Reason from Fallacy: Enhancing Large Language Models' Logical Reasoning through Logical Fallacy Understanding

Yanda Li[†], Dixuan Wang[†], Jiaqing Liang[†‡✉], Guochao Jiang[†], Qianyu He[†],
Yanghua Xiao[†‡], Deqing Yang[†‡✉]
[†]School of Data Science, Fudan University, Shanghai, China
[‡]Shanghai Key Laboratory of Data Science, Shanghai, China
[†]{ydli22, dxwang23, gcjiang22, qyhe21}@m.fudan.edu.cn
[‡]{liangjiaqing, shawyh, yangdeqing}@fudan.edu.cn

Abstract

Large Language Models (LLMs) have demonstrated good performance in many reasoning tasks, but they still struggle with some complicated reasoning tasks including logical reasoning. One non-negligible reason for LLMs' suboptimal performance on logical reasoning is their overlooking of understanding logical fallacies correctly. To evaluate LLMs' capability of logical fallacy understanding (LFU), we propose five concrete tasks from three cognitive dimensions of WHAT, WHY, and HOW in this paper. Towards these LFU tasks, we have successfully constructed a new dataset LFUD based on GPT-4 accompanied by a little human effort. Our extensive experiments justify that our LFUD can be used not only to evaluate LLMs' LFU capability, but also to fine-tune LLMs to obtain significantly enhanced performance on logical reasoning.

1 Introduction

As a cognitive process, *logical reasoning* plays an important role in many intellectual activities, such as problem solving, decision making and planning (Huang and Chang, 2022). Up to now, a lot of efforts have been dedicated to logical reasoning based on language models (Cresswell, 1973; Kowalski, 1974; Iwańska, 1993; Liu et al., 2020). More recently, the popularity of large language models (LLMs) such as ChatGPT (Ouyang et al., 2022) and GPT-4 (OpenAI, 2023) stimulates the growth of research on LLM-based logical reasoning. Compared to traditional small language models, LLMs have demonstrated better performance in many reasoning tasks.

However, LLMs still struggle with some more complex reasoning tasks including logical reasoning. One non-negligible reason for LLMs' suboptimal performance on logical reasoning is their overlooking of understanding *logical fallacies* correctly. As early as 350 BC, Aristotle first proposed the concept of logical fallacy in his work *Sophistical Refutations* (Aristotle, 2006). Since then, logical fallacies have gradually become an important issue that should be noticed in our lives. "*Thou shalt not commit logical fallacies!*" has even become a worldwide popular idiom to remind us not to commit logical fallacies. By definition, logical fallacies refer to the errors in reasoning (Tindale, 2007), and they usually happen when the premises are not relevant or sufficient to draw the conclusions. Many previous works (Liu et al., 2020; Yu et al., 2020; Joshi et al., 2020; Han et al., 2022) have focused on evaluating LLM logical reasoning capabilities from the perspective of deductive reasoning, natural language inference, reading comprehension, etc. However, few works focus on logical fallacies, which is in fact the major reason causing logical inconsistency in the sentences.

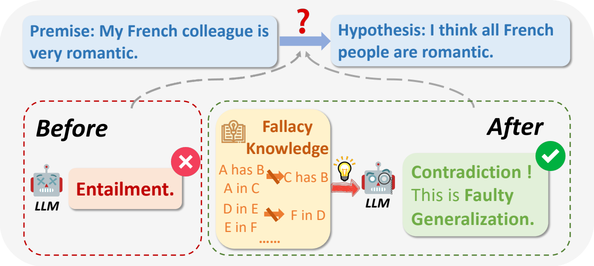
Premise: My French colleague is very romantic.
?
Hypothesis: I think all French people are romantic.
Before
LLM
Entailment.
Fallacy Knowledge
A has B
A in C
C has B
D in E
E in F
F in D
......
LLM
After
Contradiction ! This is Faulty Generalization.

Figure 1: LLMs have deficiencies in logical reasoning. Once they understand logical fallacies, they know how to avoid logical fallacies, and thus improve their performance in various logical reasoning tasks.

Chen et al. have observed that, LLMs often commit logical fallacies in logical reasoning, such as *"Either protect the environment or develop the economy."* (false dilemma) and *"Some roses are not red because not all roses are red."* (circular reasoning). It has been found that language models could avoid mistakes only when they understand what mistakes are (Chen et al., 2023a; An et al., 2023), which justifies the ancient Greek philosopher Epi-

Findings of the Association for Computational Linguistics: NAACL 2024, pages 3053–3066
June 16-21, 2024 ©2024 Association for Computational Linguistics

curus's saying "*The mistake is the first step to save yourself.*" Based on our empirical studies, we have also found that the logical reasoning capability of LLMs is closely related to their understanding of logical fallacies.

The previous studies related to logical fallacy (Jin et al., 2022; Sourati et al., 2023; bench authors, 2023) only focus on logical fallacy detection, i.e., the identification and classification of logical fallacies, rather than systematically evaluating LLMs' capability of logical fallacy understanding (LFU), not to mention improving LLMs' LFU capability. Moreover, they have not explored the relationships between LFU and logical reasoning, which is crucial to improve LLMs' capability of logical reasoning through enhancing their LFU capability. To address this problem, we focus on evaluating and enhancing LLMs' LFU capability in this paper, so as to enhance their capability of logical reasoning.

Nonetheless, our work has to face several challenges as follows. First, we need to formalize the concrete tasks for LFU, since no previous studies focus on this problem. Second, we need a new dataset specific to LFU, as the previous datasets of logical fallacies (Jin et al., 2022) only contain the logical fallacy types presenting in the sentences. To this end, we should propose a framework of constructing the LFU dataset towards the concrete LFU tasks, and then truthfully evaluating LLMs' LFU capability with the dataset.

To overcome these challenges, we primarily focus on constructing a dataset for LFU in this paper, of which the samples are generated to evaluate models' achievement on the following five LFU tasks corresponding to three cognitive dimensions of **WHAT**, **WHY**, and **HOW** (Swanborn, 2010).

1. **WHAT**-Identification (Task 1) and Classification (Task 2): identifying whether the given sentence contains a logical fallacy and which type of logical fallacy it is.
2. **WHY**-Deduction (Task 3) and Backward Deduction (Task 4): capturing the reasons causing the logical fallacy in the sentence.
3. **HOW**-Modification (Task 5): correcting the logical fallacy in the sentence.

