Are LLMs Good Zero-Shot Fallacy Classifiers?

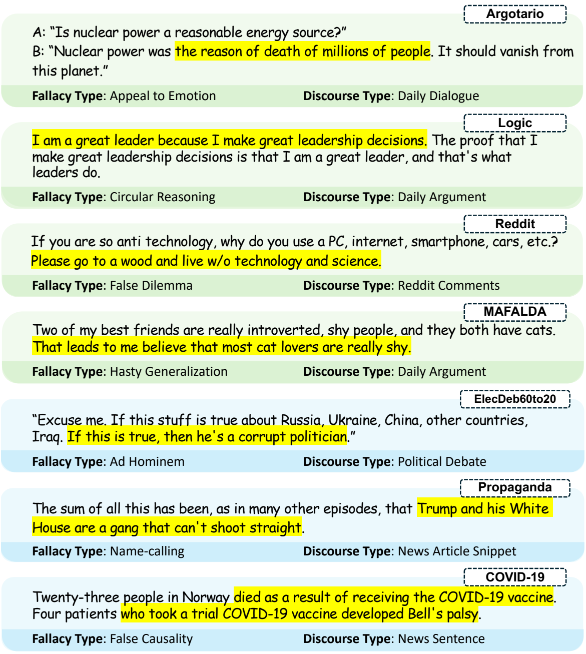
Fengjun Pan[1,2*] **Xiaobao Wu**[2*] **Zongrui Li**[1] **Anh Tuan Luu**[2✉]
[1]Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore
[2]College of Computing and Data Science, Nanyang Technological University, Singapore
panf0004@e.ntu.edu.sg xiaobao002@e.ntu.edu.sg
zongrui001@e.ntu.edu.sg anhtuan.luu@ntu.edu.sg

Abstract

Fallacies are defective arguments with faulty reasoning. Detecting and classifying them is a crucial NLP task to prevent misinformation, manipulative claims, and biased decisions. However, existing fallacy classifiers are limited by the requirement for sufficient labeled data for training, which hinders their out-of-distribution (OOD) generalization abilities. In this paper, we focus on leveraging Large Language Models (LLMs) for zero-shot fallacy classification. To elicit fallacy-related knowledge and reasoning abilities of LLMs, we propose diverse single-round and multi-round prompting schemes, applying different task-specific instructions such as extraction, summarization, and Chain-of-Thought reasoning. With comprehensive experiments on benchmark datasets, we suggest that LLMs could be potential zero-shot fallacy classifiers. In general, LLMs under single-round prompting schemes have achieved acceptable zero-shot performances compared to the best full-shot baselines and can outperform them in all OOD inference scenarios and some open-domain tasks. Our novel multi-round prompting schemes can effectively bring about more improvements, especially for small LLMs. Our analysis further underlines the future research on zero-shot fallacy classification. Codes and data are available at: `https://github.com/panFJCharlotte98/Fallacy_Detection`.

1 Introduction

A fallacy is a defective argument derived from erroneous or invalid reasoning that may appear to be reasonable but are, in fact, logically unsound or faulty (Woods, 2004; Damer, 2008; Van Eemeren et al., 2009; Hamblin, 2022). Figure 1 illustrates examples of different fallacy types, for instance, a fallacy type of Circular Reasoning: *I am a great leader because I make great leadership decisions.*



Argotario

A: "Is nuclear power a reasonable energy source?"
B: "Nuclear power was the reason of death of millions of people. It should vanish from this planet."

Fallacy Type: Appeal to Emotion **Discourse Type**: Daily Dialogue

Logic

I am a great leader because I make great leadership decisions. The proof that I make great leadership decisions is that I am a great leader, and that's what leaders do.

Fallacy Type: Circular Reasoning **Discourse Type**: Daily Argument

Reddit

If you are so anti technology, why do you use a PC, internet, smartphone, cars, etc.? Please go to a wood and live w/o technology and science.

Fallacy Type: False Dilemma **Discourse Type**: Reddit Comments

MAFALDA

Two of my best friends are really introverted, shy people, and they both have cats. That leads to me believe that most cat lovers are really shy.

Fallacy Type: Hasty Generalization **Discourse Type**: Daily Argument

ElecDeb60to20

"Excuse me. If this stuff is true about Russia, Ukraine, China, other countries, Iraq. If this is true, then he's a corrupt politician."

Fallacy Type: Ad Hominem **Discourse Type**: Political Debate

Propaganda

The sum of all this has been, as in many other episodes, that Trump and his White House are a gang that can't shoot straight.

Fallacy Type: Name-calling **Discourse Type**: News Article Snippet

COVID-19

Twenty-three people in Norway died as a result of receiving the COVID-19 vaccine. Four patients who took a trial COVID-19 vaccine developed Bell's palsy.

Fallacy Type: False Causality **Discourse Type**: News Sentence

Figure 1: Examples of fallacies and their types from existing datasets.

Fallacies commonly appear in various scenarios, such as news articles (Musi and Reed, 2022), advertisements (Danciu et al., 2014), propaganda (Walton, 1997), politics (Blassnig et al., 2019), and social media (Hidayat et al., 2020). They could be intentionally exploited to disseminate misinformation (Musi et al., 2022a), manipulate public opinions, undermine rational discussions, and influence critical decision-making (Visser et al., 2020). In consequence, detecting and classifying fallacies becomes an imperative challenge.

However, existing fallacy classification methods typically follow a full-shot supervised fine-tuning manner, limited by three issues: (**i**) They require sufficient labeled data for training (Goffredo et al., 2022; Alhindi et al., 2023a), but collecting these data is often time-consuming and expensive. This is because annotating fallacy data typically rely on expert knowledge, due to the complex and esoteric nature of fallacies (Helwe et al., 2023). (**ii**) They cannot well generalize to out-of-distribution (OOD)

---

[*] Equal contribution. ✉ Corresponding Author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 14338–14364
November 12-16, 2024 ©2024 Association for Computational Linguistics

fallacies and discourse types (Liu et al., 2021) due to the inherent limitation of supervised learning. Once given an unseen fallacy class or discourse from other domains, they have to retrain a new model from scratch, which costs considerable computational and time resources. **(iii)** They are susceptible to imbalanced data (Alhindi et al., 2023b). They could reach high performance on the dominant fallacy types while low on infrequent ones. As a result, full-shot fallacy classifiers cannot fulfill real-world application scenarios.

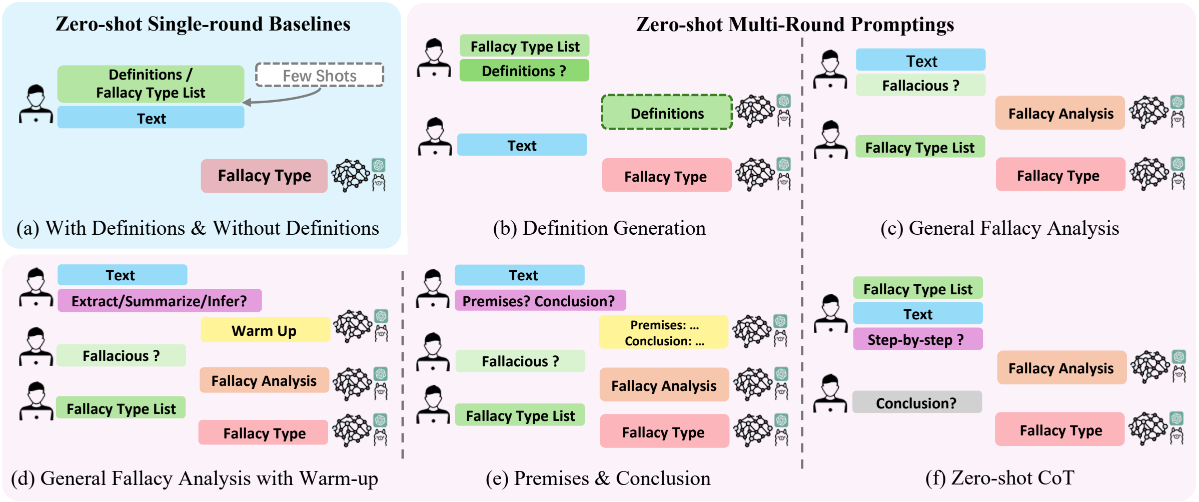
Motivated by the above issues, in this paper, we explore zero-shot fallacy classification, *i.e.*, classifies fallacies without training data. Specifically, we concentrate on Large Language Models (LLMs) as they have been extensively pretrained and possess wide knowledge and strong reasoning abilities. To investigate LLMs' performance on this task, we consider two kinds of prompting schemes. First, we employ a **single-round prompting scheme**. We simply prompt LLMs to classify fallacies either with or without manually crafted fallacy type definitions.

Second, to elicit the inherent fallacy-related knowledge and reasoning abilities of LLMs, we further propose diverse **multi-round prompting schemes**: we instruct LLMs to analyze and classify fallacies through definition generation, general fallacy analysis (with warm-up), premise & conclusion analysis, and Chain-of-Thought. Under these prompting schemes, we observe that zero-shot prompted LLMs can outperform or achieve comparable performances with SOTA full-shot fine-tuned T5 baselines on some open-domain benchmark datasets and can generally achieve sub-optimal performances on hard domain-specific datasets. More importantly, LLMs consistently demonstrate advantage over full-shot trained baselines on OOD inferences on low-resource fallacy classification tasks. Besides, our novel multi-round prompting schemes can effectively improve LLMs' classification performance compared to the single-round prompts, especially for small LLMs, *e.g.*, Llama3, Qwen2.5 and Mistral. We conclude our main contributions as below:

- We propose diverse novel prompting schemes, including both the basic single-round and the advanced multi-round ones, that are sufficiently effective in eliciting the fallacy-related knowledge and reasoning abilities of LLMs for zero-shot fallacy classification.
- We are the first to conduct extensive experiments with representative LLMs on existing fallacy benchmark datasets and provide overviews and insights concerning the boundary of LLMs' zero-shot fallacy classification performances. We show that we can resort to LLMs as potential zero-shot fallacy classifiers that solely rely on LLMs' inherent knowledge without computationally intensive training full-shot models from scratch.
- We provide detailed analysis on LLMs, datasets, and prompting schemes and summarize a general guidance for choosing the potentially most effective prompting schemes for different fallacy classification scenarios with respect to data domains and LLM types, inspiring future research on zero-shot fallacy classification and other relevant linguistic reasoning tasks.

2 Related Work

Fallacy Classification Fallacy detection and classification is an emerging natural language processing task that has received increasing attention. Many fallacy datasets across different discourse genres and various domains have been created in a line of work (Habernal et al., 2017, 2018b; Da San Martino et al., 2019; Sahai et al., 2021; Goffredo et al., 2022; Jin et al., 2022; Alhindi et al., 2023a; Musi et al., 2022b; Helwe et al., 2023). Early fallacy classification methods are mainly based on traditional machine learning models (Wu et al., 2020, 2022, 2024b,a), *e.g.*, SVM, Bi-LSTM (Habernal et al., 2018a) and CNN (Habernal et al., 2018b). Deep learning methods based on models with Transformers architecture *e.g.*, BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), T5 (Raffel et al., 2020) have been proposed in recent works (Da San Martino et al., 2019; Sahai et al., 2021; Goffredo et al., 2023; Jin et al., 2022; Alhindi et al., 2023a). However, all these methods are trained in a supervised manner that is highly reliant on the availability and sufficiency of labeled fallacy data, rendering them struggling on OOD fallacy and discourse types. Recently, LLMs' abilities related to fallacy understanding and detection have been explored (Payandeh et al., 2023; Alhindi et al., 2023b; Lim and Perrault, 2024; Lalwani et al., 2024; Ruiz-Dolz and Lawrence, 2023) but not evaluated in-depth considering prompt techniques.



