pitiful 10-mile walk in the pouring rain to get home], therefore, [the broken stirrups should be replaced by the store for free]. **Prediction:** No.✓
- Since [my horse's stirrups are broken] and [the stirrups were under warranty], therefore, [the broken stirrups should be replaced by the store for free]. **Prediction:** Yes. ✓

Informal → Insufficiency → Questionable Cause
- Since [grebos are often seen during rainstorms] and [rainstorms cause floods], therefore, [grebos cause floods]. **Prediction:** No.✓
- Since [grebos are often seen during rainstorms] and [rainstorms cause floods],therefore,[grebos can be seen during floods]. **Prediction**:No. ✗

Table 4: Contrastive reasoning steps in FALLACIES and verification predictions of GPT-4. Fallacious steps (the first sentence of each cell) are in red and correct ones (the second sentence of each cell) in green. ✓ and ✗ indicates whether the prediction matches the label or not.

entire argument could be the product of the verification performances of each individual step. Consequently, the overall verification performance of the argument might decrease exponentially with the number of steps in it. These results suggest that we need further research on existing self-verification methods to understand how they work and under what situations they can provide correct verification feedback.

Formal fallacy is more difficult than informal fallacy for LLMs. The performance of most LLMs on formal fallacies is much lower than on informal fallacies. For example, GPT-3.5 achieves 87.9% accuracy on formal fallacies, while it achieves only 74.1% accuracy on informal fallacies. As stated in Sec. 3.2, formal fallacies are more related to the logical structure of reasoning and require a greater emphasis on the understanding and utilization of formal logic. On the other hand, informal fallacies focus more on the content and semantics of the reasoning and may involve factors such as linguistic expression, semantic understanding, and semantic relevance. Therefore, the differences in the performance of LLMs across different types of fallacies may stem from their ability to understand logical structures and semantic meanings, i.e., the models present some challenges in dealing with the logical structures under the natural language, whereas they perform better in dealing with issues related to content and semantics. Taking a more fine-grained view, among formal fallacies, most models typically perform well on syllogism fallacies and poorly on proposition and probability fallacies. For informal fallacies, most models perform worse on ambiguity fallacies and have closer performance on the remaining four sub-categories.

The performance of the same model on different types of fallacies can be remarkably im-

Model	Formal					Informal						Avg.
	prop.	quant.	syl.	prob.	Avg.	amb.	incon.	irrel.	insuf.	inappr.	Avg.	
Random	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
Flan-T5-Large	15.0	1.7	0.0	0.0	4.2	0.0	0.0	7.7	4.7	3.9	3.2	3.7
Flan-T5-xl	21.7	0.0	0.0	5.0	6.7	0.0	0.0	16.4	10.7	7.8	7.0	6.8
Flan-T5-xxl	28.3	26.7	0.0	22.5	19.4	2.7	**16.7**	8.5	9.5	7.4	8.9	14.2
Llama2-7B	16.7	0.0	0.0	0.0	4.2	0.0	0.0	1.0	0.0	0.0	0.2	2.2
Llama2-13B	8.3	0.0	0.0	0.0	2.1	0.0	0.0	0.0	0.0	0.2	0.0	1.1
Llama2-70B	30.0	0.0	0.0	0.0	7.5	4.7	0.0	14.6	8.1	10.0	7.5	7.5
S Baichuan2-7B	0.0	0.0	1.2	0.0	0.3	0.0	0.0	3.5	0.9	3.0	1.5	0.9
Baichuan2-13B	0.0	0.0	0.0	2.5	0.6	0.0	0.0	13.7	3.1	4.6	4.3	2.5
ChatGLM-6B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ChatGLM2-6B	0.0	0.0	0.0	2.5	0.6	0.0	0.0	1.0	0.0	0.0	0.2	0.4
InternLM-7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.4	0.1	0.1
InternLM-20B	3.3	1.7	2.5	12.5	5.0	0.7	0.0	4.9	1.0	2.2	1.8	3.4
Falcon-7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	1.0	0.0	0.2	0.1
WizardLM-13B	15.0	0.0	0.0	7.5	5.6	0.0	0.0	1.0	0.5	0.2	0.3	3.0
Vicuna-7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.1	0.1
Vicuna-13B	31.7	0.0	0.0	0.0	7.9	0.0	0.0	2.2	0.0	0.6	0.5	4.2
Qwen-14B	10.0	0.0	0.0	20.0	7.5	0.0	0.0	5.5	2.1	0.6	1.6	4.6
GPT-3.5	40.0	20.0	3.8	22.5	21.6	16.0	0.0	22.7	16.4	9.6	12.9	17.3
GPT-4	**58.3**	**31.7**	**20.0**	**55.0**	**41.2**	**40.0**	**16.7**	**31.9**	**27.2**	**27.8**	**28.7**	**35.0**

Table 5: Accuracy results (%) on classifying the fallacy types of fallacious steps.

balanced. For example, the model Qwen-14B achieves an impressive 91.7% accuracy on the inconsistency fallacies but drops to a mere 67.5% on probability fallacies. This highlights a key observation that models may have superior verification abilities on some types of fallacy but are not necessarily equally well adapted to other types of fallacy. Such imbalance performance could be particularly important for practical applications, as different types of fallacies are not always evenly distributed in a given dataset or scenario, and certain types of fallacies might be more frequent. Therefore, we should not rely on a particular dataset when using or researching self-verification methods. Instead, we need to comprehensively consider the performance of the methods in dealing with different fallacies and scenarios.

GPT-4 achieves superior performance, particularly in identifying formal fallacies. Compared to the other models, GPT-4 achieves the best results in overall average accuracy. This gap is insignificant on informal fallacies, where models such as WizardLM-13B, Vicuna-13B, and Qwen-14B have comparable or even better average accuracies than GPT-4 on informal fallacies. However, on formal fallacies, GPT-4's accuracy is 13.5% higher than the second-best model (88.2% for GPT-4 compared to 74.7% for the second-best model Qwen-14B). The results suggest that GPT-4 demonstrates superior abilities in identifying fallacies related to logical structures than other LLMs. However, there is still room to improve the performance of GPT-4. Table 4 demonstrates the prediction results of GPT-4 in some cases. It can be seen that GPT-4 can also fail on some challenging samples.

Prompt for Classifying Fallacy Types

You are a logical fallacy classifier. Given an incorrect reasoning step, your task is to identify its type of fallacy. Answer by choosing one of these fallacies:
{(1) Affirming the Consequent
(2) Denying the Antecedent
· · · · · ·
(232) Alleged Certainty}
You should only answer the name of the fallacy.
What type of fallacy does the following reasoning step belong to?
{reasoning step}

4.4 Can LLMs distinguish types of logical fallacies?

In addition to whether LLMs can identify between correct and incorrect, we are also interested in whether LLMs can distinguish between different types of fallacies. Classifying types of fallacies requires the model to understand not only the pattern of errors in reasoning but also where the errors occur and why they may occur, which requires a higher level of reasoning ability. Given an error reasoning step, we require the model to recognize the error pattern within it and classify it as one of the 232 fallacies. We conduct experiments on the 2,320 fallacy steps in FALLACIES and calculate the macro average accuracy. Previous work has also explored this task and named it "logical fallacy