Our proposed LFU tasks simulate the human understanding process of logical fallacies. Towards these tasks, we design a pipeline framework to automatically generate and synthesize a high-quality dataset, namely **Logical Fallacy Understanding Dataset (LFUD)**, based on GPT-4 accompanied by a little human effort. Specifically, we first collect some sentences as the *propositions* (statements) which are the basic logic units and used to generate the sentences containing logical fallacies. Then, with the help of GPT-4, we generate sentences based on the propositions with twelve typical logical fallacy types (Jin et al., 2022). And for each LFU task we propose, the instances of each fallacy type are synthesized. Then, we use our LFUD to evaluate the LFU capability of some representative LLMs. For the ultimate objective of our work, i.e., enhancing LLMs' capability of logical reasoning, we further fine-tune these LLMs with the instances in LFUD. Our extensive experiments reveal that fine-tuning LLMs with LFUD can significantly enhance their logical reasoning capability.

In summary, our main contributions in this paper include:

1. Inspired by the three cognitive dimensions of **WHAT**, **WHY**, and **HOW**, we propose five concrete tasks which can truthfully evaluate LLMs' performance on LFU.

2. Towards our proposed five LFU tasks, we devise a new framework for constructing a high-quality dataset, namely LFUD, to evaluate LLMs' LFU capability, so as to enhance LLMs' performance on logical reasoning.

3. The LFUD we constructed includes 4,020 instances involving 12 logical fallacy types. Our extensive experiments have demonstrated that our LFUD can not only evaluate LLMs' LFU capability, but also improve LLMs' capability of logical reasoning through fine-tuning LLMs with LFUD samples in terms of the LFU tasks.

2 Related Work

Logical Reasoning Up to now, a lot of efforts have been dedicated to logical reasoning based on language models (Cresswell, 1973; Kowalski, 1974; Iwańska, 1993; Liu et al., 2020). In particular, how to evaluate the models' logical reasoning capability has attracted increasing attention, including deductive reasoning (Ontanon et al., 2022; Han et al., 2022), natural language inference (NLI) (Yanaka et al., 2019; Joshi et al., 2020; Liu et al., 2021) and multi-choice reading comprehension (MRC) (Liu et al., 2020; Yu et al., 2020; Wang et al., 2022). Recently, the power of LLMs has stimulated the research on logical reasoning with LLMs, including LLMs evaluation (Yu et al.,

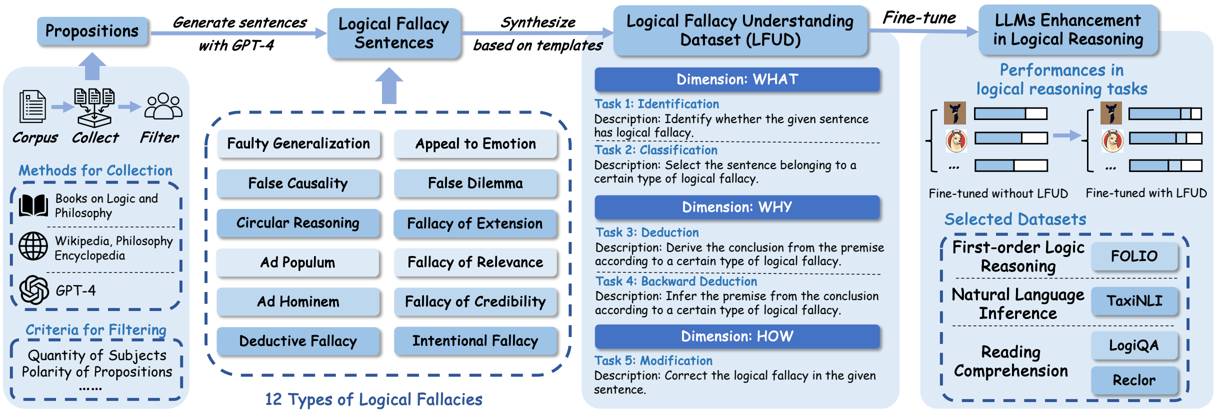
Propositions
Generate sentences with GPT-4
Logical Fallacy Sentences
Synthesize based on templates
Logical Fallacy Understanding Dataset (LFUD)
Fine-tune
LLMs Enhancement in Logical Reasoning
Corpus
Collect
Filter
Methods for Collection
Books on Logic and Philosophy
Wikipedia, Philosophy Encyclopedia
GPT-4
Criteria for Filtering
Quantity of Subjects
Polarity of Propositions
......
Faulty Generalization
Appeal to Emotion
False Causality
False Dilemma
Circular Reasoning
Fallacy of Extension
Ad Populum
Fallacy of Relevance
Ad Hominem
Fallacy of Credibility
Deductive Fallacy
Intentional Fallacy
12 Types of Logical Fallacies
Dimension: WHAT
Task 1: Identification
Description: Identify whether the given sentence has logical fallacy.
Task 2: Classification
Description: Select the sentence belonging to a certain type of logical fallacy.
Dimension: WHY
Task 3: Deduction
Description: Derive the conclusion from the premise according to a certain type of logical fallacy.
Task 4: Backward Deduction
Description: Infer the premise from the conclusion according to a certain type of logical fallacy.
Dimension: HOW
Task 5: Modification
Description: Correct the logical fallacy in the given sentence.
Performances in logical reasoning tasks
...
...
Fine-tuned without LFUD
Fine-tuned with LFUD
Selected Datasets
First-order Logic Reasoning
FOLIO
Natural Language Inference
TaxiNLI
Reading Comprehension
LogiQA
Reclor

Figure 2: Our framework of constructing LFUD and fine-tuning LLMs with LFUD to enhance logical reasoning. At first, we collected some propositions, based on which the sentences with the logical fallacies of 12 types were generated by GPT-4. Then, for the five LFU tasks we proposed, the QA instances were synthesized based on the previous generated sentences. Finally, we fine-tuned LLMs with LFUD, revealing that fine-tuning LLMs with LFUD can significantly enhance their logical reasoning capability.

2023; Blair-Stanek et al., 2023; Teng et al., 2023), and LLMs enhancement (Zhang et al., 2023; Chen et al., 2023b). Despite these works' achievements, enhancing LLMs' logical reasoning capability remains a non-negligible challenge. The major reason for logical inconsistencies in many sentences is the misunderstanding of logical fallacies, which is still under-explored in the research field of logic.