(a) With Definitions & Without Definitions (b) Definition Generation (c) General Fallacy Analysis

(d) General Fallacy Analysis with Warm-up (e) Premises & Conclusion (f) Zero-shot CoT

Figure 2: Illustration of single-round and multi-round prompting schemes. (**a**): Prompt LLMs to classify with or without fallacy type definitions. (**b**): Prompt LLMs to generate fallacy type definitions and then classify. (**c**): Prompt LLMs to analyze the input discourse and then classify. (**d**): Prompt LLMs to warm up (extract, summarize, or infer), analyze the input discourse, and then classify. (**e**): Prompt LLMs to extract the premises and conclusion, analyze the input discourse, and then classify. (**f**): Prompt LLMs to reason step by step to classify and then draw an answer.

Large Language Models The ability of LLMs in logical reasoning has been quantitatively or qualitatively evaluated, *e.g.*, commonsense causality reasoning (Kıcıman et al., 2023; Talmor et al., 2020; Willig et al., 2022) and abstract reasoning (Gendron et al., 2023; Pan et al., 2023; Wu et al., 2024c). Particularly, evidence (Bubeck et al., 2023) has verified the existence of inconsistency in its reasoning process (same for the latest LLMs like GPT-4). This motivates us to guide LLMs through the proposed multi-round prompting schemes. Our work distinguishes from the previous in two points: (**i**) instead of a general reasoning task, we focus on the fallacy classification, a more advanced reasoning task that requires not only expert knowledge of but also rigorous logical reasoning. (**ii**) Apart from the common single-round prompting, we propose the novel multi-round prompting schemes that bring along noticeable improvements for small LLMs.

3 Methodology

In this section, we apply LLMs for zero-shot fallacy classification by two kinds of prompting schemes: **Zero-shot Single-Round Prompting** and **Zero-shot Multi-Round Prompting**.

3.1 Zero-shot Single-Round Prompting

We first propose zero-shot single-round prompting schemes. Following the state-of-the-art fine-tuning baseline (Alhindi et al., 2023a), the single-round prompt consists of three components, as shown in Figure 2 (a): (i) A label space that defines a limited number of fallacy types; (ii) An input fallacious discourse to be classified; (iii) The task and output format instructions. For the label space, we consider two settings: (i) **Without Definitions**, which merely enumerates all the viable fallacy types as an option list, for example, *1. Appeal to Emotion, 2. Ad Hominem, 3. False Dilemma ....* (ii) **With Definitions**, which means that we provide the natural language definition of each fallacy type, for example, *1. Appeal to Emotion is a fallacy when someone attempts to argue or persuade by using emotive language to arouse non-rational sentiments within the intended audience. 2. Ad Hominem is a fallacy when someone attacks the others' characters or motives instead of addressing the substance of their arguments...* See the used single-round prompt templates in Table 29.

3.2 Zero-shot Multi-Round Prompting

Besides, we propose zero-shot multi-round prompting schemes. This is motivated by the complexity nature of the fallacy classification task as it involves reading comprehension, information extraction, logical reasoning, knowledge recall, and pattern recognition. Therefore, we introduce the following multi-round prompting schemes that aim to elicit LLMs' inherent knowledge and reasoning abilities on fallacy. Figure 2 summarizes all these schemes.

Definition Generation As illustrated in Figure 2 (b), we prompt LLMs to generate the definition for each fallacy given in the fallacy type list in the first round, and then classify the fallacy type of the input discourse based on these definitions in the second round. This is because LLMs probably have learned the knowledge about different fallacy types during pretraining. This scheme elicits LLMs to recall these knowledge. Besides, the generated definitions better align with LLMs' understanding compared to manually crafted definitions due to their auto-regressive paradigm, thus may enhance the reasoning process of determining fallacy types.

General Fallacy Analysis As shown in Figure 2 (c), we first instruct LLMs to analyze the input discourse and determine whether it is logically reasonable or potentially fallacious; then we ask LLMs to determine the fallacy type. The first round works as an intermediate step to offer analytical information for any detected potential fallacy, which assists the fallacy classification in the second round. Note that here we use a neutral instruction which includes both positive (logically reasonable) and negative (potentially fallacious) possibilities to avoid any biased implication.

General Fallacy Analysis with Warm Up In this scheme (Figure 2 (d)), we add a warm-up round as the first round that asks LLMs to extract, paraphrase, summarize or infer about the content and context of the input discourse, and then follow the above General Fallacy Analysis scheme. This scheme is inspired by the fact that discourses in some domain-specific datasets are truncated, which greatly hinders LLMs' understanding. The warm-up round brings more contextual information to ease LLMs' understanding, which thus benefits their predictions on fallacy types.

Premises & Conclusion Helwe et al. (2023) give a formal definition of the term "fallacy": *A fallacy is an argument where the premises do not entail the conclusion*. Following this formal definition, we use three rounds as shown in Figure 2 (e): First, we prompt LLMs to extract the premises and conclusion of the input discourse. Second, we ask LLMs to determine whether the input is fallacious by analyzing if the premises entail the conclusion according to the formal definition. Finally, LLMs predict the fallacy type.

Zero-shot CoT Lastly, we consider zero-shot CoT (Chain-of-Thought). Previous multi-round prompting schemes all apply task-specific instructions in intermediate steps. Differently as shown in Figure 2 (f) zero-shot CoT directly leverages the magic instruction *Now, let's think step by step* (Kojima et al., 2022) to prompt LLMs to classify fallacy types through step-by-step reasoning. This elicits LLMs to derive analogous reasoning chains (Besta et al., 2024).

Domain	Dataset	Splits (train/dev/test)	#FT.	DT.
Open	Argotario	909/102/312	5	QA
	Logic	1849/300/300	13	Unlimited
	Reddit	1195/342/513	8	Reddit Comments
	Mafalda	0/0/200	23	Unlimited
Politics	ElecDeb	1267/136/154	5	Debate Transcripts
News	Propaganda	1583/265/265	13	News Articles
COVID News	COVID-19	0/0/154	9	News & Posts

Table 1: Fallacy dataset statistics. **#FT.**: Number of fallacy types. **DT.**: Discourse type.

4 Experiments

In this section, we conduct extensive experiments and evaluate model performance by **Macro F1** following Alhindi et al. (2023a).

Datasets. We consider the following 7 benchmark datasets: **(i)** **Argotario** (Habernal et al., 2017) contains short QA pairs about various open topics. The answers may commit one of the 5 fallacy types or no fallacy. **(ii)** **Logic** (Jin et al., 2022) consists of fallacy examples in shot statements or conversations collected from education websites across 13 different fallacy types. **(iii)** **Reddit** (Sahai et al., 2021) collects Reddit comments submitted for different discussion topics. Each data example contains one of the 8 types of fallacies that may occur in multiple spans. Since we find this dataset to be the most balanced one, we combine its validation and test set as the inference set and hold it as an OOD dataset when testing the full-shot baseline. **(iv)** **Mafalda** (Helwe et al., 2023) merges and unifies previous fallacy datasets and annotate 200 examples in sentence level with a total of 23 fallacy types. As this dataset contains data examples with multi-labels across different sentences, we take the most dominant fallacy type with most occurrences within each example as the single fallacy label to adhere the unified multi-class single-label classification setup. We also consider this dataset to be OOD as it has insufficient instances for fine tuning. **(v)** **ElecDeb60to20** (**ElecDeb** hereinafter, Goffredo et al., 2022) proposes a fallacy corpus of political speeches in U.S. presidential election debates. Each data instance

Shot	Round	Model	ARGOTARIO	LOGIC	REDDIT	ELECDEB	PROPAGANDA	MAFALDA	COVID-19
Full	N/A	T5-3B Single-task	69.13	**64.95**	**83.20**	**62.37**	38.36	-	-
		T5-3B Multi-ALR	**72.38**	63.54	81.88	33.22	12.58	31.52	12.28
		T5-3B Multi-ALEP	70.51	61.65	56.98	56.35	**43.33**	**35.60**	**14.59**
		T5-large Single-task	58.26	55.23	77.77	41.48	38.62	-	-
		T5-large Multi-ALR	65.65	59.48	80.42	37.22	8.82	25.13	13.07
		T5-large Multi-ALEP	64.14	57.67	38.87	56.15	39.75	25.60	14.08
Zero	Single	GPT-4	78.94	50.43	79.10	42.26°	34.8°	48.74°	20.47
		GPT-3.5	63.59°	39.65	70.42°	41.01	22.39	31.27	17.45
		Qwen2.5-Instruct 14B	68.19°	41.82°	67.72°	37.28	21.89	33.03°	17.6
		Qwen2.5-Instruct 7B	61.38°	35.48°	58.58	43.34	16.03	31.27°	15.19°
		Llama3-Chat 8B	48.87	27.45°	49.41	39.36°	17.30	24.85°	11.00°
		Mistral-Instruct 7B	57.04°	28.99°	46.89°	33.23°	16.89°	23.44°	14.69
		Llama2-Chat 13B	50.20	25.11	34.15°	35.57	10.61	22.09	14.15
	Multi	GPT-4	**79.87**	**50.54**	**81.11**	41.25	**35.37**	**52.86**	**25.18**
		GPT-3.5	68.40	41.11	71.08	37.77	26.67	40.73	17.24
		Qwen2.5-Instruct 14B	68.87	45.89	77.08	34.09	26.60	45.94	23.73
		Qwen2.5-Instruct 7B	60.20	40.09	64.58	**44.55**	19.22	35.37	22.88
		Llama3-Chat 8B	61.39	35.66	57.83	40.81	21.35	34.18	19.83
		Mistral-Instruct 7B	57.26	31.43	59.70	32.91	20.41	29.08	18.53
		Llama2-Chat 13B	48.79	28.85	45.82	36.37	11.11	15.68	14.16

Table 2: Fallacy classification results of Macro-F1. The best results obtained by T5 baselines are in **bold**, and our best zero-shot results obtained by LLMs are in **bold**. Blue denotes out-of-distribution (OOD) results of T5 baselines while Red denotes the corresponding zero-shot results where single-round prompted LLMs outperform the OOD results of T5 baselines; $\circ$ denotes single-round prompting without definitions, while other single-round results are with definitions. Here we only report the best zero-shot results of LLMs across all prompting schemes. See Appendix B for detailed results.

is a shot snippet that contains one type of fallacies. We keep 5 fallacy types and remove the examples of *Slogans* from the original dataset as it may not strictly align with the definition of a fallacy as discussed in (Helwe et al., 2023). **(vi) PROPAGANDA** (Da San Martino et al., 2019) is a large corpus of propaganda techniques used in news articles annotated at sentence level. We exclude 5 propaganda-oriented classes (*Loaded Language*, *Exaggeration or Minimisation*, *Thought-terminating Cliches*, *Slogans*, *Repetition*) and keep a total of 13 fallacy types. We include the four most adjacent sentences before and after each annotated fallacious sentence as its surrounding context. **(vii) COVID-19** (Musi et al., 2022b) is a fallacy corpus of news sentences and media posts about COVID pandemic. Each data examples could be non-fallacious or contain one of the 9 types of fallacies. Considering its small size, we combine all the data instances in three available splits as the inference set and hold it as OOD for full-shot baseline. Table 1 summarizes the dataset statistics. Fallacy class distribution in each dataset can be found in Table 26.