Model	Formal	Informal	Avg.
Flan-T5-Large	62.3 (+1.3)	65.5 (-8.1)	63.9 (-3.4)
Flan-T5-xl	56.6 (-11.3)	67.1 (-7.0)	61.8 (-9.2)
Flan-T5-xxl	66.2 (+2.3)	73.1 (+0.1)	69.7 (+1.3)
Llama2-7B	50.9 (-10.4)	52.0 (-19.8)	51.5 (-15.0)
Llama2-13B	57.9 (-2.0)	53.2 (-10.0)	55.6 (-5.9)
Llama2-70B	54.0 (-19.3)	56.9 (-31.3)	55.4 (-25.3)
Baichuan2-7B	52.8 (+0.1)	58.0 (+5.7)	55.4 (+2.9)
Baichuan2-13B	50.1 (-17.9)	52.7 (-28.2)	51.4 (-23.0)
ChatGLM-6B	54.6 (-1.2)	58.0 (+3.6)	56.3 (+1.2)
ChatGLM2-6B	58.0 (-3.7)	61.9 (-13.3)	60.0 (-8.4)
InternLM-7B	55.1 (-5.5)	59.4 (-16.0)	57.2 (-10.8)
InternLM-20B	59.2 (-4.4)	68.6 (-8.9)	63.9 (-6.7)
Falcon-7B	40.2 (-2.8)	46.9 (-23.3)	43.5 (-13.1)
WizardLM-13B	74.2 (+6.7)	82.9 (-3.6)	78.5 (+1.5)
Vicuna-7B	70.4 (+2.0)	78.0 (-4.2)	74.2 (-1.1)
Vicuna-13B	61.6 (-8.2)	67.7 (-18.7)	64.6 (-13.5)
Qwen-14B	71.1 (-3.6)	79.2 (-9.0)	75.2 (-6.3)
GPT-3.5	75.1 (+1.0)	75.7 (-12.2)	75.4 (-5.6)
GPT-4	91.2 (+3.0)	84.9 (-2.3)	88.0 (+0.3)

Table 6: Accuracy results (%) of identifying fallacious steps given the definitions of corresponding fallacies. We present the performance variations in the case with the definitions in parentheses compared to the case without definitions.

detection" (Jin et al., 2022). However, they only classify over 13 types of fallacies, whereas our task requires classifying over 232 types. Our task is more challenging and allows for fine-grained and hierarchical analysis. We evaluate the model using the same prompt in a zero-shot setting.

Table 5 shows the results. First, this task is very challenging for the existing LLMs. The models' performances are poor, with less than 10% overall accuracy, except for three models, GPT-4, GPT-3.5, and Flan-T5-xxl. For example, Vicuna-13B can achieve 78.1% accuracy on identifying fallacious steps, but only 4.2% accuracy on this task. Among all the models, GPT-4 performs the best, achieving an overall accuracy of 35.0%. This indicates that GPT-4 can recognize and classify the reasoning error patterns to a certain degree, showing a stronger reasoning ability than other models. Nevertheless, there is still substantial room for improvement. Further research may be required to achieve higher accuracy and enhance reasoning ability.

It is worth noting that on this task, the models typically perform better on formal than informal fallacies. This is inconsistent with the observation in identifying fallacious steps (Table 3). When identifying fallacious steps, models typically perform worse on formal fallacies and better on informal fallacies. There could be various reasons for this inconsistency. One possible explanation is that the models might just know the names of the fallacies rather than having an in-depth understanding of what these fallacies are. In determining whether there is an error in reasoning, the models might not be relying directly on their understanding of reasoning or fallacies but on some other abilities, which thus contributes to this inconsistency.

Prompt for Identifying Fallacious Steps Given Fallacy Definition

You are a trained model capable of identifying the logical fallacy known as {fallacy}.
This is the definition for {fallacy}: {fallacy definition}
Is the following reasoning step correct? You can only answer "Yes" or "No."
{reasoning step}

4.5 Can LLMs understand fallacies better from their definitions?

We test whether LLMs can perform better in identifying fallacious steps given the fallacy definition. Specifically, for each step in FALLACIES, we add the name and definition of its corresponding fallacy to the prompt of LLMs in advance. We then ask the model to determine whether the reasoning step is correct as before. As stated in Sec. 3.3, the definitions of fallacies are gathered from academic sources. An example of the definition can be found in Figure 3 in the Appendix.

We can observe a surprising trend by analyzing the results in Table 6. When definitions of corresponding fallacies are provided in advance, most models' performance decreases rather than improves. For instance, the overall accuracy of Vicuna-13B decreases from 78.1% to 64.6%, with a 13.5% decrease. These results suggest that providing definitions may hurt models' performance.

The reasons for this phenomenon deserve further exploration. One possible reason is that, in pre-trained data, the definitions and the fallacies themselves may not co-occur frequently, resulting in a mismatch with the current setting. Moreover, the mechanism by which the models judge fallacies has not yet been fully clarified. In this case, even if the definitions are provided, the models fail to improve performance. Instead, the prompt becomes complex with the addition of definitions, possibly interfering with their decision-making process. These observations inspire us that more intensive research is called to understand what are the mechanisms by which LLMs understand the reasoning and fallacies.

5 Conclusion

In this paper, we take a closer look at the verification abilities of LLMs in logical reasoning. We collect a dataset containing 232 fallacies and propose a hierarchical taxonomy of fallacies. Our main experimental finding is that most LLMs still struggle to identify fallacies in logical reasoning accurately. This implies that it may be overly optimistic to expect LLMs to be able to inherently identify errors and conduct self-verification reasoning, at least with respect to the current state of technology. Therefore, researchers and practitioners should be more cautious in using self-verification methods. We call for more research to explore the potentials and limitations of self-verification methods to steer LLMs towards improved accuracy and reliability.

Acknowledgements

We appreciate the anonymous reviewers for their insightful comments. This work is supported by the Natural Science Foundation of China (NSFC No. 62176132) and the Guoqiang Institute of Tsinghua University (2020GQG0005).

Limitations

In this paper, we present an intensive investigation and evaluation of the verification abilities of Large Language Models (LLMs) for logical reasoning. Although we have reached some findings, we are also aware that there is still some room for improvement and future research areas worth exploring.

Firstly, we conduct experiments only on the most common LLMs. Such a limitation comes from two main reasons: one is due to the limitation of computational resources of our research team; and the other is the barrier of access to certain closed-source models. This results in our inability to perform detailed experiments on all types and all scales of models. Thus, our results may not fully reflect the abilities of all LLMs. In future studies, it is worthwhile to turn our attention to more types and scales of models to provide a more comprehensive evaluation of their abilities.

Second, our study focused mainly on the aspect of logical reasoning. Reasoning in real-world applications often encompasses other types of reasoning, such as numerical reasoning. It would be interesting to extend our research to more types of reasoning. By doing so, we can reveal the boundaries of the abilities of LLMs in these areas, which can enhance our deeper understanding of the performance of LLMs. Meanwhile, it would also help us to figure out how we can improve the accuracy and robustness of the reasoning of LLMs.

This article follows the ACL Code of Ethics. To the best of our knowledge, our work is foundational research, and we do not find obvious risks related to malicious harmful effects, environmental impact, fairness considerations, or privacy considerations.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Konstantine Arkoudas. 2023. GPT-4 can't reason. *CoRR*, abs/2308.03762.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073.

Bo Bennett. 2012. *Logically fallacious: the ultimate collection of over 300 logical fallacies (Academic Edition)*. eBookIt. com.

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukosiute, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. 2022. Measuring progress on scalable oversight for large language models. *CoRR*, abs/2211.03540.

Angelica Chen, Jérémy Scheurer, Tomasz Korbak, Jon Ander Campos, Jun Shern Chan, Samuel R. Bowman, Kyunghyun Cho, and Ethan Perez. 2023a. Improving code generation by training with natural language feedback. *CoRR*, abs/2303.16749.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023b. Iterative translation refinement with large language models. *CoRR*, abs/2306.03856.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative AI - A tool augmented framework for multi-task and multi-domain scenarios. *CoRR*, abs/2307.13528.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3882–3890. ijcai.org.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. 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 7358–7370. 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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 320–335. Association for Computational Linguistics.

Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José G. C. de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, and André F. T. Martins. 2023. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *CoRR*, abs/2305.00955.

James Fieser and Bradley Dowden. 2011. Internet encyclopedia of philosophy: Fallacies.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. CRITIC: large language models can self-correct with tool-interactive critiquing. *CoRR*, abs/2305.11738.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana. Association for Computational Linguistics.