Logical Fallacy Logical fallacy is the main reason for the logical inconsistencies presenting in our life. As early as 350 BC, Aristotle first proposed the concept of logical fallacy in his work *Sophistical Refutations* (Aristotle, 2006). Since then, logical fallacies have gradually gained attention in human society. In recent years, the studies related to logical fallacies mainly focused on dataset construction (Habernal et al., 2018; Martino et al., 2020; Jin et al., 2022) and fallacy classification (Stab and Gurevych, 2017; Goffredo et al., 2022; Jin et al., 2022; Payandeh et al., 2023). For instance, Jin et al. first proposed the task of Logical Fallacy Detection, presenting a framework of 13 logical fallacy types, and evaluated all sentence samples on a classification task. Sourati et al. proposed a Case-Based Reasoning method that classifies new cases of logical fallacy by language-modeling-driven retrieval and the adaptation of historical cases. However, there is no work to systematically evaluate LLMs' capability of logical fallacy understanding (LFU). For the first time, our work in this paper proposes a new dataset specific to LFU represented by five concrete tasks corresponding to three cognitive dimensions of **WHAT**, **WHY**, and **HOW**.

Learning from Synthetic Data Synthesizing data for model training has gradually gained popularity along with the advancements of language models. This approach is particularly beneficial for tasks that are difficult to be constructed or those with scarce data resources (Møller et al., 2023). Currently, synthetic data has been applied in various tasks such as relation extraction (Papanikolaou and Pierleoni, 2020), text classification (Chung et al., 2023), irony detection (Abaskohi et al., 2022), translation (Sennrich et al., 2015), and sentiment analysis (Maqsud, 2015). For example, Josifoski et al. proposed a strategy to design an effective synthetic data generation pipeline and applied it to closed information extraction. In addition, Li et al. conducted a series of experiments to evaluate the effectiveness of LLMs in generating synthetic data to support model training for different text classification tasks. Beyond these fundamental tasks, Eldan and Li proposed to use LLMs with synthetic data to generate short stories typically for 3 to 4-year-old only containing words. But they did not focus on logical fallacy. We are the first to focus on the data augmentation strategies in LFU.

3 Methodology of Dataset Construction

In this section, we present the pipeline of constructing our LFUD, of which the overall framework is depicted in Figure 2. Starting from the propositions, we detail the steps of synthesizing the samples towards five LFU tasks and the twelve representative

Statistic	Number
Singular Proposition	54
Particular Proposition	5
Universal Proposition	8
Affirmative Proposition	56
Negative Proposition	11
Propositions without Pronouns	58
Propositions with Pronouns	9
Propositions with Human Subjects	52
Propositions with Non-Human Subjects	15

Table 1: Some statistics of the 67 propositions in LFUD.

logical fallacy types.

3.1 Acquiring Propositions

At the first step of constructing our LFUD, we collected some propositions which were subsequently used for generating the sentences presenting various logical fallacies. According to Hurley (2000), a proposition is one sentence that is either true or false. We considered several sources of proposition collection, including some authoritative books of logic and philosophy (Hurley, 2000; Hausman, 2012), open websites such as Wikipedia and Stanford Encyclopedia of Philosophy. In addition, LLMs can be utilized to generate some propositions for enriching proposition diversity. To seek the satisfactory LLM for generating propositions, we tested some representative LLMs' identification performance on 200 instances from the Big-Bench (bench authors, 2023), consisting of correct and incorrect (logical fallacy) sentences. The results showed that GPT-4 can correctly identify in over 90% of the sentences whether they have logical fallacies, despite the limited capability in directly generating complex tasks. Thus, we leveraged GPT-4 to generate more propositions and subsequent sentences presenting logical fallacies.

The considerable propositions should be simple and intuitive, but diverse. Finally, we filtered out 67 propositions and the relevant statistics are listed in Table 1. The following sentences are the proposition examples:

1. Everyone in my family has never been to Europe.
2. X accepted Y's suggestion.
3. Michael had dinner at an Italian restaurant.

3.2 Generating Sentences with GPT-4

Given GPT-4's capability of natural language generation and logical fallacy identification, we directly used GPT-4 to generate the sentences presenting

/ Generation Instruction */*

As a logician, when presented with a proposition, your objective is to simulate the way of human thinking, generating a sentence with specific type of logical fallacy. The generation should follow these instructions:

1. Generate the sentence with **Faulty Generalization**. Faulty Generalization occurs when ... (Detailed description)
2. The sentence should have complete premise and conclusion, but try not to make it too long.

/ Three demonstration examples */*

Proposition 1: Neither of the classes I took at UF were interesting.

Result 1: A college is not a good college if none of its classes are interesting. Neither of the classes I took at UF were interesting, so UF is not a good college.

. . .

/ Input the proposition */*

Proposition: Peter visited China last year.

/ GPT-4's output */*

Result: Peter visited China last year. Peter is a European. Therefore, all Europeans have been to China.

Table 2: A prompt case for GPT-4 to generate a sentence with the given logical fallacy type.

various logical fallacies in this step. To take into account the logical fallacies existing in our life as many as possible, we refered to the thirteen typical types of logical fallacies (as listed in Table 7 and Appendix B) proposed by Jin et al. (2022).

Given a proposition and a certain logical fallacy type, we asked GPT-4 to generate a sentence of this logical fallacy type with a prompt, which contains the generation instruction and a demonstration example of the given logical fallacy type. Table 2 illustrates the prompt for GPT-4 about the type of Faulty Generalization. Specifically, due to the rather vague definition of Equivocation provided by Jin et al. (2022), and the scarcity of such fallacy instances in real life, GPT-4 can hardly understand Equivocation and generate corresponding sentences correctly. To ensure the quality of the sentences generated by GPT-4 , we neglected Equivocation fallacy type and generated the sentences for the rest twelve logical fallacy types.

To ensure that the generated sentences meet the requirements, we further manually proofread the sentences with logical fallacies generated by GPT-4. Each generated sentence was proofread with two main areas of concern: structural integrity and validity of fallacies, as described in Appendix C, to ensure that the sentences made sense and met the requirements of specific fallacy type. For each of the 67 propositions, we generated 12 sentences

Dimension	Task name	Task definition
WHAT	Task1: Identification Task2: Classification	Identify whether the given sentence has logical fallacy. Select the sentence belonging to a certain type of logical fallacy.
WHY	Task3: Deduction Task4: Backward Deduction	Derive the conclusion from the premise according to a certain type of logical fallacy. Infer the premise from the conclusion according to a certain type of logical fallacy.
HOW	Task5: Modification	Correct the logical fallacy in the given sentence.

Table 3: Five LFU tasks corresponding to three cognitive dimensions.

with GPT-4, each of which presents one logical fallacy type. Thus, we generated 804 sentences with logical fallacies in total. These sentences are used to synthesize the samples for concrete LFU tasks as follows.