Scheme	#R	Rank	%Failed
General Fallacy Analysis with Warm Up	3	4.13	12.27
Zero-shot CoT	2	4.33	5.99
General Fallacy Analysis	2	4.35	15.09
With Definitions	1	4.65	4.10
Definition Generation	2	4.77	2.85
Without Definitions	1	4.90	5.04
Premises & Conclusion	3	5.84	12.46

Table 3: Overall rankings on Macro-F1 of multi-round prompting schemes. **#R**: Number of rounds.

Baselines We consider the state-of-the-art unified fallacy classification model based on T5 (Raffel et al., 2020) proposed by Alhindi et al. (2023a) as the baseline model since it is the existing most comprehensive SOTA baseline that can be applied to diverse extant fallacy benchmark datasets. Baselines in prior works *e.g.*, (Jin et al., 2022) and (Goffredo et al., 2023) can only be applied to one specific fallacy dataset under highly specialized conditions on data annotations, fallacy forms, and additional data features, *etc.*. Due to such limited generalizability and reproducibility, these baselines are currently out of our scope. T5 models are implemented in a deployment-efficient unified multitask paradigm, thus serving as a suitable and strong full-shot baseline. We follow the instruction-based prompts and hyper-parameter setups used in their original paper (See Appendix A) and replicate three full-shot fine-tuned baseline settings: **(i) Single-task** is fine-

Domain	Base Model	Scheme	#R	Rank	%Failed
Open	GPT	DG	2	2.88	3.28
		WD	1	3.25	4.00
	Small (≤14B)	GFA-W	3	2.85	11.18
		CoT	2	3.79	5.19
Specific	GPT	WD	1	2.50	3.65
		CoT	2	3.33	7.93
	Small (≤14B)	CoT	2	3.20	7.57
		GFA-W	3	3.80	15.91

Table 4: Best two zero-shot prompting schemes for different base models and data domains based on average Macro-F1 rankings. DG: Definition Generation. CoT: Zero-shot CoT. GFA-W: General Fallacy Analysis with Warm Up. GFA: General Fallacy Analysis. WD: With Definitions. **#R**: Number of rounds.

tuned on each dataset individually. (**ii**) **Multi-ALR** is fine-tuned on three open-domain datasets Argotario, Logic and Reddit. (**iii**) **Multi-ALEP** is fine-tuned on two open-domain datasets Argotario, Logic and two domain-specific datasets ElecDeb and Propaganda. We implement each baseline setting with both **T5-3B** and **T5-large**. We further apply up-sampling Xie et al. (2022) techniques to balance the size of multitask training data for better performance during fine-tuning. We report the average results of 3-time repeated experiments for each T5 baseline.

LLMs. We consider seven recent representative LLMs that have been fine-tuned for powerful instruction following behaviors thus are capable for multi-round chat-based Q&A: (**i**) **GPT-4** (Nori et al., 2023), (**ii**) **GPT-3.5**, (**iii**) **Llama3-Chat 8B**, (**iv**) **Qwen2.5-Instruct 14B**, (**v**) **Qwen2.5-Instruct 7B**, (**vi**) **Mistral-Instruct 7B** (Jiang et al., 2023) and (**vii**) **Llama2-Chat 13B**. For reproducibility, we report the average results of 3-time repeated experiments for GPTs and the average 5-time repeated experimental results for other small LLMs.

4.1 Zero-shot Single-round Prompting Schemes vs. Fine-Tuned T5 Baselines

We report the results of fine-tuned T5 baselines and LLMs with zero-shot single-round prompting schemes in Table 2. We have the following observations. First, zero-shot single-round prompted GPT-4 can outperform both T5-3B and T5-large fine-tuned baselines on Argotario with the best result of 78.94 and can achieve highly competitive results of 79.10 on Reddit as compared to the best results of 83.20 and 80.42 obtained by T5-3B and T5-large, respectively. Its performances on Logic, ElecDeb and Propaganda are comparable to the best of full-shot T5-large baselines but are still behind the best of T5-3B. Except GPT-4, the other LLMs fail to outperform the best fine-tuned T5 baselines across all datasets under a fully zero-shot setting, which evidences the difficulty of zero-shot fallacy classification.

Second, zero-shot prompted LLMs demonstrate stronger generalization abilities than the fine-tuned T5 baselines. As can be seen, LLMs can largely outperform T5 baselines on all the OOD holdout datasets of Mafalda, Covid-19, Reddit, ElecDeb, and Propaganda. Specifically, GPT-4, GPT-3.5 and Qwen2.5 consistently outperform all multitask fine-tuned T5 baselines on four OOD inference scenarios. Llama3 and Mistral can also outperform on two datasets when set as OOD for T5 baselines. This validates the poor generalization ability of fine-tuned T5 baselines since they heavily rely on the amount and beneficial diversity of annotated training data thus may struggle in generalizing to OOD scenarios with unseen fallacy classes and discourse types. Besides, when T5 is scaled up, multitask fine-tuning of T5 could be susceptible to the composition of tasks and does not always translate to improvements for all the tasks in training but could result in deteriorated performance compared to single-task training. In contrast, resorting to zero-shot prompted LLMs can bypass these limitations.

Third, although zero-shot single-round prompted LLMs show considerably promising performances on two open-domain datasets Argotario and Reddit, they are struggling with Logic and Mafalda and the other three domain-specific benchmarks. We consider the following reasons: (**i**) Argotario and Reddit are two most balanced and comprehensible fallacy datasets that have fallacious discourses with common fallacy types delivered in a relatively intuitive, casual, informal utterances about daily contexts. These features may align better with the chat-optimized behaviors of instruction-tuned LLMs. (**ii**) Although Logic and Mafalda contain open-domain contexts, they are difficult because of their large label space (13 and 23 respectively) with edge fallacy types, such as "*Fallacy of Converse*", "*Doubt Credibility*", "*Intentional*", *etc.*. In addition, since Logic is collected from online education websites such as study.com and Quizlet, the noisy contexts in its discourse segments may cause confusion with

Model	Macro-F1 Infomal Def.	 Formal Def.	%Failed Infomal Def.	 Formal Def.
GPT-4	48.38	**49.78**	0.67	0.83
GPT-3.5	31.27	27.10	2.50	4.00
Qwen2.5-14Bf	31.11	**34.64**	0.88	0.70
Qwen2.5-7Bf	30.62	29.43	5.00	2.20
Llama3-8Bf	24.43	**25.71**	12.90	12.10
Mistral-7Bf	22.23	18.37	3.50	3.60
Llama2-13Bf	22.09	21.30	27.70	8.10

Table 5: Informal vs. formal fallacy definitions in the single-round prompting scheme on Mafalda.

Model	Shot	Logic	ElecDeb	Propaganda
GPT-4	One	**54.48**	**44.36**	33.18
	Two	**55.14**	**44.64**	31.57
	Zero°	48.45	42.26	34.80
	Zero*	50.54^{GFA}	42.26°	**35.37**^{CoT}

Table 6: Few-shot vs. zero-shot results of GPT-4. °: Zero-shot single-round results without definitions. *: Best zero-shot results. See Table 23 for other LLMs.

our instructions. (**iii**) In-domain corpora of news and political speeches are delivered in formal language or specific utterance styles and rely on sufficient contexts for comprehension. It could be hard to infer about the fallacy type if the understanding of the content and context is challenged by the truncation of discourses. Besides, the fallacious examples in the three in-domain datasets, though may share the same fallacy names with other datasets, could have nuanced differences in definitions that deviate from the common ones (Alhindi et al., 2023a) and refer to specific language use cases tied to the domain context.

4.2 Single-Round vs. Multi-Round

As shown in Table 2, our multi-round prompting schemes are effective in further improving LLMs' zero-shot performances compared to the basic single-round prompts. In detail (See Table 24), they improve on 83.7% of the classification experiments compared to the basic single-round prompting schemes, with an average improvement across the datasets of 1.63, 2.46, 4.95, 3.67, 7.54, 4.02 and 1.27 points for GPT-4, GPT-3.5, Qwen2.5-Instruct (14B), Qwen2.5-Instruct (7B), Llama3-Chat (8B), Mistral-Instruct (7B) and Llama2-Chat (13B), respectively, indicating the effectiveness of our proposed multi-round prompts, especially for capable small LLMs such as Llama3 and Qwen2.5. Multi-round prompted GPT-4 is also further improved on most tasks compared to its best single-round prompting performance and achieves new state-of-the-art scores on Argotario, Mafalda and Covid-19.

4.3 Multi-Round Prompting Schemes Analysis

We present the overall average Macro-F1 ranking of each multi-round prompting scheme in Table 3. It shows that three of our proposed multi-round prompting schemes can be superior to the basic single-round prompts. General Fallacy Analysis with or without the warm-up round (GFA-W and GFA) ranks within the top three best-performed multi-round schemes, indicating that explicit fallacy analysis can benefit LLMs by eliciting their internalized fallacy knowledge to guide the classification process. However, all the proposed multi-round prompting schemes that require LLMs to first perform tasks that indirectly address the focal classification problem could be prone to failed classification *i.e.*, predicting a fallacy type that is outside the given fallacy label space or predicting as non-fallacious. This is predicable since it is possible that some data examples do commit more than one type of fallacy or can be interpreted as certain out-of-scope fallacies. Having LLMs to make analysis first without limiting them within a specific label space increases the probability that LLMs bias to their first judgements. An additional warm-up round that allows LLMs to first reflect on the content and context of input discourses can not only improve the overall classification performance but also help alleviate the failure rate (12.27% vs. 15.09%). Zero-shot CoT generally ranks as the second-best multi-round prompting scheme, suggesting that CoT's general contribution to improved reasoning in LLMs is also transferable to fallacy classification tasks. Besides, Zero-shot CoT has significant advantages over GFA(-W) if taking the average failure rate into account, which makes Zero-shot CoT the potentially most effective and robust multi-round prompting scheme in practice.