Hangfeng He, Hongming Zhang, and Dan Roth. 2023. Rethinking with retrieval: Faithful large language model inference. *CoRR*, abs/2301.00303.

Ruixin Hong, Hongming Zhang, Xintong Yu, and Changshui Zhang. 2022. METGEN: A module-based entailment tree generation framework for answer explanation. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1887–1905. Association for Computational Linguistics.

Ruixin Hong, Hongming Zhang, Hong Zhao, Dong Yu, and Changshui Zhang. 2023. Faithful question answering with monte-carlo planning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages

3944–3965. Association for Computational Linguistics.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *CoRR*, abs/2210.11610.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

William Stanley Jevons. 1872. *Elementary lessons in logic: Deductive and inductive: With copious questions and examples, and a vocabulary of logical terms*. Macmillan.

Fangkai Jiao, Yangyang Guo, Xuemeng Song, and Liqiang Nie. 2022. Merit: Meta-path guided contrastive learning for logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3496–3509. Association for Computational Linguistics.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1266–1279. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Measuring faithfulness in chain-of-thought reasoning. *CoRR*, abs/2307.13702.

Joe YF Lau. 2011. *An introduction to critical thinking and creativity: Think more, think better*. John Wiley & Sons.

Jan Leike and Ilya Sutskever. 2023. Introducing superalignment. *OpenAI*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *CoRR*, abs/2305.20050.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive verification of chain-of-thought reasoning. *CoRR*, abs/2306.03872.

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

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):195:1–195:35.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *CoRR*, abs/2303.17651.

PD Magnus. 2005. forall x: An introduction to formal logic.

John McCarthy. 1989. Artificial intelligence, logic and formalizing common sense. *Philosophical Logic and Artificial Intelligence*, pages 161–190.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2023. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. *CoRR*, abs/2308.00436.

Ansong Ni, Srini Iyer, Dragomir Radev, Veselin Stoyanov, Wen-Tau Yih, Sida I. Wang, and Xi Victoria Lin. 2023. LEVER: learning to verify language-to-code generation with execution. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 26106–26128. PMLR.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *CoRR*, abs/2305.12295.

Amirreza Payandeh, Dan Pluth, Jordan Hosier, Xuesu Xiao, and Vijay K. Gurbani. 2023. How susceptible are llms to logical fallacies? *CoRR*, abs/2308.09853.

Andrew Peng, Michael Wu, John Allard, Logan Kilpatrick, and Steven Heidel. 2023. Gpt-3.5 turbo fine-tuning and api updates. `https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates`.

Xinyu Pi, Wanjun Zhong, Yan Gao, Nan Duan, and Jian-Guang Lou. 2022. Logigan: Learning logical reasoning via adversarial pre-training. In *NeurIPS*.

Nicholas Rescher and Morton L. Schagrin. 2023. Fallacy. Encyclopedia Britannica.

Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2023. Training language models with language feedback at scale. *CoRR*, abs/2303.16755.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Christian Stab and Iryna Gurevych. 2017. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, Valencia, Spain. Association for Computational Linguistics.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3621–3634. Association for Computational Linguistics.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. 2023. Can large language models really improve by self-critiquing their own plans? *CoRR*, abs/2310.08118.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13484–13508. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Yixuan Weng, Minjun Zhu, Shizhu He, Kang Liu, and Jun Zhao. 2022. Large language models are reasoners with self-verification. *CoRR*, abs/2212.09561.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, Min-Yen Kan, Junxian He, and Qizhe Xie. 2023. Decomposition enhances reasoning via self-evaluation guided decoding. *CoRR*, abs/2305.00633.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *CoRR*, abs/2304.12244.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023a. Baichuan 2: Open large-scale language models. *CoRR*, abs/2309.10305.

Kaiyu Yang, Jia Deng, and Danqi Chen. 2022. Generating natural language proofs with verifier-guided search. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 89–105. Association for Computational Linguistics.

Zonglin Yang, Xinya Du, Rui Mao, Jinjie Ni, and Erik Cambria. 2023b. Logical reasoning over natural language as knowledge representation: A survey. *CoRR*, abs/2303.12023.

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

Fei Yu, Hongbo Zhang, and Benyou Wang. 2023. Nature language reasoning, A survey. *CoRR*, abs/2303.14725.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. In *NeurIPS*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685.

A Details of Data Collection

To generate candidate fallacious steps, we prompt GPT-4 with the following content. Figure 3 shows an example of the fallacy definition. Note that the steps generated by GPT-4 are only used as candidates. We then have human experts proofread and revise them to ensure the quality of the data. We invited 10 well-trained graduate students from universities as human expert annotators. They have passed the graduate school entrance exam and have well logical reasoning skills. They have undergone rigorous training in areas including Mathematical Logic, Computer Science, Programming, and Statistics.

Prompt for Generating Candidate Fallacious Steps

You are a faulty reasoner. I will describe a logical fallacy to you, and then you generate a reasoning step that belongs to this logical fallacy.
Here is the description: {fallacy definition}
Now, generate a reasoning step that contains this type of logical fallacy. The generated step should be related to this element: {entity}
Use square brackets [] to include propositions. The format of the output is "Since [XXX] and [XXX], therefore, [XXX]

In the end of the appendix, Table 11 shows the detailed categorizations and descriptions of the 232 fallacies in our dataset.

B Details of Experiments

B.1 Models

Table 7 shows the version and source URL of the LLMs used in our experiments. For all series of LLMs, we select their versions of instruction fine-tuned or chat fine-tuned, since these models are closest to realistic applications. We follow the licences of these models to use them.

B.2 Prompt Selection

To select suitable prompts and to explore the impact of prompts on model performance, we conduct an ablation study on the prompts. Table 8 demonstrates the prompts we used. Prompt 1 is the simplest prompt. The model is expected to answer "Yes" for the correct reasoning steps and "No" for the incorrect reasoning steps. Prompt 2, alternatively, replaces the response "Yes/No" with "True/False." Prompt 3 adopts a Chain-of-Thought-style prompt (Kojima et al., 2022) that allows the model to generate some relevant rationales before giving the prediction. Prompt 4 describes the task in more detail. Moreover, in contrast to Prompt 1, Prompt 4 requires the model to answer "Yes" for incorrect reasoning steps (which contain logical fallacies) and "No" to correct reasoning steps (which do not contain logical fallacies).

Table 9 demonstrates the results. We can find that Prompt 1 works well for all models. For Prompt 2, some of the models show significant performance degradation after replacing the response words. Prompt 3 introduces chains of thought, but does not achieve significant performance gains on all models. For Prompt 4, although it describes the task in more detail, it does not seem to help improve the LLMs' performance. Among all the models, GPT-4 exhibits the strongest robustness to prompts, achieving similar performance with different prompts. Comprehensively, we finally chose Prompt 1 as our prompt.

B.3 Few-shot Setting

We also conduct experiments under few-shot setting. We include four demonstrations in the prompt, covering the correct and fallacious steps related to formal and informal fallacies.

Prompt for Identifying Fallacious Steps under Few-shot Setting

Is the following reasoning step correct? You can only answer "Yes" or "No."
Since [If it's raining then the streets are wet] and [It's raining now], therefore, [The streets are wet].
Yes.
Since [I found a shell on the beach] and [this shell was beautifully shaped and colored], therefore, [all shells are beautifully shaped and colored].
No.
Since [I am at home or I am in the city] and [I am at home], therefore, [I am not in the city].
No.
Since [heavy snowfall often leads to traffic jams] and [traffic jams cause delays], therefore, [heavy snowfall can lead to delays].
Yes.
{reasoning step}

Table 10 demonstrates the performance of identifying fallacious steps under few-shot setting. We can find that most of the LLMs achieve comparable performance under few-shot setting compared to that under zero-shot setting. Moreover, our findings in Sec 4.3 stand under the few-shot setting as well.