3.3 Proposing LFU Tasks and Synthesizing Task Instances

To evaluate LLMs' capability of LFU, we need to design concrete evaluation tasks. According to the principles of cognitive science (Swanborn, 2010), humans generally understand objects from three dimensions: **WHAT** it is, **WHY** it is, and **HOW** it operates, which are interconnected and progressive cognition levels. Inspired by these dimensions, we propose five concrete tasks which are used to verify models' capability of LFU. Table 3 lists the definitions of the five tasks. Wherein, Task 1 and Task 2 belong to WHAT dimension, which identify whether the given sentence has the logical fallacy (of a certain type). Task 3 and Task 4 belong to WHY dimension, which verify whether the model captures the reason causing the logical fallacy in the sentence. The last Task 5 belongs to HOW dimension, which requires correcting the logical fallacy of the given type in the sentence. Specifically, we synthesized multiple-choice questions for the first four tasks, and sentence generation questions for Task 5. We further provided one toy example for each task in Appendix A.

In fact, previous studies (Jin et al., 2022; bench authors, 2023) have focused on the two tasks of WHAT dimension, i.e., understanding what the logical fallacy in the sentence is. To the best of our knowledge, there are no studies concerning the tasks of WHY and HOW dimensions by now. But notably, the ultimate goal of LFU is to avoid logical fallacies, which requires us to understand the reasons causing logical fallacies and correct logical fallacies. Therefore, we paid more attention to the tasks of WHY and HOW dimension in this paper.

For each sentence with one of the twelve logical fallacy types generated in the previous step, we synthesized one QA instance for every LFU task with the question templates. For each LFU task, the question stems (without question options) of all instances are generated according to some templates, as shown in Appendix A. Particularly, for Task 3 and Task 4, we need to identify the premise and conclusion for the given sentence, and further provide question options. Thus, we directly asked GPT-4 to generate the results as we needed.

To minimize the impact of instruction design when asking LLMs to achieve these tasks, we first designed some candidate question templates to constitute a template pool in fact, and then randomly chose one template from the pool to generate the question for a certain LFU task. In addtion, we also shuffled the orders of question options. Finally, our LFUD contains 4,020 (QA) instances in total, involving 5 LFU tasks and 12 logical fallacy types, which stem from the 67 propositions and 804 sentences with logical fallacies.[1]

4 Evaluation

4.1 Experiment Setup

Datasets To evaluate LLMs' performance on logical reasoning, we used four representative datasets including FOLIO (Han et al., 2022), TaxiNLI (Joshi et al., 2020), LogiQA (Liu et al., 2020), and Reclor (Yu et al., 2020) in our experiments.

FOLIO focuses on first-order logic reasoning (FOL) that is a classical deductive reasoning task. TaxiNLI is specific to natural language inference (NLI) that tests the logical relationship between a premise and a hypothesis. LogiQA and Reclor are the multi-choice reading comprehension (MRC) datasets, which choose the most suitable answer corresponding to the given text, could better reflect comprehensive logical reasoning abilities. The instances of the four datasets are shown in Appendix D. In addition to the training data in above four datasets and our LFUD, we also used the logical

[1] LFUD is provided at https://github.com/YandaGo/LFUD

Datasets	FT Data	LLaMA2-13B		LLaMA2-7B		Vicuna-13B		Vicuna-7B		Orca2-7B	
		Acc.	Δ%	Acc.	Δ%	Acc.	Δ%	Acc.	Δ%	Acc.	Δ%
LogiQA2.0	Origin	45.55	-	42.30	-	52.74	-	47.71	-	54.39	-
	Origin+LOGIC	44.66	-1.95	35.62	-15.79	53.37	1.19	45.10	-5.47	52.93	-2.68
	Origin+LFUD	**47.90**	5.16	**43.13**	1.96	**55.85**	5.90	**47.84**	0.27	**56.55**	3.97
Reclor	Origin	47.20	-	40.40	-	54.40	-	49.20	-	55.80	-
	Origin+LOGIC	46.20	-2.12	42.20	4.46	54.00	-0.74	47.80	-2.85	55.80	0.00
	Origin+LFUD	**50.20**	6.36	**46.40**	14.85	**57.00**	4.78	**51.80**	5.28	**58.20**	4.30
TaxiNLI	Origin	68.54	-	62.68	-	78.91	-	77.47	-	82.33	-
	Origin+LOGIC	40.60	-40.76	58.80	-6.19	77.92	-1.25	76.18	-1.67	82.18	-0.18
	Origin+LFUD	**73.70**	7.53	**67.26**	7.31	**79.76**	1.08	**77.77**	0.39	**84.02**	2.05
FOLIO	Origin	61.76	-	50.98	-	36.76	-	50.49	-	72.55	-
	Origin+LOGIC	62.25	0.79	52.45	2.88	36.28	-1.31	45.10	-10.68	73.53	1.35
	Origin+LFUD	**66.18**	7.16	**59.31**	16.34	**44.61**	21.35	**56.37**	11.65	**76.47**	5.40

Table 4: LLMs' accuracy(%) on the four logical reasoning tasks (datasets) after being fine-tuned with different data. Origin represents fine-tuning the LLMs with the original training data in the logical reasoning datasets. Δ% is accuracy improvement relative to Origin. The best accuracy scores are **bolded** and the second best scores are underlined.

fallacy data LOGIC (Jin et al., 2022) to fine-tune LLMs. LOGIC (including LOGIC-CLIMATE) contains thirteen types of logical fallacy sentences, as shown in Appendix B.

LLMs We selected five popular LLMs in our experiments, including LLaMA2-7B, LLaMA2-13B (Touvron et al., 2023), Vicuna-7B, Vicuna-13B (Chiang et al., 2023) and Orca2-7B (Mitra et al., 2023). When fine-tuning these LLMs, we set the learning rate to 2.5e-5 and the batch size to 8. To ensure the robustness of our results, we repeated all experiments for three times and reported the average performance (accuracy) scores.

Dataset Split For the 4,020 synthesized instances in our LFUD, we randomly selected 3,000 instances (corresponding to 600 sentences with logical fallacies) as the training set and the remaining 1,020 instances (corresponding to 204 sentences with logical fallacies) as the test set. Given the instances of Task 1–4 (choice questions) have fixed answers, we only used the training samples (2,500 instances) of Task 1–4 to fine-tune the five LLMs. And we directly used some test samples of Task 5 to evaluate LLMs' cross-task learning capability on LFU, as presented in Subsection 4.3. To balance the labels of logical right and fallacy in Task 1 instances, we appended 500 logically correct sentences of Big-Bench (bench authors, 2023), and thus collected 2,900 training samples in our LFUD in total.