Although on average Definition Generation slightly underperforms the single-round With Definitions, it is still better than prompting LLMs with only the label space. Human crafted fallacy definitions that could be more detailed and relevant to the focused context, if available, are usually more reliable than LLMs' self-generated definitions. It is notable that both Zero-shot CoT and Definition Generation have lower failure rates, proba-

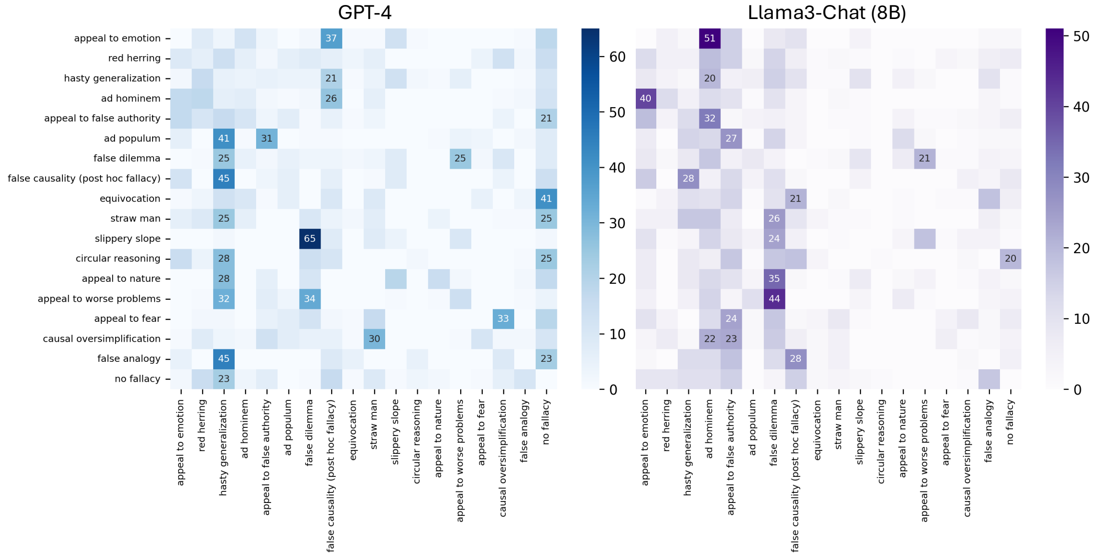


Figure 3: Misclassification confusion matrix of common fallacy types given by GPT-4 and Llama3-Chat (8B). Rows are the ground truth fallacy types and columns are predicted fallacy types. Cell values represent the percentages of row fallacy types that are misclassified as column fallacy types.

bly because they have fewer chat rounds and can immediately address the given label space at the first place, confining LLMs to more restricted contexts that are less likely to be directed to unlimited or irrelevant association. In particular, Definition Generation has the lowest failure rate, indicating that LLMs could be more robust in generating instruction-specified labels when relying on their self-generated contents. In contrast, the three-round Premises & Conclusion that explicitly applies the formal definition of the term “fallacy” is not always a robust choice to improve the fallacy classification performance of LLMs. It turns out that the formal academic conception cannot be well translated to eliciting LLMs’ knowledge and abilities for distinguishing among different fallacy types but mislead LLMs to a deviated focus on checking “whether the premises entail the conclusion” rather than recognizing the specific fallacy type. (See Appendix D.1)

We present the top two best performed prompting schemes for different classification scenarios with respect to data domains and model types in Table 4. This summarization may serve as a general guidance for fallacy classification in practice. As suggested by the results, when applying LLMs to open-domain fallacy classification tasks in a zero-shot setting, the inherent pre-trained fallacy knowledge of advanced GPT models could be effective when explicitly prompted for self-generated fallacy definitions or directly referring to human crafted defintions. For smaller LLMs, they can be prompted to reflect on the input context before giving a general fallacy analysis to better elicit and leverage their internalized relevant knowledge. When tackling domain-specific fallacy classification tasks, to address the nuances in fallacy definitions for in-domain characteristics, it is more effective to require GPT models to directly refer to the manually crafted definitions, while for small LLMs, they can still benefit from Zero-shot CoT prompting. See Table 21 and Table 22 for more detailed rankings and the comparison between best zero-shot results and the guidance-suggested results in Table 25.

4.4 Informal vs. Formal Definitions

We substitute the informal fallacy definitions (Table 27) with formal fallacy definitions (See Table 28) of MAFALDA summarized by Helwe et al. (2023) in the single-round prompting scheme. As shown in Table 5, we can only observe a limited improvement of performance on more capable LLMs of GPT-4, Qwen2.5-Instruct (14B) and Llama3-Chat (8B) when providing them with formal definitions of fallacies. The results of the other LLMs remain comparable with or are slightly worse than that with informal definitions. The failure rates of the two prompting schemes are not significantly different. While formal definitions that incorporate some symbolic representations could be less ambiguous, they are not necessarily easier for LLMs to comprehend and relate to the exact occurrences of fallacies in actual discourses.

4.5 Few-shot vs. Zero-shot

To verify whether LLMs' weak performance on some datasets (*i.e.*, LOGIC, ELECDEB and PROPAGANDA) could have been resulted from the challenging fully zero-shot prompt setting, we conduct N-way few-shot experiments that randomly present LLMs with 1-shot or 2-shot examples of each fallacy classes from the holdout data splits. Experiment results in Table 6 show that only GPT-4's performance can be further improved on LOGIC and ELECDEB. The rest of few-shot results by other LLMs all fail to surpass the previous best-performed zero-shot prompting results, suggesting the limitation of LLMs' in-context learning abilities in handling hard and in-domain fallacy classification tasks.

4.6 Error Analysis

To view how LLMs perform on different fallacy types, we present the confusion matrices of GPT-4 and Llama3-Chat (8B) on each dataset in Appendix D.3. To further analyze classification errors, we aggregate the experimental results of GPT-4 and Llama3 on each dataset under their top three best performed zero-shot prompting schemes and select a total of 18 common fallacy types (including *No Fallacy*) that occur in more than two datasets to present the confusion matrix in misclassification percentage as shown in Figure 3. It is notable that GPT-4 is generally more inclined to classify fallacy examples to be non-fallacious compared to Llama3. As revealed in Figure 7 (a) and Figure 9 (e), GPT-4 can better recall actual *No Fallacy* examples but severely over predict most examples in COVID-19 to be non-fallacious as shown in Figure 8 (b). In addition, some fallacy types frequently confused by LLMs reflect the inherent limitation in the natural language definitions of informal fallacies, which may overlap with each other with inevitable ambiguity. For example, GPT-4 tends to classify many fallacy types as *Hasty Generalization*. We find that GPT-4's self-generated definition for *Hasty Generalization* that "*A conclusion based on insufficient or biased evidence; rushing to a conclusion before you have all relevant facts*." can be applied to interpret many fallacies *e.g.*, *Ad Populum*, *Post Hoc Fallacy* and *False Analogy*. We find similar explanations towards both GPT-4's and Llama3's misclassifations of *Appeal to Worse Problems*, *Slippery Slope* as *False Dilemma* and *Ad Populum* as *Appeal to False Authority*. As for Llama3, we find that it tends to overpredict the fallacy types of *False Dilemma*, *Ad Hominem* and *Appeal to False Authority*. Interestingly, as shown in the corpus of U.S. presidential debate ELECDEB in Figure 8 (d), Llama3 predicts more *Appeal to Emotion*, *Ad Hominem* and *Appeal to False Authority*, which are more basic and superficial fallacy types, while GPT-4 predicts more *Slippery Slope* and *Post Hoc Fallacy* that are fallacy conceptions more related to causal reasoning. This further suggests that the task of fallacy classification requires advanced reasoning abilities. We further analyze the potential reason for the relatively poor performance of GPT-based models under the zero-shot multi-round prompting scheme of Premises & Conclusion in Appendix D.1.

5 Conclusion

In this paper, we explore LLMs' performance on zero-shot fallacy classification. We propose both single-round and multi-round prompting schemes to fully elicit LLMs' fallacy classification ability. Through extensive experiments on benchmark datasets, we demonstrate that LLMs with a zero-shot single-round prompting scheme can outperform or achieve highly comparable performances with SOTA full-shot fine-tuned T5 baselines on some open-domain benchmark datasets and can generally achieve sub-optimal performances on hard domain-specific datasets. Besides, our multi-round prompting schemes can effectively enhance the performances, especially for small LLMs. The performance gaps between zero-shot prompted LLMs and full-shot fine-tuned baselines are acceptable, indicating LLMs' potential for further improvements for this task. Our detailed analysis moreover highlights the future research for zero-shot fallacy classification.

Limitations

We consider the following limitations for future work. First, we only focus on single-label fallacy classification. For instance, we only use the dominant types in the dataset MAFALDA as the ground-truth labels. We may later explore the more complex multi-label fallacy classification task. Second, since we only conduct few-shot experiments on instruction-tuned LLMs that are optimized for chat-based applications, it is yet to be explored whether vanilla LLMs could better leverage their in-context learning abilities for the few-shot classification.

References

Tariq Alhindi, Tuhin Chakrabarty, Elena Musi, and Smaranda Muresan. 2023a. Multitask instruction-based prompting for fallacy recognition. *arXiv preprint arXiv:2301.09992*.

Tariq Alhindi, Smaranda Muresan, and Preslav Nakov. 2023b. Large language models are few-shot training example generators: A case study in fallacy recognition. *arXiv preprint arXiv:2311.09552*.

Maciej Besta, Florim Memedi, Zhenyu Zhang, Robert Gerstenberger, Nils Blach, Piotr Nyczyk, Marcin Copik, Grzegorz Kwaśniewski, Jürgen Müller, Lukas Gianinazzi, et al. 2024. Topologies of reasoning: Demystifying chains, trees, and graphs of thoughts. *arXiv preprint arXiv:2401.14295*.

Sina Blassnig, Florin Büchel, Nicole Ernst, and Sven Engesser. 2019. Populism and informal fallacies: An analysis of right-wing populist rhetoric in election campaigns. *Argumentation*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

T Damer. 2008. *Attacking faulty reasoning: A practical guide to fallacy-free arguments*. Nelson Education.

Victor Danciu et al. 2014. Manipulative marketing: persuasion and manipulation of the consumer through advertising. *Theoretical and Applied Economics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gael Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. 2023. Large language models are not strong abstract reasoners. *arXiv preprint arXiv:2305.19555*.

Pierpaolo Goffredo, Mariana Espinoza, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *IJCAI*.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. *arXiv preprint arXiv:1707.06002*.

Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018a. Adapting serious game for fallacious argumentation to german: Pitfalls, insights, and best practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018b. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. *arXiv preprint arXiv:1802.06613*.

Charles Leonard Hamblin. 2022. *Fallacies*. Advanced Reasoning Forum.

Chadi Helwe, Tom Calamai, Pierre-Henri Paris, Chloé Clavel, and Fabian Suchanek. 2023. Mafalda: A benchmark and comprehensive study of fallacy detection and classification. *arXiv preprint arXiv:2311.09761*.

Didin Nuruddin Hidayat, Maya Defianty, Ummi Kultsum, Agus Sufyan, et al. 2020. Logical fallacies in social media: A discourse analysis in political debate. In *2020 8th International Conference on Cyber and IT Service Management (CITSM)*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. *arXiv preprint arXiv:2202.13758*.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*.

Abhinav Lalwani, Lovish Chopra, Christopher Hahn, Caroline Trippel, Zhijing Jin, and Mrinmaya Sachan. 2024. Nl2fol: Translating natural language to first-order logic for logical fallacy detection. *arXiv preprint arXiv:2405.02318*.

Gionnieve Lim and Simon T Perrault. 2024. Evaluation of an llm in identifying logical fallacies: A call for rigor when adopting llms in hci research. *arXiv preprint arXiv:2404.05213*.

Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2021. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Elena Musi, Myrto Aloumpi, Elinor Carmi, Simeon Yates, and Kay O'Halloran. 2022a. Developing fake news immunity: fallacies as misinformation triggers during the pandemic. *Online Journal of Communication and Media Technologies*.

Elena Musi, Myrto Aloumpi, Elinor Carmi, Simeon Yates, and Kay O'Halloran. 2022b. Developing fake news immunity: fallacies as misinformation triggers during the pandemic. *Online Journal of Communication and Media Technologies*.

Elena Musi and Chris Reed. 2022. From fallacies to semi-fake news: Improving the identification of misinformation triggers across digital media. *Discourse & Society*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004.

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*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*.

Ramon Ruiz-Dolz and John Lawrence. 2023. Detecting argumentative fallacies in the wild: Problems and limitations of large language models. In *Proceedings of the 10th Workshop on Argument Mining*.

Saumya Sahai, Oana Balalau, and Roxana Horincar. 2021. Breaking down the invisible wall of informal fallacies in online discussions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. 2020. Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge. *Advances in Neural Information Processing Systems*.

Frans H Van Eemeren, Bart Garssen, and Bert Meuffels. 2009. *Fallacies and judgments of reasonableness: Empirical research concerning the pragma-dialectical discussion rules*. Springer Science & Business Media.