The **description** of the fallacy of type "Affirming the Consequent" is "An error in formal logic where if the consequent is said to be true, the antecedent is said to be true, as a result."

The **abstract logical form** is

If P then Q.
Q.
Therefore, P.

The following are **some examples and their explanations** of the fallacy of type "Affirming the Consequent."

Example #1:

If taxes are lowered, I will have more money to spend.
I have more money to spend.
Therefore, taxes must have been lowered.

Explanation: I could have had more money to spend simply because I gave up crack-cocaine, prostitute solicitation, and baby-seal-clubbing expeditions.

Example #2:

If it's brown, flush it down.
I flushed it down.
Therefore, it was brown.

Explanation: No! I did not have to follow the, "if it's yellow, let it mellow" rule -- in fact, if I did follow that rule I would probably still be single. The stated rule is simply, "if it's brown" (the antecedent), then (implied), "flush it down" (the consequent). From this, we cannot imply that we can ONLY flush it down if it is brown.

Figure 3: The definition of the fallacy of "Affirming the Consequent", one of 232 types of fallacies in our dataset.

Model	**Version**	**URL**
Flan-T5-Large (Chung et al., 2022)	google/flan-t5-large	https://huggingface.co/google/flan-t5-large
Flan-T5-xl (Chung et al., 2022)	google/flan-t5-xl	https://huggingface.co/google/flan-t5-xl
Flan-T5-xxl (Chung et al., 2022)	google/flan-t5-xxl	https://huggingface.co/google/flan-t5-xxl
Llama2-7B (Touvron et al., 2023)	meta-llama/Llama-2-7b-chat-hf	https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
Llama2-13B (Touvron et al., 2023)	meta-llama/Llama-2-13b-chat-hf	https://huggingface.co/meta-llama/Llama-2-13b-chat-hf
Llama2-70B (Touvron et al., 2023)	meta-llama/Llama-2-70b-chat-hf	https://huggingface.co/meta-llama/Llama-2-70b-chat-hf
Baichuan2-7B (Yang et al., 2023a)	baichuan-inc/Baichuan2-7B-Chat	https://huggingface.co/baichuan-inc/Baichuan2-7B-Chat
Baichuan2-13B (Yang et al., 2023a)	baichuan-inc/Baichuan2-13B-Chat	https://huggingface.co/baichuan-inc/Baichuan2-13B-Chat
ChatGLM-6B (Du et al., 2022)	THUDM/chatglm-6b	https://huggingface.co/THUDM/chatglm-6b
ChatGLM2-6B (Du et al., 2022)	THUDM/chatglm2-6b	https://huggingface.co/THUDM/chatglm2-6b
InternLM-7B (Team, 2023)	internlm/internlm-chat-7b-v1_1	https://huggingface.co/internlm/internlm-chat-7b-v1_1
InternLM-20B (Team, 2023)	internlm/internlm-chat-20b	https://huggingface.co/internlm/internlm-chat-20b
Falcon-7B (Almazrouei et al., 2023)	tiiuae/falcon-7b-instruct	https://huggingface.co/tiiuae/falcon-7b
WizardLM-13B (Xu et al., 2023)	WizardLM/WizardLM-13B-V1.2	https://huggingface.co/WizardLM/WizardLM-13B-V1.2
Vicuna-7B (Zheng et al., 2023)	lmsys/vicuna-7b-v1.5	https://huggingface.co/lmsys/vicuna-7b-v1.5
Vicuna-13B (Zheng et al., 2023)	lmsys/vicuna-13b-v1.5	https://huggingface.co/lmsys/vicuna-13b-v1.5
Qwen-14B (Bai et al., 2023)	Qwen/Qwen-14B-Chat	https://huggingface.co/Qwen/Qwen-14B-Chat
GPT-3.5 (Peng et al., 2023)	gpt-3.5-turbo	https://platform.openai.com/docs/models/gpt-3-5
GPT-4 (OpenAI, 2023)	gpt-4	https://platform.openai.com/docs/models/gpt-4

Table 7: Detailed information about the models we experiment with.

No.	Content
Prompt 1	Is the following reasoning step correct? \n You can only answer "Yes" or "No".\n {reasoning step}
Prompt 2	Is the following reasoning step correct? \n You can only answer "True" or "False".\n {reasoning step}
Prompt 3	Is the following reasoning step correct? \n Let's think step by step and then answer "Yes" or "No".\n {reasoning step}
Prompt 4	You are a trained model capable of detecting reasoning errors and logical fallacies. \n As a detector, your task is to analyze the given reasoning steps and determine whether they involve any logical fallacies. \n If a logical fallacy is present, your response should be "Yes". \n If no logical fallacies are detected, your response should be "No".\n You can only answer "Yes" or "No". \n {reasoning step}

Table 8: The different prompts used to prompt large language models to identify fallacious steps.

Model	Prompt 1			Prompt 2			Prompt 3			Prompt 4		
	Formal	Informal	Avg.	Formal	Informal	Avg.	Formal	Informal	Avg.	Formal	Informal	Avg.
Flan-T5-Large	61.0	73.6	67.3	59.5	71.7	65.6	60.9	73.9	67.4	40.1	41.3	40.7
Flan-T5-xl	67.9	74.1	71.0	68.2	68.5	68.4	67.3	74.7	71.0	41.0	47.6	44.3
Flan-T5-xxl	63.9	73.0	68.4	65.1	74.2	69.6	63.4	73.6	68.5	43.1	59.1	51.1
Llama2-7B	61.3	71.8	66.5	50.7	53.3	52.0	49.8	52.2	51.0	50.0	49.9	50.0
Llama2-13B	59.9	63.2	61.5	57.5	67.3	62.4	49.8	51.0	50.4	50.2	51.1	50.7
Llama2-70B	73.3	88.2	80.7	72.3	87.8	80.1	63.8	67.7	65.7	54.7	59.1	56.9
Baichuan2-7B	52.7	52.3	52.5	60.6	75.0	67.8	54.4	71.1	62.7	45.3	46.5	45.9
Baichuan2-13B	68.0	80.9	74.4	61.7	82.2	72.0	65.9	75.9	70.9	50.9	49.9	50.4
ChatGLM-6B	55.8	54.4	55.1	42.9	48.1	45.5	53.3	61.3	57.3	48.2	46.3	47.3
ChatGLM2-6B	61.7	75.2	68.4	59.1	76.5	67.8	61.8	71.3	66.5	50.5	50.5	50.5
InternLM-7B	60.6	75.4	68.0	57.6	72.6	65.1	59.5	75.4	67.4	56.7	68.3	62.5
InternLM-20B	63.6	77.5	70.6	68.4	81.3	74.8	64.6	77.8	71.2	59.6	77.1	68.3
Falcon-7B	43.0	70.2	56.6	0.0	0.0	0.0	56.5	73.1	64.8	44.4	43.3	43.9
WizardLM-13B	67.5	86.5	77.0	44.1	73.5	58.8	68.5	85.9	77.2	56.6	76.2	66.4
Vicuna-7B	68.4	82.2	75.3	0.0	1.0	0.5	66.0	81.4	73.7	55.8	57.3	56.5
Vicuna-13B	69.8	86.4	78.1	74.2	85.0	79.6	60.6	81.4	71.0	55.7	67.6	61.7
Qwen-14B	74.7	88.2	81.5	64.1	67.9	66.0	60.2	79.3	69.7	57.8	72.3	65.1
GPT-3.5	74.1	87.9	81.0	80.9	88.2	84.6	78.6	88.8	83.7	53.1	57.7	55.4
GPT-4	88.2	87.2	87.7	89.8	87.4	88.6	89.0	87.0	88.0	82.3	90.8	86.6

Table 9: Accuracy results (%) on identifying fallacious steps using different prompts.