4.2 Effectiveness on Enhancing LLMs' Logical Reasoning

4.2.1 Overall Performance

To justify the value of our LFUD instances on enhancing LLMs' logical reasoning capability, we merged LFUD training samples with the original training samples in the four logical reasoning datasets, denoted by Origin, to fine-tune LLMs. We compared such a fine-tuning method with the method of fine-tuning LLMs only with Origin. In addition, we also compared the method of fine-tuning LLMs with Origin and some samples in LOGIC (Jin et al., 2022), which have the same number as the training samples in LFUD.

Table 4 lists the accuracy(%) scores of all five LLMs on the four logical reasoning tasks (datasets) which were fine-tuned with Origin, Origin+LOGIC and Origin+LFUD, respectively. And the performance improvements of Origin+LOGIC and Origin+LFUD relative to Orign are also listed. Based on the results in this table, we have the following observations and analysis.

1. Appending the training samples in our LFUD to Origin when fine-tuning LLMs significantly enhances their performance on all logical reasoning tasks. It shows that learning the LFU tasks we proposed is indeed helpful to improve LLMs' capability of various logical reasoning.

2. Although the samples in LOGIC are also the sentences with various logical fallacies, Origin+LOGIC cannot obtain the significant performance improvements of logical reasoning. Even

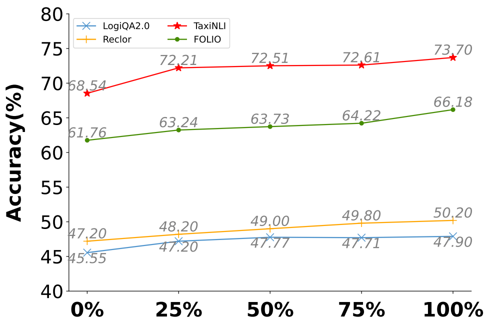

Figure 3: LLaMA2-13B's performance on the four logical reasoning tasks with different scales of LFUD training samples.

Models	WHAT		WHY		HOW
	Task1	Task2	Task3	Task4	Task5
LLaMA2-7B	46.84	47.86	24.75	29.35	40.00
LLaMA2-13B	55.76	53.61	52.36	54.48	50.00
Vicuna-7B	32.96	57.71	60.70	56.59	46.00
Vicuna-13B	50.76	58.46	58.33	61.32	56.00
ChatGPT	54.66	73.88	62.94	70.65	60.00
GPT-4	**86.35**	**86.19**	**78.61**	**85.70**	**88.00**

Table 5: Six representative LLMs' performance on our proposed five LFU tasks. To evaluate Task 5, we manually assessed LLMs' outputs for 50 randomly selected samples.

Task Category	LogiQA2.0	Reclor	TaxiNLI	FOLIO
No Tasks	45.55	47.20	68.54	61.76
w/o Task1	46.69	49.80	69.53	63.73
w/o Task2	45.74	48.00	69.88	65.20
w/o Task3	47.46	49.00	72.01	64.22
w/o Task4	46.44	48.80	69.28	65.20
All Tasks	**47.90**	**50.20**	**73.70**	**66.18**

Table 6: LLaMA2-13B's performance on the four logical reasoning tasks when excluding different LFU task's training instances.

Fallacy Type	LogiQA2.0	Reclor	TaxiNLI	FOLIO
No Fallacy Data	45.55	47.20	68.54	61.76
w/o Faulty Generalization	46.56	49.80	71.91	64.71
w/o False Causality	46.69	47.60	72.56	62.75
w/o Circular Reasoning	46.12	49.80	72.95	64.22
w/o Ad Populum	46.25	47.60	72.85	64.22
w/o Ad hominem	46.95	48.60	69.53	65.20
w/o Deductive Fallacy	45.87	49.40	73.78	62.75
w/o Appeal to Emotion	47.65	49.80	69.93	63.73
w/o False Dilemma	46.12	50.00	73.10	63.24
w/o Fallacy of Extension	45.93	49.40	72.51	64.71
w/o Fallacy of Relevance	47.65	50.20	70.92	61.27
w/o Fallacy of Credibility	47.58	48.60	72.06	62.75
w/o Intentional Fallacy	46.88	49.80	69.48	65.69
All Fallacy Types	**47.90**	**50.20**	**73.70**	**66.18**

Table 7: LLaMA2-13B's performance on the four logical reasoning tasks when excluding different logical fallacy type's training instances.

worse, it degrades LLMs' logical reasoning performance, compared with Origin in some cases. It implies that, unlike our LFU tasks from WHAT, WHY and HOW, only identifying the logical fallacy presented by the sentences in LOGIC cannot result in LLMs' really capability of LFU. In addition, The samples in LOGIC are raw and unclean, with some examples consisting of even fallacy questions and fallacy definitions.

4.2.2 Impacts of Different Factors in LFUD

To further validate LFUD's effectiveness on enhancing LLMs' logical reasoning capability, we also investigated the impacts of different factors in LFUD, including the scale of training data, LFU tasks and logical fallacy types. Due to space limitation, we only display the results of LLaMA2-13B.

Training Data Scale To verify the impacts of training data scale, we respectively extracted 25%, 50%, and 75% of the LFUD training data accompanied with Origin to fine-tune LLaMA2-13B, and then tested its performance on the four logical reasoning tasks. From Figure 3 we can see that, LLaMA2-13B's performance improvement becomes more apparent as the training data scale increases, showing that even only a small part of LFUD samples is also valuable.

LFU Task We fine-tuned LLaMA2-13B again with the training data excluding the instances of Task 1, Task 2, Task 3 and Task4, respectively. As shown in Table 6, excluding any task's instances would lead to the performance decline of LLaMA2-13B.

Logical Fallacy Type Similarly, we respectively excluded the instances of each logical fallacy type from LFUD training data, and then tested LLaMA2-13B's performance. The results in Table 7 indicate that every logical fallacy type contributes positively to LLM's logical reasoning capability.

4.3 LFU Performance of LLMs

Next, we validate LLMs' capability of LFU through evaluating their performance on the LFU tasks. We want to investigate LLMs' inherent capability on LFU, thus we directly used all instances of each LFU task in LFUD as the test samples without fine-tuning them with the training data.

Performance on Each LFU Task Besides the previous four LLMs, we additionally considered ChatGPT (Ouyang et al., 2022) and the latest GPT-4 (OpenAI, 2023) (using OpenAI API with tem-

Models	Accuracy(%)
LLaMA-2-7B	0.92 (6/654)
LLaMA-2-13B	1.99 (13/654)
Vicuna-7B	7.95 (52/654)
Vicuna-13B	26.61 (174/654)
ChatGPT	37.92 (248/654)
GPT-4	95.57 (625/654)

Table 8: LLMs' Performance on identifying 654 logically correct sentences of Task 1.