Jacky Visser, John Lawrence, and Chris Reed. 2020. Reason-checking fake news. *Commun. ACM*.

Douglas Walton. 1997. What is propaganda, and what exactly is wrong with it. *Public Affairs Quarterly*.

Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. 2022. Can foundation models talk causality? *arXiv preprint arXiv:2206.10591*.

John Woods. 2004. Who cares about the fallacies? *The Death of Argument: Fallacies in Agent Based Reasoning*.

Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1772–1782, Online.

Xiaobao Wu, Anh Tuan Luu, and Xinshuai Dong. 2022. Mitigating data sparsity for short text topic modeling by topic-semantic contrastive learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2748–2760, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiaobao Wu, Fengjun Pan, and Anh Tuan Luu. 2024a. Towards the TopMost: A topic modeling system toolkit. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41, Bangkok, Thailand. Association for Computational Linguistics.

Xiaobao Wu, Fengjun Pan, Thong Nguyen, Yichao Feng, Chaoqun Liu, Cong-Duy Nguyen, and Anh Tuan Luu. 2024b. On the affinity, rationality, and diversity of hierarchical topic modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Xiaobao Wu, Liangming Pan, William Yang Wang, and Anh Tuan Luu. 2024c. Updating language models with unstructured facts: Towards practical knowledge editing. *arXiv preprint arXiv:2402.18909*.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I Wang, et al. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *arXiv preprint arXiv:2201.05966*.

A Implementation Details

We fine-tune all T5 baselines and run inference with small LLMs (Qwen2.5-Instruct 14B, Qwen2.5-Instruct 7B, Llama3-Chat 8B, Llama2-Chat 13B and Mistral-Instruct 7B) on two RTX A6000 48GB GPUs.

We basically follow the hyperparameter setting used in Alhindi et al. (2023a) for the single/multi-task training of T5-large and T5-3B. We train all T5 models for 5 epochs and select the best performed checkpoints on validation sets as the final models for inference. We use a constant 1e-4 learning rate with warm-up and set batch sizes and gradient accumulation steps with respect to training data size accordingly. We report the results given by the Adafactor optimizer as it performs better than AdamW in our setting. The maximum input length is 1024 so contexts beyond the limit are truncated. The maximum generation length is 64 and the number of beams is 1. The generated output is compared with the ground truth based on strict string matching. Fine-tuning the T5-3B for 5 epochs takes 4 to 6 hours on average, varying with the size of the dataset.

We follow the default generation hyperparameter configurations for inference with all LLMs *e.g.*, sampling is used for generation, with temperature of 0.75 (0.6 for Llama3), top p of 0.9 and top k of 50 *etc.*.

Due to the lack of published preprocessed dataset splits in previous work, we have to adopt different dataset setups tailored to our experiments to ensure that our test splits have acceptable sizes of sufficient statistical power to test for LLMs' performance. In particular, the complete dataset of COVID-19 has become proprietary by the time we start this work. We only have access to a small piece of sample dataset that is publicly available. The results of this dataset may vary if more data are included. We exclude some fallacy classes in ELECDEB and PROPAGANDA following (Helwe et al., 2023) to exclude propaganda-like techniques that do not align well with the definition of "fallacy". As a result of the above changes in setting, our T5 baseline results under similar conditions do not strictly reproduce the results of previous works.

We experimented with two versions of the three-round prompting schemes Premises & Conclusion to thoroughly examine the effects of prompting LLMs with the formal definition of "fallacy" (Helwe et al., 2023). We aim to explore whether explicitly prompting LLMs to reason based on this standard definition could enhance LLMs' understanding of fallacies and translate to improvement on the classification performances. The two versions differ in the position of the definition text and the instruction utterance to cope with any potential ambiguity. Based on our analysis of results derived from both of these two versions, we provide our interpretations of the reasons why this multi-round prompting scheme underperforms.

B Experiment Results

B.1 Zero-shot Single-round Prompting Results

We report detailed experimental results (including accuracy scores and failure rates) of single-round prompting schemes with (**w/**) or without (**w/o**) definitions for each LLM on each dataset compared to the best results of the T5-3B and T5-large baselines.

Setting	Model	Def.	Macro-F1	Acc.	%Failed
Full	T5-3B Multi-ALR	w/	72.38	73.08	0.00
	T5-large Multi-ALR	w/	65.65	67.63	0.00
Zero-shot Single-round	GPT-4	w/	78.94	78.96	0.64
		w/o	78.62	78.53	1.39
	GPT-3.5	w/o	63.59	63.59	1.54
		w/	61.72	62.50	5.82
	Qwen2.5-14Bf	w/	68.19	67.95	2.88
		w/o	67.84	68.40	2.05
	Qwen2.5-7Bf	w/	61.38	63.40	0.00
		w/o	59.59	62.18	0.06
	Llama3-8Bf	w/	48.87	53.53	1.02
		w/o	46.57	52.24	1.34
	Mistral-7Bf	w/o	57.04	60.13	2.88
		w/	50.92	56.60	1.99
	Llama2-13Bf	w/	50.20	55.00	0.13
		w/o	47.39	53.27	0.51

Table 7: Single-round results on ARGOTARIO

Setting	Model	Def.	Macro-F1	Acc.	%Failed
Full	T5-3B Single-task	w/	64.95	70.89	0.33
	T5-large Multi-ALR	w/	59.48	65.22	0.00
Zero-shot Single-round	GPT-4	w/	50.43	59.78	5.89
		w/o	48.45	58.11	9.44
	GPT-3.5	w/	39.65	49.44	11.45
		w/o	38.60	47.78	11.22
	Qwen2.5-14Bf	w/o	41.82	52.07	4.27
		w/	38.36	50.20	4.00
	Qwen2.5-7Bf	w/o	35.48	38.53	4.47
		w/	30.99	39.27	3.13
	Llama3-8Bf	w/o	27.45	40.67	3.27
		w/	27.34	39.80	1.67
	Mistral-7Bf	w/o	28.99	40.47	8.40
		w/	26.07	37.41	5.67
	Llama2-13Bf	w/	25.11	36.27	1.93
		w/o	24.06	32.73	6.60

Table 8: Single-round results on LOGIC

Setting	Model	Def.	Macro-F1	Acc.	%Failed
Full	T5-3B Single-task	w/	83.20	83.30	0.06
	T5-large Multi-ALR	w/	80.42	80.64	0.13
Zero-shot Single-round	GPT-4	w/	79.10	79.73	2.34
		w/o	77.70	78.49	3.70
	GPT-3.5	w/o	70.42	70.76	2.40
		w/	68.90	67.90	2.66
	Qwen2.5-14Bf	w/o	67.72	68.13	3.17
		w/	66.01	66.86	7.02
	Qwen2.5-7Bf	w/	58.58	59.61	0.62
		w/o	58.14	59.96	2.42
	Llama3-8Bf	w/	49.41	52.09	2.53
		w/o	45.21	48.26	4.72
	Mistral-7Bf	w/o	46.89	46.20	8.38
		w/	45.92	45.03	9.01
	Llama2-13Bf	w/o	34.15	38.21	21.52
		w/	34.14	38.87	15.36

Table 9: Single-round results on REDDIT

Setting	Model	Def.	Macro-F1	Acc.	%Failed
Full	T5-3B Single-task	w/	62.37	79.78	0.00
	T5-large Multi-ALEP	w/	56.15	71.78	0.00
Zero-shot Single-round	GPT-4	w/o	42.26	45.56	6.71
		w/	41.93	45.78	9.96
	GPT-3.5	w/	41.01	43.33	3.47
		w/o	37.10	37.11	1.30
	Qwen2.5-14Bf	w/	37.28	41.20	8.70
		w/o	35.76	43.87	12.99
	Qwen2.5-7Bf	w/	43.34	55.33	4.42
		w/o	42.28	53.33	1.82
	Llama3-8Bf	w/o	39.36	63.60	0.26
		w/	36.00	53.60	0.00
	Mistral-7Bf	w/o	33.23	40.93	4.80
		w/	28.01	36.83	4.06
	Llama2-13Bf	w/	35.57	55.07	0.78
		w/o	32.69	59.87	3.12

Table 10: Single-round results on ELECDEB

Setting	Model	Def.	Macro-F1	Acc.	%Failed
Full	T5-3B Multi-ALEP	w/	43.33	76.73	0.25
	T5-large Multi-ALEP	w/	39.75	76.48	0.00
Zero-shot Single-round	GPT-4	w/o	34.80	58.49	2.39
		w/	33.45	56.10	2.39
	GPT-3.5	w/	22.39	38.11	3.90
		w/o	21.84	34.59	4.28
	Qwen2.5-14Bf	w/	21.89	41.13	6.61
		w/o	16.04	29.81	16.23
	Qwen2.5-7Bf	w/	16.03	33.66	1.28
		w/o	15.66	24.53	0.00
	Llama3-8Bf	w/	17.30	28.68	0.15
		w/o	16.03	33.73	1.58
	Mistral-7Bf	w/o	16.89	24.60	10.72
		w/	11.75	19.93	14.19
	Llama2-13Bf	w/	10.61	11.32	3.09
		w/o	4.41	8.91	9.06

Table 11: Single-round results on PROPAGANDA

Setting	Model	Def.	Macro-F1	Acc.	%Failed
Full	T5-3B Multi-ALEP	w/	35.60	31.83	17.50
	T5-large Multi-ALEP	w/	25.60	27.33	7.17
Zero-shot Single-round	GPT-4	w/o	48.74	63.50	1.50
		w/	48.38	65.50	0.67
	GPT-3.5	w/	31.27	44.67	2.50
		w/o	28.97	42.00	2.33
	Qwen2.5-14Bf	w/o	33.03	50.70	0.70
		w/	31.11	49.00	0.88
	Qwen2.5-7Bf	w/o	31.27	41.50	2.80
		w/	30.62	37.00	5.00
	Llama3-8Bf	w/o	24.85	29.10	10.80
		w/	24.43	27.40	12.90
	Mistral-7Bf	w/o	23.44	41.50	4.80
		w/	22.23	36.30	3.50
	Llama2-13Bf	w/	22.09	28.30	27.70
		w/o	17.60	27.40	32.00

Table 12: Single-round results on MAFALDA

Setting	Model	Def.	Macro-F1	Acc.	%Failed
Full	T5-3B Multi-ALEP	w/	14.59	19.05	0.65
	T5-large Multi-ALEP	w/	14.08	18.40	0.65
Zero-shot Single-round	GPT-4	w/	20.47	34.63	0.22
		w/o	16.74	31.17	0.00
	GPT-3.5	w/	17.45	20.35	1.95
		w/o	14.88	20.78	2.17
	Qwen2.5-14Bf	w/	17.60	27.27	1.56
		w/o	15.97	26.10	1.30
	Qwen2.5-7Bf	w/o	15.19	30.52	1.30
		w/	13.48	27.66	1.69
	Llama3-8Bf	w/o	11.00	18.70	0.26
		w/	9.67	16.88	0.00
	Mistral-7Bf	w/	14.69	18.83	2.27
		w/o	10.23	19.35	4.29
	Llama2-13Bf	w/	14.15	15.32	1.69
		w/o	13.09	22.47	18.05

Table 13: Single-round results on Covid-19

Setting	Model	Scheme	Macro-F1	Acc.	%Failed
Full	T5-3B Multi-ALR	WD	72.38	73.08	0.00
	T5-large Multi-ALR	WD	65.65	67.63	0.00
Zero-shot Multi-round	GPT-4	DG	79.87	80.24	0.32
		CoT	76.85	76.60	0.19
	GPT-3.5	GFA	68.40	67.84	2.24
		DG	66.77	68.19	0.40
	Qwen2.5-14Bf	CoT	68.87	70.19	0.00
		GFA	67.63	66.92	3.01
	Qwen2.5-7Bf	DG	60.20	63.20	0.26
		GFA-W	57.71	59.36	2.05
	Llama3-8Bf	P&C^2	61.39	63.59	17.57
		GFA-W	59.25	61.22	16.86
	Mistral-7Bf	GFA-W	57.26	58.97	7.24
		GFA	55.84	56.99	12.88
	Llama2-13Bf	GFA	48.79	54.30	5.90
		DG	45.27	50.51	0.13

Table 14: Multi-round results on Argotario

B.2 Zero-shot Multi-round Prompting Results

We report the detailed experiment results of the best two multi-round prompting schemes for each LLM on each dataset compared to the best results of the T5-3B and T5-large baselines. Notations for multi-round prompting schemes: **DG** for Definition Generation. **GFA** for General Fallacy Analysis. **GFA-W** for General Fallacy Analysis with Warm up. **P&C** for Premises & Conclusion. **CoT** for Zero-shot CoT.