Model	Formal					Informal						Avg.
	prop.	quant.	syl.	prob.	Avg.	amb.	incon.	irrel.	insuf.	inappr.	Avg.	
Random	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Flan-T5-Large	58.3	64.2	71.9	62.5	64.2	62.3	85.0	73.8	73.3	69.7	72.8	68.5
Flan-T5-xl	47.5	74.2	81.9	68.8	68.1	65.3	85.0	70.0	70.3	70.6	72.3	70.2
Flan-T5-xxl	44.2	70.8	79.4	56.2	62.7	64.0	76.7	73.5	74.6	71.4	72.0	67.3
Llama2-7B	65.0	66.7	56.2	63.8	62.9	80.3	75.0	81.5	87.0	81.8	81.1	72.0
Llama2-13B	65.8	65.8	60.0	62.5	63.5	69.0	73.3	78.6	80.3	80.6	76.4	70.0
Llama2-70B	64.2	77.5	78.8	72.5	73.2	83.3	81.7	86.8	90.9	88.1	86.2	79.7
Baichuan2-7B	55.0	70.0	62.5	62.5	62.5	73.3	76.7	72.8	77.4	73.9	74.8	68.7
Baichuan2-13B	60.8	72.5	68.8	62.5	66.1	76.3	75.0	78.7	82.3	79.1	78.3	72.2
ChatGLM-6B	71.7	72.5	68.1	61.3	68.4	70.7	81.7	73.7	78.5	76.9	76.3	72.3
ChatGLM2-6B	64.2	65.8	78.8	57.5	66.6	74.0	83.3	79.7	81.6	76.2	79.0	72.8
InternLM-7B	60.0	68.3	76.9	47.5	63.2	69.3	80.0	74.1	75.6	71.8	74.2	68.7
InternLM-20B	58.3	69.2	78.8	60.0	66.6	69.3	80.0	74.9	77.6	75.8	75.5	71.1
Falcon-7B	56.7	58.3	55.0	53.8	55.9	70.0	71.7	74.0	73.0	72.5	72.2	64.1
WizardLM-13B	76.7	72.5	77.5	71.2	74.5	82.3	80.0	85.8	90.8	84.4	84.7	79.6
Vicuna-7B	59.2	72.5	73.1	60.0	66.2	76.7	81.7	84.2	87.2	83.5	82.6	74.4
Vicuna-13B	67.5	75.0	83.1	62.5	72.0	82.7	85.0	85.8	90.8	85.5	85.9	79.0
Qwen-14B	68.3	75.0	83.1	67.5	73.5	81.7	88.3	87.8	91.2	85.2	86.8	80.2
GPT-3.5	72.5	81.7	78.1	78.8	77.8	86.0	86.7	87.9	89.4	85.6	87.1	82.4
GPT-4	91.7	84.2	87.5	88.8	88.0	84.7	85.0	89.2	93.0	87.8	87.9	88.0

Table 10: Accuracy results (%) of identifying fallacious steps under few-shot setting.

Category	Name of Fallacy	Description of Fallacy
Formal-Proposition	Affirming the Consequent	An error in formal logic where if the consequent is said to be true, the antecedent is said to be true, as a result.
	Denying the Antecedent	It is a fallacy where in a standard if/then premise, the antecedent (what comes after the "if") is made not true, then it is concluded that the consequent (what comes after the "then") is not true.
	Negating Antecedent and Consequent	The valid form of this argument is as follows: If P then Q. Therefore, if not-Q then not-P.
	Commutation of Conditionals	Switching the antecedent and the consequent in a logical argument.
	Affirming a Disjunct	Making the false assumption that when presented with an either/or possibility, that if one of the options is true that the other one must be false.
	Denying a Conjunct	A formal fallacy in which the first premise states that at least one of the two conjuncts (antecedent and consequent) is false and concludes that the other conjunct must be true.
Formal-Quantification	False Conversion	The formal fallacy where the subject and the predicate terms of the proposition are switched (conversion) in the conclusion, in a proposition that uses "all" in its premise (type "A" forms), or "some/not" (type "O" forms).
	Unwarranted Contrast	Assuming that implicature means implication, when it logically does not.
	Quantifier Shift Fallacy	A fallacy of reversing the order of two quantifiers.
	Existential Fallacy	A formal logical fallacy, which is committed when a categorical syllogism employs two universal premises ("all") to arrive at a particular ("some") conclusion.
	Fallacy of Every and All	When an argument contains both universal quantifiers and existential quantifiers (all, some, none, every) with different meanings, and the order of the quantifiers is reversed.
	Illicit Contraposition	A formal fallacy where switching the subject and predicate terms of a categorical proposition, then negating each, results in an invalid argument form.
Formal-Syllogism	Fallacy of the Undistributed Middle	A formal fallacy in a categorical syllogism where the middle term, or the term that does not appear in the conclusion, is not distributed to the other two terms.
	Exclusive Premises	A standard form categorical syllogism that has two negative premises either in the form of "no X are Y" or "some X are not Y".
	Fallacy of Four Terms	This fallacy occurs in a categorical syllogism when the syllogism has four terms rather than the requisite three.
	Illicit Substitution of Identicals	A fallacy due to confusing the knowing of a thing (extension) with the knowing of it under all its various names or descriptions (intension).
	Illicit Minor	Any form of a categorical syllogism in which the minor term is distributed in the conclusion, but not in the minor premise.
	Illicit Major	Any form of a categorical syllogism in which the major term is distributed in the conclusion, but not in the major premise.
	Negative Conclusion from Affirmative Premises	The conclusion of a standard form categorical syllogism is negative, but both of the premises are positive.
	Affirmative Conclusion from a Negative Premise	The conclusion of a standard form categorical syllogism is affirmative, but at least one of the premises is negative.
Formal-Probability	Gamblers Fallacy	Reasoning that, in a situation that is pure random chance, the outcome can be affected by previous outcomes.
	Hot Hand Fallacy	The irrational belief that if you win or lose several chance games in a row, you are either "hot" or "cold," respectively, meaning that the streak is likely to continue and has to do with something other than pure probability.

	Conjunction Fallacy	This occurs when one estimates a conjunctive statement (this and that) to be more probable than at least one of its component statements.
	Disjunction Fallacy	It occurs when one estimates a disjunctive statement (this or that) to be less probable than at least one of its component statements.
Informal-Ambiguity	Argument of the Beard	When one argues that no useful distinction can be made between two extremes, just because there is no definable moment or point on the spectrum where the two extremes meet.
	Appeal to Extremes	Erroneously attempting to make a reasonable argument into an absurd one, by taking the argument to the extremes.
	Type Token Fallacy	The type-token fallacy is committed when a word can refer to either a type (representing an abstract descriptive concept) or a token (representing an object that instantiates a concept) and is used in a way that makes it unclear which it refers to.
	Use Mention Error	Confusing the word used to describe a thing, with the thing itself.
	Reification	When an abstraction is treated as if it were a concrete, real event or physical entity – when an idea is treated as if had a real existence.
	Fake Precision	Using implausibly precise statistics to give the appearance of truth and certainty, or using a negligible difference in data to draw incorrect inferences.
	No True Scotsman	When a universal ("all", "every", etc.) claim is refuted, rather than conceding the point or meaningfully revising the claim, the claim is altered by going from universal to specific, and failing to give any objective criteria for the specificity.
	Contextomy	Removing a passage from its surrounding matter in such a way as to distort its intended meaning.
	Stolen Concept Fallacy	Requiring the truth of the something that you are simultaneously trying to disprove.
	Anthropomorphism	The attributing of human characteristics and purposes to inanimate objects, animals, plants, or other natural phenomena, or to gods.
	Accent Fallacy	When the meaning of a word, sentence, or entire idea is interpreted differently by changing where the accent falls.
	Ambiguity Fallacy	When an unclear phrase with multiple definitions is used within the argument; therefore, does not support the conclusion.
	Alphabet Soup	The deliberate and excessive use of acronyms and abbreviations to appear more knowledgeable in the subject or confuse others.
	Equivocation	Using an ambiguous term in more than one sense, thus making an argument misleading.
	Modal Scope Fallacy	Making a formal argument invalid by confusing the scope of what is actually necessary or possible.
Informal-Inconsistency	Inconsistency	In terms of a fallacious argument, two or more propositions are asserted that cannot both possibly be true.
	Conflicting Conditions	When the argument is self-contradictory and cannot possibly be true.
	Kettle Logic	Making (usually) multiple, contradicting arguments, in an attempt to support a single point or idea.
Informal-Irrelevance	Political Correctness Fallacy	It is the assumption or admission that two or more groups, individuals, or ideas of groups or individuals, are equal, of equal value, or both true, based on the recent phenomenon of political correctness.
	Appeal to Complexity	Concluding that because you don't understand something, it must not be true, it's improbable, or the argument must be flawed.
	Statement of Conversion	Accepting the truth of a claim based on a conversion story without considering any evidence for the truth of the claim.
	Appeal to the Moon	Using the argument, "If we can put a man on the moon, we could..." as evidence for the argument.
	Quantum Physics Fallacy	Using quantum physics in an attempt to support your claim, when in no way is your claim related to quantum physics.