Fallacy Type	Task 1	Task 2	Task 3	Task 4
Faulty Generalization	76.12	**89.55**	59.70	50.75
False Causality	61.19	70.15	67.16	65.67
Circular Reasoning	34.33	52.24	55.22	62.69
Ad Populum	65.67	80.60	**79.10**	**79.10**
Ad hominem	**77.61**	**89.55**	59.70	59.70
Deductive Fallacy	40.30	49.25	62.69	77.61
Appeal to Emotion	16.42	77.61	64.18	77.61
False Dilemma	29.85	44.78	62.69	50.75
Fallacy of Extension	53.73	25.37	41.79	38.81
Fallacy of Relevance	25.37	37.31	11.94	44.78
Fallacy of Credibility	40.30	53.73	61.19	64.18
Intentional Fallacy	68.66	31.34	47.76	64.18

Table 9: Vicuna-13B's performance on Task 1–4 specific to each type of logical fallacies. The best accuracy scores are **bolded** and the second best scores are underlined.

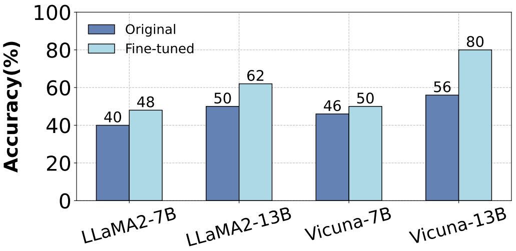

Figure 4: LLMs' Performance on Task 5 without fine-tuning (denoted as Original) or after being fine-tuned with training data of Task 1–4.

perature 0.7) in LFU performance evaluation. To balance the labels of Task 1, we added all 654 correct sentences in Big-Bench into Task 1's test data. Thus, we have a total of 1,458 instances for Task 1's evaluation. In addition, as Task 5 is to generate a new sentence rather than a fixed answer, we randomly selected 50 samples from its instances and manually assessed LLMs' outputs. All tested LLMs' performance is listed in Table 5, showing that different LLMs' performance varies significantly on the five LFU tasks. Among the LLMs, GPT-4 has much better performance than others on all tasks, justifying its strong capability of LFU. By contrast, LLaMA2-7B has the worst performance that is even worse than random selection.

Identifying Logical Correctness To further investigate whether LLMs really understand logical fallacies, we also asked LLMs to achieve Task 1 for the 654 sentences from Big-Bench that are logically correct (without logical fallacies). Their accuracy scores are listed in Table 8, showing that only GPT-4 has the satisfactory performance for this task. In fact, the rest LLMs tended to recognize the sentences as having logical fallacies for catering to Task 1's question. In addition, we also found these LLMs except for GPT-4 are easily influenced by the order of question options when achieving Task 1–4, indicating that they cannot well understand logical fallacies.

In Terms of Logical Fallacy Type Besides, we evaluated LLMs' LFU performance on Task 1–4 in terms of a specific logical fallacy type. The results listed in Table 9 show that, LLMs exhibited better performance on the tasks of *Faulty Generalization*, *False Causality*, *Ad populum* and *Ad Hominem*. These four types of logical fallacies are more distinctive and more frequently present in our life, resulting in that LLMs have encountered more sentences with these logical fallacy types during their pre-training.

Cross-task Learning Performance Compared with Task 1–4, Task 5 belongs to the higher cognition dimension **HOW**, and is more difficult for LLMs since it requires to generate a sentence satisfying the demand. An interesting research question is that, whether LLMs can well achieve Task 5 after learning the previous four tasks? To answer this question, for each of Task 1–4 we sampled 60 instances from its training data, and mixed them with the equal amount (240) of general conversation instances from lmsys-chat-1m to fine-tune LLMs, which was used to guarantee LLMs' generative ability. Then, we evaluated the fine-tuned LLMs' performance on Task 5, of which the performance is depicted in Figure 4. As well, LLMs' performance without fine-tuning, denoted as Original, is also displayed in the figure. The results indicate that, all the tested LLMs indeed enhanced their LFU performance through learning Task 1–4, also justifying their good cross-task learning capability of LFU tasks.

5 Conclusion

To evaluate LLMs' LFU performance, we propose five concrete tasks from three cognition dimensions WHAT, WHY, and HOW. Towards these tasks, we constructed a high quality dataset LFUD, which has been proven helpful by our extensive experiments to enhance LLMs' capability of logical reasoning. We hope our work in this paper is instructive and our LFUD becomes a valuable resource for further research on LFU.

Limitations

Although we argue that enhancing LLMs' logical reasoning capability through enabling LLMs to understand logical fallacies is language-independent, we should still acknowledge that the data and experiments of our work were only in English. As we know, LLMs might have different performance on many tasks including logical reasoning, across different languages. Therefore, the effectiveness of our solution proposed in this paper may vary when applied to other languages.

Ethical Considerations

At first, all authors of this work abide by the provided Code of Ethics. The quality of manual proofreading for logical fallacy sentences is ensured through a double-check strategy outlined in Appendix C. We ensure that the privacy rights of all members for proofreading are respected in the process. Besides, synthetic data generated by LLMs may involve potential ethical risks regarding fairness and bias (Bommasani et al., 2021; Blodgett et al., 2020), which results in further consideration when they are employed in downstream tasks. Although our dataset LFUD was built for better understanding logical fallacies, which is not intended for safety-critical applications, we still asked our members for proofreading to refine the offensive and harmful data generated by GPT-4. Despite these considerations, there may still be some unsatisfactory data that goes unnoticed in our final dataset.

Acknowledgements

This work was supported by Chinese NSF Major Research Plan (No. 92270121), Youth Fund (No. 62102095), Shanghai Science and Technology Innovation Action Plan (No. 21511100401). The computations in this research were performed using the CFFF platform of Fudan University.

References

Amirhossein Abaskohi, Arash Rasouli, Tanin Zeraati, and Behnam Bahrak. 2022. UTNLP at SemEval-2022 task 6: A comparative analysis of sarcasm detection using generative-based and mutation-based data augmentation. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 962–969, Seattle, United States. Association for Computational Linguistics.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2023. Learning from mistakes makes llm better reasoner. *arXiv preprint arXiv:2310.20689*.