Setting	Model	Scheme	Macro-F1	Acc.	%Failed
Full	T5-3B Single-task	WD	64.95	70.89	0.33
	T5-large Multi-ALR	WD	59.48	65.22	0.00
Zero-shot Multi-round	GPT-4	GFA	50.54	58.78	8.22
		DG	50.03	59.67	9.11
	GPT-3.5	GFA	41.11	51.45	17.11
		GFA-W	39.09	50.89	15.33
	Qwen2.5-14Bf	CoT	45.89	54.00	6.25
		GFA	42.50	50.67	19.53
	Qwen2.5-7Bf	CoT	40.09	47.27	10.87
		GFA-W	36.22	46.13	14.07
	Llama3-8Bf	GFA-W	35.66	47.27	16.40
		GFA	35.18	45.40	15.80
	Mistral-7Bf	GFA-W	31.43	44.13	16.80
		GFA	31.42	44.33	15.47
	Llama2-13Bf	GFA	28.85	39.80	28.00
		GFA-W	28.24	39.80	21.87

Table 15: Multi-round results on Logic

Setting	Model	Scheme	Macro-F1	Acc.	%Failed
Full	T5-3B Single-task	WD	62.37	79.78	0.00
	T5-large Multi-ALEP	WD	56.15	71.78	0.00
Zero-shot Multi-round	GPT-4	DG	41.25	43.55	8.01
		GFA-W	36.27	40.22	25.76
	GPT-3.5	DG	37.77	40.22	2.17
		CoT	34.55	35.67	3.41
	Qwen2.5-14Bf	GFA-W	34.09	42.27	25.20
		GFA	33.82	42.40	28.57
	Qwen2.5-7Bf	DG	44.55	55.87	0.91
		GFA-W	31.20	35.86	17.01
	Llama3-8Bf	CoT	40.81	57.07	2.60
		DG	36.09	56.67	0.26
	Mistral-7Bf	DG	32.91	42.40	4.54
		CoT	27.82	42.53	23.77
	Llama2-13Bf	DG	36.37	49.07	0.78
		GFA-W	29.93	35.07	17.01

Table 16: Multi-round results on ElecDeb

Setting	Model	Scheme	Macro-F1	Acc.	%Failed
Full	T5-3B Multi-ALEP	WD	43.33	76.73	0.25
	T5-large Multi-ALEP	WD	39.75	76.48	0.00
Zero-shot Multi-round	GPT-4	CoT	35.37	52.08	5.66
		DG	33.23	57.23	4.28
	GPT-3.5	CoT	26.67	37.55	3.87
		DG	23.98	40.13	1.89
	Qwen2.5-14Bf	CoT	26.60	46.51	10.38
		GFA-W	22.54	32.23	13.21
	Qwen2.5-7Bf	GFA	19.22	30.79	13.05
		GFA-W	16.88	24.60	13.96
	Llama3-8Bf	CoT	21.35	35.32	5.89
		GFA-W	18.95	35.40	26.34
	Mistral-7Bf	CoT	20.41	40.53	6.42
		DG	15.68	26.34	7.25
	Llama2-13Bf	GFA-W	11.11	12.53	18.94
		GFA	10.13	13.36	35.70

Table 17: Multi-round results on Propaganda

Setting	Model	Scheme	Macro-F1	Acc.	%Failed
Full	T5-3B Multi-ALEP	WD	35.60	31.83	17.50
	T5-large Multi-ALEP	WD	25.60	27.33	7.17
Zero-shot Multi-round	GPT-4	CoT	52.86	58.50	1.83
		GFA-W	48.55	57.50	0.50
	GPT-3.5	GFA-W	40.73	52.33	2.50
		GFA	36.85	44.67	4.17
	Qwen2.5-14Bf	P&C^2	45.94	58.00	0.25
		CoT	43.17	52.25	0.50
	Qwen2.5-7Bf	P&C^1	35.37	52.50	0.30
		GFA	34.98	44.40	0.30
	Llama3-8Bf	P&C^1	34.18	41.33	9.00
		GFA-W	33.71	39.90	14.10
	Mistral-7Bf	P&C^1	29.08	49.50	1.62
		GFA-W	28.55	45.20	2.60
	Llama2-13Bf	GFA-W	15.68	39.70	1.00
		CoT	15.50	17.70	8.70

Table 18: Multi-round results on Mafalda

Setting	Model	Scheme	Macro-F1	Acc.	%Failed
Full	T5-3B Single-task	WD	83.20	83.30	0.06
	T5-large Multi-ALR	WD	80.42	80.64	0.13
Zero-shot Multi-round	GPT-4	CoT	81.11	81.68	3.70
		DG	79.34	80.25	2.08
	GPT-3.5	DG	71.08	71.41	3.70
		CoT	68.64	68.88	2.14
	Qwen2.5-14Bf	CoT	77.08	77.29	1.95
		GFA-W	71.08	71.15	7.29
	Qwen2.5-7Bf	P&C^2	64.58	65.81	2.49
		CoT	64.08	65.58	4.52
	Llama3-8Bf	P&C^1	57.83	58.64	19.77
		GFA-W	57.45	57.66	20.00
	Mistral-7Bf	CoT	59.70	60.82	12.52
		DG	49.27	48.62	4.44
	Llama2-13Bf	GFA	45.82	46.70	22.96
		P&C^2	44.49	45.11	11.70

Table 19: Multi-round results on REDDIT

Setting	Model	Scheme	Macro-F1	Acc.	%Failed
Full	T5-3B Multi-ALEP	WD	14.59	19.05	0.65
	T5-large Multi-ALEP	WD	14.08	18.40	0.65
Zero-shot Multi-round	GPT-4	CoT	25.18	33.12	0.00
		GFA-W	23.59	40.69	1.73
	GPT-3.5	GFA	17.24	30.74	5.41
		GFA-W	16.04	29.22	2.38
	Qwen2.5-14Bf	GFA	23.73	30.00	4.67
		CoT	23.31	34.09	0.00
	Qwen2.5-7Bf	GFA	22.88	33.77	0.39
		CoT	19.71	31.69	0.00
	Llama3-8Bf	CoT	19.83	28.44	0.13
		GFA	16.20	19.22	30.00
	Mistral-7Bf	CoT	18.53	26.62	0.78
		P&C^1	16.41	27.92	5.58
	Llama2-13Bf	DG	14.16	30.26	0.00
		P&C^1	12.76	29.22	3.64

Table 20: Multi-round results on COVID-19

B.3 Prompting Scheme Rankings

Data Domain	LLM	Scheme	#R	Rank	%Failed
Open	GPT-4	Zero-shot CoT	2	2.50	3.43
		With Definitions	1	2.75	2.38
	GPT-3.5	General Fallacy Analysis	2	2.50	9.40
		Definition Generation	2	2.75	3.51
	Qwen2.5-7/14Bf	Zero-shot CoT	2	2.25	4.13
		General Fallacy Analysis with Warm Up	3	3.50	7.06
	Llama3-8Bf	Premises & Conclusion	3	2.62	15.58
		General Fallacy Analysis with Warm Up	3	2.80	16.81
	Mistral-7Bf	General Fallacy Analysis with Warm Up	3	2.25	12.76
		Zero-shot CoT	2	2.67	9.13
	Llama2-13Bf	With Definitions	1	4.83	7.84
		Definition Generation	2	5.50	1.43
Specific	GPT-4	Zero-shot CoT	2	2.33	13.36
		With Definitions	1	3.33	4.19
	GPT-3.5	With Definitions	1	1.67	3.11
		Definition Generation	2	3.00	2.22
	Qwen2.5-7/14Bf	General Fallacy Analysis	2	2.83	14.22
		General Fallacy Analysis with Warm Up	3	3.00	12.10
	Llama3-8Bf	Zero-shot CoT	2	1.00	2.87
		General Fallacy Analysis with Warm Up	3	3.67	18.74
	Mistral-7Bf	Zero-shot CoT	2	2.00	10.32
		Without Definitions	1	3.67	6.60
	Llama2-13Bf	With Definitions	1	2.00	1.85
		Definition Generation	2	2.67	0.76

Table 21: The best two prompting schemes for different LLMs and data domains based on average Macro-F1 rankings. **#R**: Number of rounds. **Rank**: Average ranking on Macro-F1, the lower the better.

Domain	Task	Scheme	#R	Rank	%Failed
Open	ARGOTARIO	Definition Generation	2	5.40	0.30
		With Definitions	1	5.60	1.44
	LOGIC	General Fallacy Analysis	2	1.71	17.60
		General Fallacy Analysis with Warm Up	3	2.57	15.13
	MAFALDA	General Fallacy Analysis with Warm Up	3	2.57	3.21
		General Fallacy Analysis	2	3.67	5.12
	REDDIT	Zero-shot CoT	2	3.00	4.98
		General Fallacy Analysis with Warm Up	3	3.71	15.03
Specific	COVID-19	General Fallacy Analysis	2	2.86	6.99
		Zero-shot CoT	2	3.00	0.18
	ELECDEB	Without Definitions	1	2.14	4.43
		With Definitions	1	2.14	4.48
	PROPAGANDA	Zero-shot CoT	2	2.43	6.87
		General Fallacy Analysis with Warm Up	3	3.00	19.14

Table 22: The best two prompting schemes for each benchmark dataset based on average Macro-F1 rankings. **#R**: Number of rounds. **Rank**: Average ranking on Macro-F1, the lower the better.