fact to fiction fallacy	Attempting to support a narrative or argument with facts that don't support the narrative or argument.
Non Sequitur	Evidence or reason is irrelevant or adds very little support to the conclusion.
Inflation of Conflict	Reasoning that because authorities cannot agree precisely on an issue, no conclusions can be reached at all, and minimizing the credibility of the authorities, as a result.
Argument by Fast Talking	When fast talking is seen as intelligence and/or confidence in the truth of one's argument; therefore, seen as evidence of the truth of the argument itself.
Appeal to Intuition	Evaluating an argument based on "intuition" or "gut feeling" that is unable to be articulated, rather than evaluating the argument using reason.
Appeal to Closure	Accepting evidence on the basis of wanting closure, or to be done with the issue.
Appeal to Definition	Using a dictionary's limited definition of a term as evidence that term cannot have another meaning, expanded meaning, or even conflicting meaning.
Spiritual Fallacy	Insisting that something meant to be literal is actually "spiritual" as an explanation or justification for something that otherwise would not fit in an explanation.
gish gallop	Overwhelming an interlocutor with as many arguments as possible, without regard for accuracy or strength of the arguments.
Denying the Correlative	Introducing alternatives when, in fact, there are none.
Red Herring	Attempting to redirect the argument to another issue to which the person doing the redirecting can better respond.
Strawman Fallacy	Substituting a person's actual position or argument with a distorted, exaggerated, or misrepresented version of the position of the argument.
Avoiding the Issue	When an arguer responds to an argument by not addressing the points of the argument.
Logic Chopping	Using the technical tools of logic in an unhelpful and pedantic manner by focusing on trivial details instead of directly addressing the main issue in dispute.
Meaningless Question	Asking a question that cannot be answered with any sort of rational meaning.
Failure to Elucidate	When the definition is made more difficult to understand than the word or concept being defined.
Argument by Gibberish	When incomprehensible jargon or plain incoherent gibberish is used to give the appearance of a strong argument, in place of evidence or valid reasons to accept the argument.
Hypnotic Bait and Switch	Stating several uncontroversially true statements in succession, followed by a claim that the arguer wants the audience to accept as true.
Traitorous Critic Fallacy	Responding to criticism by attacking a person's perceived favorability to an out-group or dislike to the in-group as the underlying reason for the criticism rather than addressing the criticism itself, and suggesting that they stay away from the issue and/or leave the in-group.
Having Your Cake	Making an argument, or responding to one, in such a way that it does not make it at all clear what your position is.
Appeal to Common Belief	When the claim that most or many people in general or of a particular group accept a belief as true is presented as evidence for the claim.
Appeal to Popularity	Using the popularity of a premise or proposition as evidence for its truthfulness.
Appeal to Common Sense	Asserting that your conclusion or facts are just "common sense" when, in fact, they are not.
Appeal to Common Folk	In place of evidence, attempting to establish a connection to the audience based on being a "regular person" just like each of them.

Appeal to Trust	The belief that if a source is considered trustworthy or untrustworthy, then any information from that source must be true or false, respectively.
Argument from Age	The misconception that previous generations had superior wisdom to modern man, thus conclusions that rely on this wisdom are seen accepted as true or more true than they actually are.
Appeal to Heaven	Asserting the conclusion must be accepted because it is the "will of God" or "the will of the gods".
Appeal to Tradition	Using historical preferences of the people (tradition), either in general or as specific as the historical preferences of a single individual, as evidence that the historical preference is correct.
Etymological Fallacy	The assumption that the present-day meaning of a word should be/is similar to the historical meaning.
Genetic Fallacy	Basing the truth claim of an argument on the origin of its claims or premises.
Appeal to Celebrity	Accepting a claim of a celebrity based on his or her celebrity status, not on the strength of the argument.
Appeal to Authority	Insisting that a claim is true simply because a valid authority or expert on the issue said it was true, without any other supporting evidence offered.
Appeal to False Authority	Using an alleged authority as evidence in your argument when the authority is not really an authority on the facts relevant to the argument.
Argument from False Authority	When a person making a claim is presented as an expert who should be trusted when his or her expertise is not in the area being discussed.
Blind Authority Fallacy	Asserting that a proposition is true solely on the authority making the claim.
Argument by Personal Charm	When an argument is made stronger by the personal characteristics of the person making the argument, often referred to as "charm".
Argument to the Purse	Concluding that the truth value of the argument is true or false based on the financial status of the author of the argument or the money value associated with the truth.
Ad Hominem Circumstantial	Suggesting that the person who is making the argument is biased or predisposed to take a particular stance, and therefore, the argument is necessarily invalid.
Gadarene Swine Fallacy	The assumption that because an individual is not in formation with the group, that the individual must be the one off course.
Ad Hominem Tu quoque	Claiming the argument is flawed by pointing out that the one making the argument is not acting consistently with the claims of the argument.
Bulverism	It is the assumption and assertion that an argument is flawed or false because of the arguer's characteristic associated with the arguer's identity.
Righteousness Fallacy	Assuming that just because a person's intentions are good, they have the truth or facts on their side.
Self Righteousness Fallacy	Assuming that just because your intentions are good, you have the truth or facts on your side.
Reductio ad Hitlerum	The attempt to make an argument analogous with Hitler or the Nazi party.
Ad Hominem Guilt by Association	When the source is viewed negatively because of its association with another person or group who is already viewed negatively.
Identity Fallacy	When one's argument is evaluated based on their physical or social identity when the strength of the argument is independent of identity.
Appeal to Stupidity	Attempting to get the audience to devalue reason and intellectual discourse, or devaluing reason and intellectual discourse based on the rhetoric of an arguer.