Aristotle. 2006. *On sophistical refutations*. ReadHowYouWant. com.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can gpt-3 perform statutory reasoning? *arXiv preprint arXiv:2302.06100*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Kai Chen, Chunwei Wang, Kuo Yang, Jianhua Han, Lanqing Hong, Fei Mi, Hang Xu, Zhengying Liu, Wenyong Huang, Zhenguo Li, et al. 2023a. Gaining wisdom from setbacks: Aligning large language models via mistake analysis. *arXiv preprint arXiv:2310.10477*.

Meiqi Chen, Yubo Ma, Kaitao Song, Yixin Cao, Yan Zhang, and Dongsheng Li. 2023b. Learning to teach large language models logical reasoning.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

John Joon Young Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. *arXiv preprint arXiv:2306.04140*.

Maxwell John Cresswell. 1973. *Logics and Languages*. Routledge, London, England.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4143–4149.

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.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, et al. 2022. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*.

Alan Hausman. 2012. *Logic and Philosophy: A Modern Introduction*. Wadsworth, Cengage Learning, Boston, MA.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Patrick J. Hurley. 2000. *A Concise Introduction to Logic*. Wadsworth, Belmont, CA.

Lucja Iwańska. 1993. Logical reasoning in natural language: It is all about knowledge. *Minds and Machines*, 3:475–510.

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.

Pratik Joshi, Somak Aditya, Aalok Sathe, and Monojit Choudhury. 2020. Taxinli: Taking a ride up the nlu hill. *arXiv preprint arXiv:2009.14505*.

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: Synthie and the case of information extraction. *arXiv preprint arXiv:2303.04132*.

Robert Kowalski. 1974. *Logic for problem solving*. Department of Computational Logic, Edinburgh University.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv preprint arXiv:2310.07849*.

Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. 2021. Natural language inference in context-investigating contextual reasoning over long texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13388–13396.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Umar Maqsud. 2015. Synthetic text generation for sentiment analysis. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 156–161.

G Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. *arXiv preprint arXiv:2009.02696*.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agrawal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason.

Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2023. Is a prompt and a few samples all you need? using gpt-4 for data augmentation in low-resource classification tasks. *arXiv preprint arXiv:2304.13861*.

Santiago Ontanon, Joshua Ainslie, Vaclav Cvicek, and Zachary Fisher. 2022. Logicinference: A new dataset for teaching logical inference to seq2seq models. *arXiv preprint arXiv:2203.15099*.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Yannis Papanikolaou and Andrea Pierleoni. 2020. Dare: Data augmented relation extraction with gpt-2. *arXiv preprint arXiv:2004.13845*.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Zhivar Sourati, Filip Ilievski, Hông-Ân Sandlin, and Alain Mermoud. 2023. Case-based reasoning with language models for classification of logical fallacies. *arXiv preprint arXiv:2301.11879*.

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.

Peter Swanborn. 2010. Case study research: What, why and how? *Case study research*, pages 1–192.

Zhiyang Teng, Ruoxi Ning, Jian Liu, Qiji Zhou, Yue Zhang, et al. 2023. Glore: Evaluating logical reasoning of large language models. *arXiv preprint arXiv:2310.09107*.

Christopher W Tindale. 2007. *Fallacies and argument appraisal*. Cambridge University Press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Siyuan Wang, Zhongkun Liu, Wanjun Zhong, Ming Zhou, Zhongyu Wei, Zhumin Chen, and Nan Duan. 2022. From lsat: The progress and challenges of complex reasoning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2201–2216.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. Help: A dataset for identifying shortcomings of neural models in monotonicity reasoning. *arXiv preprint arXiv:1904.12166*.

Fei Yu, Hongbo Zhang, and Benyou Wang. 2023. Nature language reasoning, a survey. *arXiv preprint arXiv:2303.14725*.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. *arXiv preprint arXiv:2002.04326*.

Hanlin Zhang, Jiani Huang, Ziyang Li, Mayur Naik, and Eric Xing. 2023. Improved logical reasoning of language models via differentiable symbolic programming. *arXiv preprint arXiv:2305.03742*.

A Details of Five LFU Tasks

We list the definitions and examples of our five tasks below.

Dimension: WHAT

- **Task1:** Identification
- **Definition:** Identify whether the given sentence has logical fallacy.
- **Example:**
  Sentence: Many people believe most museums will be closed on Mondays, therefore it's a fact.
  Identify if there is any logical fallacy in the sentence.
  A) Yes, there is a logical fallacy.
  B) No, there is no logical fallacy.

- **Task2:** Classification
- **Definition:** Select the sentence belonging to a certain type of logical fallacy.
- **Example:**
  Circular reasoning occurs when an argument uses the claim it is trying to prove as proof that the claim is true.
  Select which among the following options demonstrates the logical fallacy of circular reasoning.
  A) Most people believe that Rebecca doesn't like spicy food, therefore it must be true.
  B) Rebecca, a renowned food critic, does not like spicy food. Hence, spicy food is not good.
  C) Rebecca either refrains from spicy food due to discomfort it causes her, or she lacks well-developed taste buds.
  D) Rebecca doesn't like spicy food because she dislikes spicy food.

Dimension: WHY

- **Task3:** Deduction
- **Definition:** Derive the conclusion from the premise according to a certain type of logical fallacy.
- **Example:**
  Faulty generalization occurs when a conclusion about all or many instances of a phenomenon is drawn from one or a few instances of that phenomenon.
  The premise is known: Bob painted his house green and he is a homeowner.
  With which of the two conclusions can the premise be coupled to create logical fallacy of