B.4 Few-shot vs. Zero-shot Results

Model	Shot	Logic	ElecDeb	Propaganda
GPT-3.5	One	37.13	31.41	12.23
	Two	34.55	24.07	12.08
	Zero°	38.6	37.1	21.84
	Zero*	**41.11**^{GFA}	**41.01**^{WD}	**26.67**^{CoT}
Qwen2.5-14Bf	One	43.08	27.91	20.54
	Two	44.47	31.07	20.73
	Zero°	41.82	35.76	16.04
	Zero*	**45.89**^{CoT}	**37.28**^{WD}	**26.6**^{CoT}
Qwen2.5-7Bf	One	33.22	35.53	13.67
	Two	28.62	31.69	15.58
	Zero°	35.48	42.28	15.66
	Zero*	**40.09**^{CoT}	**44.55**^{DG}	**18.95**^{GFA}
Llama3-8Bf	One	22.94	27.2	7.55
	Two	24.64	24.16	6.32
	Zero°	27.45	39.36	16.03
	Zero*	**35.66**$^{\text{GFA-W}}$	**40.81**^{CoT}	**21.35**^{CoT}
Mistral-7Bf	One	30.08	25.32	11.61
	Two	27.04	24.95	8.55
	Zero°	28.99	33.23	16.89
	Zero*	**31.43**$^{\text{GFA-W}}$	**33.23**^{WoD}	**20.41**^{CoT}
Llama2-13Bf	One	1.76	3.73	1.47
	Two	1.74	7.22	3.15
	Zero°	24.06	32.69	4.41
	Zero*	**28.85**^{GFA}	**36.37**^{DG}	**11.11**$^{\text{GFA-W}}$

Table 23: Few-shot vs. zero-shot results. °: Zero-shot single-round results without definitions. *: Best zero-shot results.

Model	Argotario	Logic	Reddit	ElecDebate	Propaganda	Mafalda	Covid-19	Avg.
GPT-4	0.93	0.11	2.01	-1.01	0.57	4.12	4.71	**1.63**
GPT-3.5	4.81	1.46	0.66	-3.24	4.28	9.46	-0.21	**2.46**
Qwen2.5-Instruct 14B	0.68	4.07	9.36	-3.19	4.71	12.91	6.13	**4.95**
Qwen2.5-Instruct 7B	-1.18	4.61	6.00	1.21	3.19	4.10	7.69	**3.66**
Llama3-Chat 8B	12.52	8.21	8.42	1.45	4.05	9.33	8.83	**7.54**
Mistral-Instruct 7B	0.22	2.44	12.81	-0.32	3.52	5.64	3.84	**4.02**
Llama2-Chat 13B	-1.41	3.74	11.67	0.80	0.50	-6.41	0.01	**1.27**

Table 24: Performance improvements of the best performed multi-round prompting schemes compared to the best performed single-round schemes. The average performance gains across datasets for each LLM are in **bold**.

Model	Argotario	Logic	Reddit	ElecDeb	Propaganda	Mafalda	Covid-19
GPT-4	**79.87**^{DG} 79.87^{DG}	**50.54**^{GFA} 50.03^{DG}	**81.11**^{CoT} 79.34^{DG}	**41.25**^{DG} 41.93^{WD}	**35.37**^{CoT} 33.45^{WD}	**52.86**^{CoT} 45.85^{DG}	**25.18**^{CoT} 20.47^{WD}
GPT-3.5	**68.4**^{GFA} 66.77^{DG}	**41.11**^{GFA} 37.2^{DG}	**71.08**^{DG} 71.08^{DG}	**37.77**^{DG} 41.01^{WD}	**26.67**^{CoT} 22.39^{WD}	**40.73**$^{\text{GFA-W}}$ 35.25^{DG}	**17.24**^{GFA} 17.45^{WD}
Qwen2.5-14Bf	**68.87**^{CoT} 66.91$^{\text{GFA-W}}$	**45.89**^{CoT} 41.64$^{\text{GFA-W}}$	**77.08**^{CoT} 71.08$^{\text{GFA-W}}$	**34.09**$^{\text{GFA-W}}$ 32.31^{CoT}	**26.6**^{CoT} 26.6^{CoT}	**45.94**$^{\text{P\&C}^2}$ 40.53$^{\text{GFA-W}}$	**23.73**^{GFA} 23.31^{CoT}
Qwen2.5-7Bf	**60.2**^{DG} 57.71$^{\text{GFA-W}}$	**40.09**^{CoT} 36.22$^{\text{GFA-W}}$	**64.58**$^{\text{P\&C}^2}$ 61.69$^{\text{GFA-W}}$	**44.55**^{DG} 30.89^{CoT}	**19.22**^{GFA} 15.57^{CoT}	**35.37**$^{\text{P\&C}^1}$ 33.2$^{\text{GFA-W}}$	**22.88**^{GFA} 19.71^{CoT}
Llama3-8Bf	**61.39**$^{\text{P\&C}^2}$ 59.25$^{\text{GFA-W}}$	**35.66**$^{\text{GFA-W}}$ 35.66$^{\text{GFA-W}}$	**57.83**$^{\text{P\&C}^1}$ 57.45$^{\text{GFA-W}}$	**40.81**^{CoT} 40.81^{CoT}	**21.35**^{CoT} 21.35^{CoT}	**34.18**$^{\text{P\&C}^1}$ 33.71$^{\text{GFA-W}}$	**19.83**^{CoT} 19.83^{CoT}
Mistral-7Bf	**57.26**$^{\text{GFA-W}}$ 57.26$^{\text{GFA-W}}$	**31.43**$^{\text{GFA-W}}$ 31.43$^{\text{GFA-W}}$	**59.7**^{CoT} 44.84$^{\text{GFA-W}}$	**32.91**^{DG} 27.82^{CoT}	**20.41**^{CoT} 20.41^{CoT}	**29.08**$^{\text{P\&C}^1}$ 28.55$^{\text{GFA-W}}$	**18.53**^{CoT} 18.53^{CoT}
Llama2-13Bf	**48.79**^{GFA} 44.42$^{\text{GFA-W}}$	**28.85**^{GFA} 28.24$^{\text{GFA-W}}$	**45.82**^{GFA} 42.68$^{\text{GFA-W}}$	**36.37**^{DG} 27.44^{CoT}	**11.11**$^{\text{GFA-W}}$ 7.72^{CoT}	**15.68**$^{\text{GFA-W}}$ 15.68$^{\text{GFA-W}}$	**14.16**^{DG} 8.82^{CoT}

Table 25: Best zero-shot results vs. results given by the general domain-model guidance based on prompt scheme rankings. The best zero-shot results are in **bold**.

Fallacy Type	ARGOTARIO	LOGIC	REDDIT	MAFALDA	ELECDEB	PROPAGANDA	COVID-19
Appeal to False Authority	40		64	6	22	10	10
Ad Hominem	42	41		7	21		
Appeal to Emotion	69	23		5	96		
Red Herring	37	24				3	9
Hasty Generalization	53		60	20			7
Ad Populum		30	59	12		1	
False Dilemma		12	63	6		10	
False Causality (Post Hoc Fallacy)		18		12	6		19
Equivocation		5		4		1	7
Straw Man		21		2		2	5
Slippery Slope			69	9	5		
Circular Reasoning		19		8			
Appeal to Nature			63	9			
Appeal to Worse Problems			72	5			
Doubt Credibility		17				47	
Fallacy of Converse (Affirming the Consequent)		14					
Appeal to Fear				5		10	
Causal Oversimplification				16		19	
Faulty Generalization		61					
Intentional (Intentionally Wrong Argument)		15					
Appeal to Tradition			63				
Appeal to Ridicule				7			
Fallacy of Division				2			
False Analogy				5			8
Guilt by Association				3			
Tu Quoque				3			
Flag-Waving						21	
Name-calling						120	
Reductio Ad Hitlerum						5	
Whataboutism						5	
Cherry Picking							13
Evading the Burden of Proof							14

Table 26: Distribution of included fallacies on each dataset.

Fallacy	**Definition**
Hasty Generalization	a fallacy when someone makes generalizations based on incomplete observations that cannot represent or generalize to other situations if other relevant factors are taken into account.
Red Herring	a fallacy when someone introduces irrelevant or confusing information in arguments to diverge attention from the main topic being discussed to irrelevant issues.
Circular Reasoning	(LOGIC)a fallacy when the conclusion of an argument is a restatement of the assumption, or the argument assumes the very thing it is trying to prove. (MAFALDA)This fallacy occurs when an argument assumes the very thing it is trying to prove, resulting in a circular and logically invalid argument.
Causal Oversimplification	(PROPAGANDA)a fallacy that assumes only a single cause or reason for an issue when there were actually multiple ones. (MAFALDA)This fallacy occurs when a complex issue is reduced to a single cause and effect, oversimplifying the actual relationships between events or factors.
Doubt Credibility	a propaganda technique that attacks or questions the credibility of someone or something in order to discredit the opponent's argument.
Appeal to False Authority	(ARGOTARIO, ELECDEB, COVID-19,PROPAGANDA) a fallacy when someone attempts to argue or persuade by referring to the opinions or statements of a questionable authority who lacks sufficient credibility in the discussed matter because the authority's expertise may be inadequate/irrelevant or the authority is attributed a statement that has been tweaked. (MAFALDA)This fallacy occurs when an argument relies on the opinion or endorsement of an authority figure who may not have relevant expertise or whose expertise is questionable. When applicable, a scientific consensus is not an appeal to authority.
False Causality (Post Hoc Fallacy)	a fallacy that incorrectly assumes that one event causes another solely based on the observation of a temporal order or correlation that one event came before the other, rather than a proven causal relationship.

Continued on next page

Table 27: Fallacy definitions

Ad Hominem	(Argotario)a fallacy when someone attacks the others' characters or motives instead of addressing the substance of their arguments. (Elecdeb, Logic)a fallacy when the arguer directly attacks the opponent's characters, positions or motives instead of addressing the substance of their arguments. (Mafalda)This fallacy involves attacking a person's character or motives instead of addressing the substance of their argument.
Appeal to Emotion	a fallacy when the arguer attempts to argue or persuade by using emotive language to arouse non-rational sentiments within the intended audience.
Ad Populum	(Logic,Reddit,Propaganda)a fallacy when the arguer claims that an idea or action is true, valid, correct, or better simply because it is popular or widely accepted by the majority of people, or something is unreal, invalid or bad because few people believe in it. (Mafalda)This fallacy involves claiming that an idea or action is valid because it is popular or widely accepted.
False Dilemma	(Logic,Reddit,Propaganda)a fallacy that forces a conclusion by presenting or implying an incomplete list (usually two) of options or sides, even though in fact there are more that can be chosen from. (Mafalda)This fallacy occurs when only two options are presented in an argument, even though more options may exist.
Equivocation	(Logic)a fallacy when an argument uses ambiguous language or changing the meaning of a term from time to time in an attempt to confuse or obfuscate. (Propaganda, COVID-19)a fallacy that uses confused, unclear or intenationally vague languages to disguise the argument. (Mafalda)This fallacy involves using ambiguous language or changing the meaning of a term within an argument, leading to confusion and false conclusions.
Straw Man	(Logic, Propaganda, COVID-19)a fallacy when the arguer substitutes an opponent's argument with a distorted, exaggerated or misinterpreted version to make it more easily to be attacked or discredited. (Mafalda)This fallacy involves misrepresenting an opponent's argument, making it easier to attack and discredit.
Slippery Slope	(Reddit, Elecdeb)a fallacy when the arguer claims that a small or insignificant action step will inevitably lead to a chain of events that result in significant negative consequences, while the connection between such events or steps is unwarranted or improbable. (Mafalda)This fallacy occurs when it is claimed that a small step will inevitably lead to a chain of events, resulting in a significant negative outcome.
Appeal to Nature	a fallacy that occurs when something is assumed to be good or desirable simply because it is natural, while its unnatural counterpart is assumed to be bad or undesirable.
Appeal to Worse Problems	This fallacy involves dismissing an issue or problem by claiming that there are more important issues to deal with, instead of addressing the argument at hand. This fallacy is also known as the "relative privation" fallacy.
Fallacy of Converse (Affirming the Consequent)	a fallacy that takes a true conditional statement under certain assumptions but invalidly infers its converse even though the conversed statement may not be true under the same assumptions.
Appeal to Fear or Prejudice	a fallacy that uses fear, not based on evidence or reason, as the primary motivator to get others to accept an idea, proposition, or conclusion.
Faulty Generalization	a fallacy when someone makes generalizations based on incomplete/partial observations that cannot represent or generalize to all the possible situations (or the whole population) if other relevant factors are taken into account.
Intentional (Intentionally Wrong Argument)	a fallacy that uses some intentional or subconscious actions or choices to incorrectly support an argument.
Appeal to Tradition	(Reddit)a fallacy that involves arguing that something should continue to be done a certain way because it has always been done that way, rather than evaluating its merits. (Mafalda)This fallacy involves arguing that something should continue to be done a certain way because it has always been done that way, rather than evaluating its merits.
Appeal to Ridicule	This fallacy occurs when an opponent's argument is portrayed as absurd or ridiculous with the intention of discrediting it.
Fallacy of Division	This fallacy involves assuming that if something is true for a whole, it must also be true of all or some of its parts.