Ad Hominem Abusive	Attacking the person making the argument, rather than the argument itself, when the attack on the person is completely irrelevant to the argument the person is making.
Ad Fidentia	Attacking the person's self-confidence in place of the argument or the evidence.
appeal to loyalty	When one is either implicitly or explicitly encouraged to consider loyalty when evaluating the argument when the truth of the argument is independent of loyalty.
Appeal to Accomplishment	When the argument being made is sheltered from criticism based on the level of accomplishment of the one making the argument.
Scapegoating	Unfairly blaming an unpopular person or group of people for a problem or a person or group that is an easy target for such blame.
Fallacy of Opposition	Asserting that those who disagree with you must be wrong and not thinking straight, primarily based on the fact that they are the opposition.
Proof by Intimidation	Making an argument purposely difficult to understand in an attempt to intimidate your audience into accepting it, or accepting an argument without evidence or being intimidated to question the authority or a priori assumptions of the one making the argument.
Poisoning the Well	To prime the audience with adverse information about the opponent from the start, in an attempt to make your claim more acceptable or discount the credibility of your opponent's claim.
Wishful Thinking	When the desire for something to be true is used in place of/or as evidence for the truthfulness of the claim.
Appeal to Faith	It is the assertion that one must have (the right kind of) faith in order to understand the argument.
Notable Effort	Accepting good effort as a valid reason to accept the truth of the conclusion, even though the effort is unrelated to the truth.
Prejudicial Language	Loaded or emotive terms used to attach value or moral goodness to believing the proposition.
Special Pleading	Applying standards, principles, and/or rules to other people or circumstances, while making oneself or certain circumstances exempt from the same critical criteria, without providing adequate justification.
If By Whiskey	A response to a question that is contingent on the questioner's opinions and makes use of words with strong connotations.
Overextended Outrage	One or more statistically rare cases are implied to be the norm or the trend (without evidence) for the purpose of expressing or inciting outrage toward an entire group.
Appeal to Ridicule	Presenting the argument in such a way that makes the argument look ridiculous, usually by misrepresenting the argument or the use of exaggeration.
Argument by Emotive Language	Substituting facts and evidence with words that stir up emotion, with the attempt to manipulate others into accepting the truth of the argument.
Style Over Substance	When the arguer embellishes the argument with compelling language or rhetoric, and/or visual aesthetics.
Appeal to Anger	When the emotions of anger, hatred, or rage are substituted for evidence in an argument.
Appeal to Pity	The attempt to distract from the truth of the conclusion by the use of pity.
Appeal to Emotion	This is the general category of many fallacies that use emotion in place of reason in order to attempt to win the argument.
Appeal to Flattery	When an attempt is made to win support for an argument, not by the strength of the argument, but by using flattery on those whom you want to accept your argument.
Appeal to Spite	Substituting spite (petty ill will or hatred with the disposition to irritate, annoy, or thwart) for evidence in an argument, or as a reason to support or reject a claim.

	pragmatic fallacy	Claiming that something is true because the person making the claim has experienced, or is referring to someone who has experienced, some practical benefit from believing the thing to be true.
	Appeal to Force	When force, coercion, or even a threat of force is used in place of a reason in an attempt to justify a conclusion.
	Appeal to Fear	When fear, not based on evidence or reason, is being used as the primary motivator to get others to accept an idea, proposition, or conclusion.
Informal-Insufficiency	Fallacy of Composition	Inferring that something is true of the whole from the fact that it is true of some part of the whole.
	Fallacy of Division	Inferring that something is true of one or more of the parts from the fact that it is true of the whole.
	Stereotyping the fallacy	The general beliefs that we use to categorize people, objects, and events while assuming those beliefs are accurate generalizations of the whole group.
	Ecological Fallacy	The interpretation of statistical data where inferences about the nature of individuals are deduced from inference for the group to which those individuals belong.
	Oversimplified Cause Fallacy	When a contributing factor is assumed to be the cause, or when a complex array of causal factors is reduced to a single cause.
	Accident Fallacy	When an attempt is made to apply a general rule to all situations when clearly there are exceptions to the rule.
	mcnamara fallacy	When a decision is based solely on quantitative observations (i.e., metrics, hard data, statistics) and all qualitative factors are ignored.
	Overwhelming Exception	A generalization that is technically accurate, but has one or more qualifications which eliminate so many cases that the resulting argument is significantly weaker than the arguer implies.
	Reductio ad Absurdum	A mode of argumentation or a form of argument in which a proposition is disproven by following its implications logically to an absurd conclusion.
	Nirvana Fallacy	Comparing a realistic solution with an idealized one, and discounting or even dismissing the realistic solution.
	Relative Privation	Trying to make a scenario appear better or worse by comparing it to the best or worst case scenario.
	imposter fallacy	When one suggests or claims, with insufficient evidence, that the group outliers who are viewed as damaging to the group are primarily made up of infiltrators of another group with the purpose of making the infiltrated group look bad.
	Misleading Vividness	A small number of dramatic and vivid events are taken to outweigh a significant amount of statistical evidence.
	Appeal to Possibility	When a conclusion is assumed not because it is probably true, but because it is possible that it is true, no matter how improbable.
	Rights To Ought Fallacy	When one conflates a reason for one's rights (constitutional or other) with what one should do.
	Psychogenetic Fallacy	Inferring some psychological reason why an argument is made then assuming it is invalid, as a result.
	Weak Analogy	When an analogy is used to prove or disprove an argument, but the analogy is too dissimilar to be effective, that is, it is unlike the argument more than it is like the argument.
	Extended Analogy	Suggesting that because two things are alike in some way and one of those things is like something else, then both things must be like that "something else".
	Appeal to Equality	An assertion is deemed true or false based on an assumed pretense of equality, where what exactly is "equal" is not made clear, and not supported by the argument.
	False Equivalence	An argument or claim in which two completely opposing arguments appear to be logically equivalent when in fact they are not.

Galileo Fallacy	The claim that because an idea is forbidden, prosecuted, detested, or otherwise mocked, it must be true, or should be given more credibility.
Post Designation	Drawing a conclusion from correlations observed in a given sample, but only after the sample has already been drawn, and without declaring in advance what correlations the experimenter was expecting to find.
Just In Case Fallacy	Making an argument based on the worst-case scenario rather than the most probable scenario, allowing fear to prevail over reason.
Selective Attention	Focusing your attention on certain aspects of the argument while completely ignoring or missing other parts.
nutpicking fallacy	When someone presents an atypical or weak member of a group as if they are a typical or strong representative.
Biased Sample Fallacy	Drawing a conclusion about a population based on a sample that is biased, or chosen in order to make it appear the population on average is different than it actually is.
Survivorship Fallacy	In its general form, the survivorship fallacy is basing a conclusion on a limited number of "winner" testimonies due to the fact we cannot or do not hear the testimonies of the losers.
Spotlight Fallacy	Assuming that the media's coverage of a certain class or category is representative of the class or category in whole.
Hasty Generalization	Drawing a conclusion based on a small sample size, rather than looking at statistics that are much more in line with the typical or average situation.
Incomplete Comparison	An incomplete assertion that cannot possibly be refuted. This is popular in advertising.
Texas Sharpshooter Fallacy	Ignoring the difference while focusing on the similarities, thus coming to an inaccurate conclusion.
Faulty Comparison	Comparing one thing to another that is really not related, in order to make one thing look more or less desirable than it really is.
Base Rate Fallacy	Ignoring statistical information in favor of using irrelevant information, that one incorrectly believes to be relevant, to make a judgment.
Least Plausible Hypothesis	Choosing more unreasonable explanations for phenomena over more defensible ones.
Far Fetched Hypothesis	Offering a bizarre (far-fetched) hypothesis as the correct explanation without first ruling out more mundane explanations.
Cherry Picking	When only select evidence is presented in order to persuade the audience to accept a position, and evidence that would go against the position is withheld.
Argument by Selective Reading	When a series of arguments or claims is made and the opponent acts as if the weakest argument was the best one made.
deceptive sharing	Sharing an article, post, or meme on social media with the intent to influence public perception to perceive a statistically rare event as a common event.
Multiple Comparisons Fallacy	Claiming that unexpected trends that occur through random chance alone in a data set with a large number of variables are meaningful.
Magical Thinking	Making causal connections or correlations between two events not based on logic or evidence, but primarily based on superstition.
Slippery Slope	When a relatively insignificant first event is suggested to lead to a more significant event, which in turn leads to a more significant event, and so on, until some ultimate, significant event is reached, where the connection of each event is not only unwarranted but with each step it becomes more and more improbable.
Sunk Cost Fallacy	Reasoning that further investment is warranted on the fact that the resources already invested will be lost otherwise, not taking into consideration the overall losses involved in the further investment.
Jumping to Conclusions	Drawing a conclusion without taking the needed time to evaluate the evidence or reason through the argument.