Fallacy Type	Description	Example
Faulty Generalization	Faulty generalization occurs when a conclusion about all or many instances of a phenomenon is drawn from one or a few instances of that phenomenon.	Kevin, who is a teenager, enjoys playing chess. Therefore, all teenagers must enjoy playing chess.
False Causality	False causality occurs when an argument jumps to a conclusion implying a causal relationship without supporting evidence.	Whenever David goes hiking in the mountains, it's a sunny day. Clearly, David's hiking trips cause sunny weather.
Circular Claim	Circular reasoning occurs when an argument uses the claim it is trying to prove as proof that the claim is true.	Some students are not serious about their studies because they do not focus on their studies.
Ad Populum	Ad populum occurs when an argument is based on affirming that something is real or better because the majority thinks so.	It's widely believed that Nancy relocated to another city, so it must be true.
Ad Hominem	Ad hominem is an irrelevant attack towards the person or some aspect of the person who is making the argument, instead of addressing the argument or position directly.	John claims that all people should obey the rules of the road. But John has received several speeding tickets in the past. Therefore, it's not necessary to obey the rules of the road.
Deductive Fallacy	Deductive fallacy occurs when there is a logical flaw in the reasoning behind the argument, such as Affirming the consequent, Denying the antecedent, Affirming a disjunct and so on.	Should Lucy feel alone, she will surely adopt a puppy. It's evident Lucy has adopted a puppy. Therefore, it must be that Lucy is feeling lonely.
Appeal to Emotion	Appeal to emotion is when emotion is used in place of reason to support an argument in place of reason, such as pity, fear, anger, etc.	Jack had his wallet stolen at the concert, think about how desperate and helpless Jack is now, how can we not help him?
False Dilemma	False dilemma occurs when incorrect limitations are made on the possible options in a scenario when there could be other options.	Most museums will be closed on Mondays either due to low visitor turnout, or due to their disregard for public interest.
Equivocation	Equivocation is an argument which uses a key term or phrase in an ambiguous way, with one meaning in one portion of the argument and then another meaning in another portion of the argument.	All stars are exploding balls of gas. Miley Cyrus is a star. Therefore, Miley Cyrus is an exploding ball of gas.
Fallacy of Extension	Fallacy of extension is an argument that attacks an exaggerated or caricatured version of your opponent's position.	Alex: All flowers don't stay open forever. Jamie: So you're saying that all flowers die instantly after they bloom?
Fallacy of Relevance	Fallacy of relevance, which is also known as Red Herring, occurs when the speaker attempts to divert attention from the primary argument by offering a point that does not suffice as counterpoint/supporting evidence (even if it is true).	A portion of the inhabitants of this city have a fever, but have you considered the high unemployment rate?
Fallacy of Credibility	Fallacy of credibility is when an appeal is made to some form of ethics, authority, or credibility.	Sharon, an acclaimed pianist with years of experience, claims that practicing every day will increase your piano skills by 50%. She's an expert, therefore we should believe her.
Intentional Fallacy	Intentional fallacy is a custom category for when an argument has some element that shows the intent of a speaker to win an argument without actual supporting evidence.	Since no one can prove that Peter didn't come to China last year, he must have.

Table 10: Descriptions and examples of 13 logical fallacy types

faulty generalization?
A) Green is the most popular house color.
B) All homeowners paint their houses green.

- **Task4:** Backward Deduction
- **Definition:** Infer the premise from the conclusion according to a certain type of logical fallacy.
- **Example:**
  Ad populum occurs when an argument is based on affirming that something is real or better because the majority thinks so.
  The conclusion is known: Cynthia's painting must be a masterpiece.
  With which of the two premises can the conclusion be coupled to create the logical fallacy of ad populum?
  A) People widely agree that Cynthia made a beautiful painting.
  B) A famous art critic praised Cynthia's painting.

Dimension: HOW

- **Task5:** Modification
- **Definition:** Correct the logical fallacy in the given sentence.
- **Example:**
  Original sentence: Person A: The garden needs watering. Person B: So you're saying we should neglect everything else and just focus on the garden?
  Correct the logical fallacy in the original sentence and output the modified sentence without any logical fallacy.

B Details of Logic Fallacy Types

In Table 10, we showcase the description and examples of 13 logical fallacy types.

C Details of Manual Proofreading

The evaluation standard is strictly classified into two main categories: structural integrity and validity of fallacies. Structural integrity focuses on the correctness of grammar, the accuracy of punctuation, and the proper use of syntax. On the other hand, the validity of fallacies ensures that, under specific contexts or themes, the sentences satisfy the need for specific type of logical fallacy. Besides, any offensive and harmful data will be refined during the proofreading.

In this process, we assembled an expert team proficient in linguistics and logic. This team comprises four members, including one logician and three graduate students, who are engaged in linguistics, logic, and computer science respectively. They each have the ability to understand and classify various types of logical fallacies.

To enhance the efficiency of this process, each sentence was initially processed through Grammarly, eliminating basic grammatical and lexical errors. Subsequently, our expert team manually reviewed the content. Each sentence was assigned to two team members for review. A consensus confirmed the sentence met the requirements, but in case of disagreement, the third team member would be consulted. If three members cannot achieve consensus, the logician will make the final decision.

D Examples of Logical Reasoning Datasets

We illustrate data examples of four logical reasoning datasets selected in our experiments, including FOLIO, TaxiNLI, LogiQA, and Reclor.

FOLIO

Premise: Beasts of Prey is either a fantasy novel or a science fiction novel. Science fiction novels are not about mythological creatures. Beasts of Prey Is about a creature known as the Shetani. Shetanis are mythological.

Conclusion: Beasts of prey isn't a science fiction novel.

Answer: True

TaxiNLI

Premise: Even if auditors do not follow such other standards and methodologies, they may still serve as a useful source of guidance to auditors in planning their work under GAGAS.

Hypothesis: Auditors should ignore them when they follow other standards and methodologies.

Label: Contradiction

LogiQA2.0

Passage: For a television program about astrology, investigators went into the street and found twenty volunteers born under the sign of Gemini who were willing to be interviewed on the program and to take a personality test. The test confirmed the investigators' personal impressions that each of the volunteers was more sociable and extroverted than people are on average. This modest investigation thus supports the claim that one's astrological birth sign influences one's personality.

Question: Which one of the following, if true, indicates the most serious flaw in the method used by the investigators?

A. People born under astrological signs other than Gemini have been judged by astrologers to be much less sociable than those born under Gemini.

B. There is not likely to be a greater proportion of people born under the sign of Gemini on the street than in the population as a whole.

C. People who are not sociable and extroverted are not likely to agree to participate in such an investigation.

D. The personal impressions the investigators first formed of other people have tended to be confirmed by the investigators' later experience of those people.

Answer: C

Reclor

Context: Geologist: A new method for forecasting earthquakes has reliably predicted several earthquakes. Unfortunately, this method can predict only that an earthquake will fall somewhere within a range of two and a half points on the Richter scale. Thus, since a difference of two and a half points can be the difference between a marginally perceptible shaking and a quake that causes considerable damage, the new method is unlikely to be useful.

Question: Which one of the following, if assumed, enables the geologist's conclusion to be properly inferred?

A. An earthquake-forecasting method is unlikely to be useful unless its predictions always differentiate earthquakes that are barely noticeable from ones that result in substantial destruction.

B. Several well-established methods for forecasting earthquakes can predict within much narrower ranges than two and a half points on the Richter scale.

C. Even if an earthquake-forecasting method makes predictions within a very narrow range on the Richter scale, this method is not likely to be useful unless its predictions are reliable.

D. An earthquake-forecasting method has not been shown to be useful until it has been used to reliably predict a large number of earthquakes.

Answer: A