	Argument from Silence	Drawing a conclusion based on the silence of the opponent, when the opponent is refusing to give evidence for any reason.
	Argument from Hearsay	Presenting the testimony of a source that is not an eyewitness to the event in question.
	Anonymous Authority	When an unspecified source is used as evidence for the claim.
	Insignificant Cause	An explanation that posits one minor factor, out of several that contributed, as its sole cause.
	Just Because Fallacy	Refusing to respond to give reasons or evidence for a claim by stating yourself as the ultimate authority on the matter.
	Appeal to the Law	When following the law is assumed to be the morally correct thing to do, without justification, or when breaking the law is assumed to be the morally wrong thing to do, without justification.
	Appeal to Normality	Using social norms to determine what is good or bad.
	False Effect	Claiming that the cause is true or false based on what we know about the effect in a claim of causality that has not been properly established.
	Appeal to Consequences	Concluding that an idea or proposition is true or false because the consequences of it being true or false are desirable or undesirable.
	Retrogressive Causation	Invoking the cause to eliminate the effect, or calling on the source to relieve the effect of the source.
	Confusing Currently Unexplained with Unexplainable	Making the assumption that what cannot currently be explained is, therefore, unexplainable (impossible to explain).
	Appeal to Desperation	Arguing that your conclusion, solution, or proposition is right based on the fact that something must be done, and your solution is "something."
	Regression Fallacy	Ascribing a cause where none exists in situations where natural fluctuations exist while failing to account for these natural fluctuations.
	Causal Reductionism	Assuming a single cause or reason when there were actually multiple causes or reasons.
	Questionable Cause	Concluding that one thing caused another, simply because they are regularly associated.
Informal-Inappropriate Presumption	Hedging	Refining your claim simply to avoid counter evidence and then acting as if your revised claim is the same as the original.
	Circular Definition	A circular definition is defining a term by using the term in the definition.
	Homunculus Fallacy	An argument that accounts for a phenomenon in terms of the very phenomenon that it is supposed to explain, which results in an infinite regress.
	Circular Reasoning	A type of reasoning in which the proposition is supported by the premises, which is supported by the proposition.
	Tokenism	Interpreting a token gesture as an adequate substitute for the real thing.
	Appeal to Novelty	Claiming that something that is new or modern is superior to the status quo, based exclusively on its newness.
	Two Wrongs Make a Right	When a person attempts to justify an action against another person because the other person did take or would take the same action against him or her.
	Appeal to Nature	When used as a fallacy, the belief or suggestion that “natural” is better than “unnatural” based on its naturalness.
	Naturalistic Fallacy	When the conclusion expresses what ought to be, based only on what is, or what ought not to be, based on what is not.
	Moralistic Fallacy	When the conclusion expresses what is, based only on what one believes ought to be, or what isn’t is based on what one believes ought not to be.
	Suppressed Correlative	The attempt to redefine a correlative (one of two mutually exclusive options) so that one alternative encompasses the other, i.e. making one alternative impossible.

Historians Fallacy	Judging a person's decision in the light of new information not available at the time.
Willed Ignorance	Refusing to change one's mind or consider conflicting information based on a desire to maintain one's existing beliefs.
Appeal to Coincidence	Concluding that a result is due to chance when the evidence strongly suggests otherwise.
Argument from Incredulity	Concluding that because you can't or refuse to believe something, it must not be true, improbable, or the argument must be flawed.
Argument by Pigheadedness	This is a refusal to accept a well-proven argument for one of many reasons related to stubbornness.
Argument by Repetition	Repeating an argument or a premise over and over again in place of better supporting evidence.
Definist Fallacy	Defining a term in such a way that makes one's position much easier to defend.
Limited Scope	The theory doesn't explain anything other than the phenomenon it explains (that one thing), and at best, is likely to be incomplete.
Moving the Goalposts	Demanding from an opponent that he or she address more and more points after the initial counter-argument has been satisfied refusing to concede or accept the opponent's argument.
Argument from Fallacy	Concluding that the truth value of an argument is false based on the fact that the argument contains a fallacy.
False Dilemma	When only two choices are presented yet more exist, or a spectrum of possible choices exists between two extremes.
Argument from Ignorance	The assumption of a conclusion or fact based primarily on lack of evidence to the contrary.
Alternative Advance	When one is presented with just two choices, both of which are essentially the same, just worded differently.
Shifting of the Burden of Proof	Making a claim that needs justification, then demanding that the opponent justifies the opposite of the claim.
Proving Non Existence	Demanding that one proves the non-existence of something in place of providing adequate evidence for the existence of that something.
Proof Surrogate	A claim masquerading as proof or evidence, when no such proof or evidence is actually being offered.
Rationalization	Offering false or inauthentic excuses for our claim because we know the real reasons are much less persuasive or more embarrassing to share, or harsher than the manufactured ones given.
Spin Doctoring	Presenting information in a deceptive way that results in others interpreting the information in such a way that does not reflect reality but is how you want the information to be interpreted.
Lying with Statistics	Presenting statistical data in a very biased way.
Ad Hoc Rescue	When we desperately want to be right and hold on to certain beliefs, we begin to make up excuses without no real evidence as to why our belief could be.
False Attribution	Appealing to an irrelevant, unqualified, unidentified, biased, or fabricated source in support of an argument (modern usage).
Amazing Familiarity	The argument contains information that seems impossible to have obtained—like it came from an omniscient author.
Ludic Fallacy	Assuming flawless statistical models apply to situations where they actually don't.
Missing Data Fallacy	Refusing to admit ignorance to the hypothesis and/or the conclusion, but insisting that your ignorance has to do with missing data that validate both the hypothesis and conclusion.
Begging the Question	Any form of argument where the conclusion is assumed in one of the premises.
Complex Question Fallacy	A question that has a presupposition built in, which implies something but protects the one asking the question from accusations of false claims.

	Package Deal Fallacy	Assuming things that are often grouped together must always be grouped together, or the assumption that the ungrouping will have significantly more severe effects than anticipated.
	Subjectivist Fallacy	Claiming something is true for one person, but not for someone else when, in fact, it is true for everyone (objective) as demonstrated by empirical evidence.
	Distinction Without a Difference	The assertion that a position is different from another position based on the language when, in fact, both positions are the same – at least in practice or practical terms.
	Hypothesis Contrary to Fact	Offering a poorly supported claim about what might have happened in the past or future, if (the hypothetical part) circumstances or conditions were different.
	Shoehorning	The process of force-fitting some current affair into one's personal, political, or religious agenda.
	Appeal to Self-evident Truth	Making the claim that something is "self-evident" when it is not self-evident in place of arguing a claim with reason.
	Subverted Support	The attempt to explain some phenomenon that does not actually occur or there is no evidence that it does.
	Double Standard	Judging two situations by different standards when, in fact, you should be using the same standard.
	Fantasy Projection	Confusing subjective experiences, usually very emotionally charged, with objective reality, then suggesting or demanding that others accept the subjective experience as objective reality.
	Argument to Moderation	Asserting that given any two positions, there exists a compromise between them that must be correct.
	Broken Window Fallacy	The illusion that destruction and money spent in recovery from destruction, is a net-benefit to society.
	Self Sealing Argument	An argument or position is self-sealing if and only if no evidence can be brought against it no matter what.
	Unfalsifiability	Confidently asserting that a theory or hypothesis is true or false even though the theory or hypothesis cannot possibly be contradicted by an observation or the outcome of any physical experiment, usually without strong evidence or good reasons.
	Conspiracy Theory	Explaining that your claim cannot be proven or verified because the truth is being hidden and/or evidence destroyed by a group of two or more people.
	Confusing an Explanation with an Excuse	Treating an explanation of a fact as if it were a justification of the fact, a valid reason for the fact, or evidence for the fact.
	Limited Depth	Failing to appeal to an underlying cause, and instead simply appealing to membership in a category.
	Alleged Certainty	Asserting a conclusion without evidence or premises, through a statement that makes the conclusion appear certain when, in fact, it is not.

Table 11: Categorizations and descriptions of the fallacies in our dataset.