Continued on next page

Table 27: Fallacy definitions (Continued)

False Analogy	a fallacy that involves making an analogy between two elements based on superficial resemblance.
Guilt by association	This fallacy involves discrediting an idea or person based on their association with another person, group, or idea that is viewed negatively.
Tu Quoque	This fallacy occurs when someone's argument is dismissed because they are accused of acting inconsistently with their claim, rather than addressing the argument itself.
Flag-Waving	a fallacy that plays on strong feeling (e.g. patriotism or nationalism) for a group/nation/country to justify or promote an action or idea.
Name-calling	a fallacy that labels the object of campaign as either something the target audience fears, hates, finds undesirable or loves, praises.
Reductio Ad Hitlerum	a fallacy that persuades an audience to disapprove an action or idea by suggesting that the idea is popular with groups hated in contempt (e.g., Hitler or the Nazi party) by the target audience.
Whataboutism	a fallacy that responses to an accusation of wrongdoing by claiming that an offense committed by another is similar or worse, attempting to discredit the allegation.
Cherry Picking	a fallacy that chooses among competing evidence only for those that support a given position or argument, ignoring or dismissing findings which do not support it.
Evading the Burden of Proof	a fallacy that persuades an argument without any support of evidence as if it was self-evident

Table 27: Fallacy definitions

Fallacy	Formal Definition
Ad Hominem	E claims P. E's character is attacked (A). Therefore, negate P.
Tu Quoque	E claims P, but E is acting as if E negates P. Therefore, negate P.
Guilt by Association	E1 claims P. Also, E2 claims P, and E2's character is attacked (A). Therefore, negate P. OR, E1 claims P. E2's character is attacked (A) and is similar to E1. Therefore, negate P.
Ad Populum	A lot of people believe/do P. Therefore, P. OR only a few people believe/do P. Therefore, negate P.
Appeal to Nature	P1 is natural. P2 is not natural. Therefore, P1 is better than P2. OR P1 is natural, therefore P1 is good.
Appeal to Tradition	We have been doing P for generations. Therefore, we should keep doing P. OR our ancestors thought P. Therefore, P.
Appeal to False Authority	E claims P (when E is seen as an authority on the facts relevant to P). Therefore, P.
Causal Oversimplification	P1 caused C (although P2, P3, P4, etc. also contributed to C).
Hasty Generalization	Sample E1 is taken from population E. (Sample E1 is a very small part of population E.) Conclusion C is drawn from sample E1.
False Causality	P is associated with C (when the link is mostly temporal and not logical). Therefore, P causes C.
False Analogy	E1 is like E2. E2 has property P. Therefore, E1 has property P. (but E1 really is not too much like E2)
False Dilemma	Either P1 or P2, while there are other possibilities. OR either P1, P2, or P3, while there are other possibilities.
Slippery Slope	P1 implies P2, then P2 implies P3,... then C which is negative. Therefore, negate P1.
Fallacy of Division	E1 is part of E, E has property P. Therefore, E1 has property P.
Straw Man	E1 claims P. E2 restates E1's claim (in a distorted way P'). E2 attacks (A) P'. Therefore, negate P.
Circular Reasoning	C because of P. P because of C. OR C because C.
Equivocation	No logical form: P1 uses a term T that has a meaning M1. P2 uses the term T with the meaning M2 to mislead.
Appeal to Positive Emotion	P is positive. Therefore, P.
Appeal to Anger	E claims P. E is outraged. Therefore, P. Or E1 claims P. E2 is outraged by P. Therefore, P (or negate P depending on the situation).

Continued on next page

Table 28: Formal fallacy definitions for MAFALDA

Appeal to Fear	If negate P1, something terrible P2 will happen. Therefore, P1.
Appeal to Pity	P which is pitiful, therefore C, with only a superficial link between P and C.
Appeal to Ridicule	E1 claims P. E2 makes P look ridiculous, by misrepresenting P (P'). Therefore, negate P.
Appeal to Worse Problems	P1 is presented. P2 is presented as a best-case. Therefore, P1 is not that good. OR P1 is presented. P2 is presented as a worst-case. Therefore, P1 is very good.

Table 28: Formal fallacy definitions for MAFALDA

C Prompt Templates

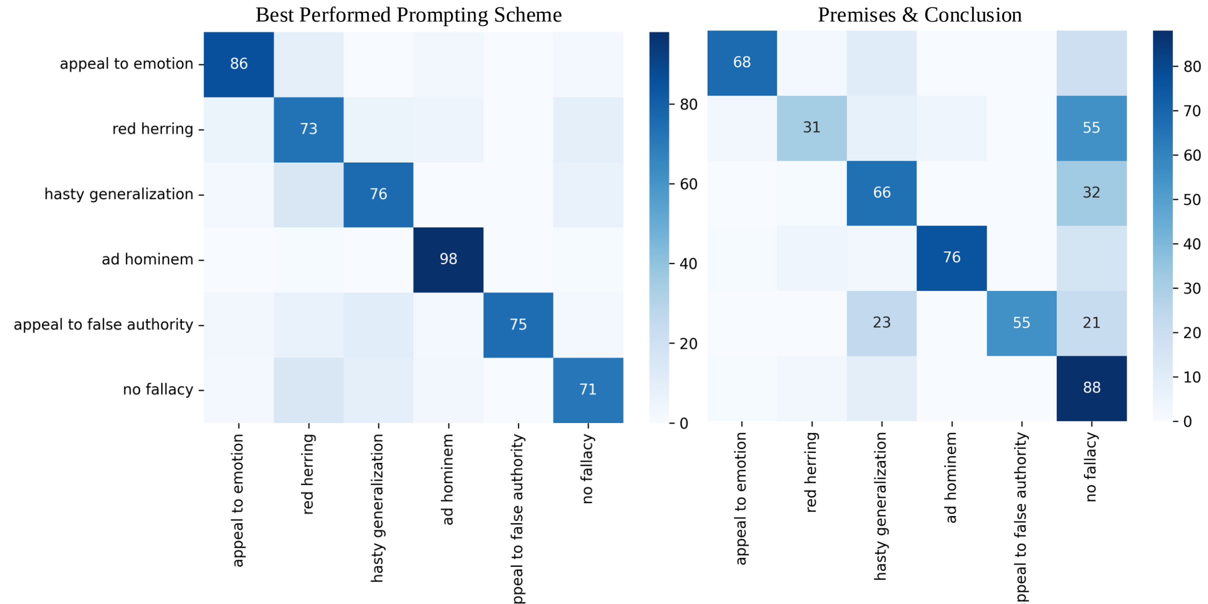
Scheme	Round	Prompt Template
WD	Single	Given [#Classes] types of fallacies, namely, [Fallacy List], and <Discourse Type> below, determine whether any / which of the fallacies given is present in <Discourse Segment to be Classified>? [Discourse] Output your answer in JSON format {"fallacy": name_of_the_fallacy, "explanation": in_a_sentence_or_two.} (If none of the fallacies is found, output {"fallacy": "No Fallacy", "explanation": in_a_sentence_or_two}). Only output JSON.
WD	Single	Based on the following definitions of fallacies, [Fallacy Definitions], given <Discourse Type> below, determine whether any / which of the fallacies defined above is present in <Discourse Segment to be Classified>? [Discourse] Output your answer in JSON format {"fallacy": name_of_the_fallacy, "explanation": in_a_sentence_or_two.} (If none of the fallacies is found, output {"fallacy": "No Fallacy", "explanation": in_a_sentence_or_two}). Only output JSON.
CoT	R1	Given the following <Discourse Type>, [Discourse] and the following [#Fallacy] types of fallacies, namely, [Fallacy List], which of the listed fallacies is present in <Discourse Segment to be Classified>? Now, let's think step by step.
	R2	Output your previous conclusion in JSON format {"fallacy": name_of_the_fallacy. Only output JSON.}
DG	R1	Give a definition to each of the following [#Fallacy] types of fallacies in [Fallacy List].
	R2	Based on the definitions you provided, given the <Discourse Type> below, determine which of the defined fallacies is present in <Discourse Segment to be Classified>? <Discourse Type>: [Discourse] Output your answer in JSON format {"fallacy": name_of_the_fallacy. Only output JSON.}
GFA	R1	Given the following <Discourse Type>, is the <Discourse Segment to be Classified> logically reasonable or potentially fallacious? <Discourse Type>: [Discourse] Give your analysis.
	R2	According to your previous analysis, considering [#Fallacy] of fallacies [Fallacy List], determine which of these listed fallacies is present in the <Discourse Segment to be Classified>? Output your answer in JSON format {"fallacy": name_of_the_fallacy. Only output JSON.}
GFA-W	R1	(LOGIC) Extract and summarize the focal argument in the following segment of discourse: [Discourse]
	R2	Is <Discourse Segment to be Classified> logically reasonable or potentially fallacious? Give your analysis.
	R3	According to your previous analysis, considering [#Fallacy] of fallacies [Fallacy List], determine which of these listed fallacies is present in the <Discourse Segment to be Classified>? Output your answer in JSON format {"fallacy": name_of_the_fallacy. Only output JSON.}
P&C^1	R1	Given the following <Discourse Type> [Discourse] extract and summarize <Discourse Argument> by pointing out the premise(s) and conclusion of the argument.
	R2	A fallacy is defined as an argument where the premises do not entail the conclusion. According to this definition, is <Discourse Segment to be Classified> logically reasonable or potentially fallacious? Give your analysis.
	R3	According to your previous analysis, considering [#Fallacy] of fallacies [Fallacy List], determine which of these listed fallacies is present in the <Discourse Segment to be Classified>? Output your answer in JSON format {"fallacy": name_of_the_fallacy. Only output JSON.}
P&C^2	R1	Given the following <Discourse Type> [Discourse] extract and summarize <Discourse Argument> by pointing out the premise(s) and conclusion of the argument.
	R2	Whether or not the premise(s) of <Discourse Argument> entail the conclusion? Give your analysis.
	R3	A fallacy is defined as an argument where the premises do not entail the conclusion. According to your previous analysis, considering [#Fallacy] of fallacies [Fallacy List], determine which of these listed fallacies is present in the <Discourse Segment to be Classified>? Output your answer in JSON format {"fallacy": name_of_the_fallacy. Only output JSON.}

Table 29: Sample templates of our proposed single-round and multi-round prompting schemes. Notations: **DG** for Definition Generation. **GFA** for General Fallacy Analysis. **GFA-W** for General Fallacy Analysis with Warm up. **P&C** for Premises & Conclusion. **CoT** for Zero-shot CoT.

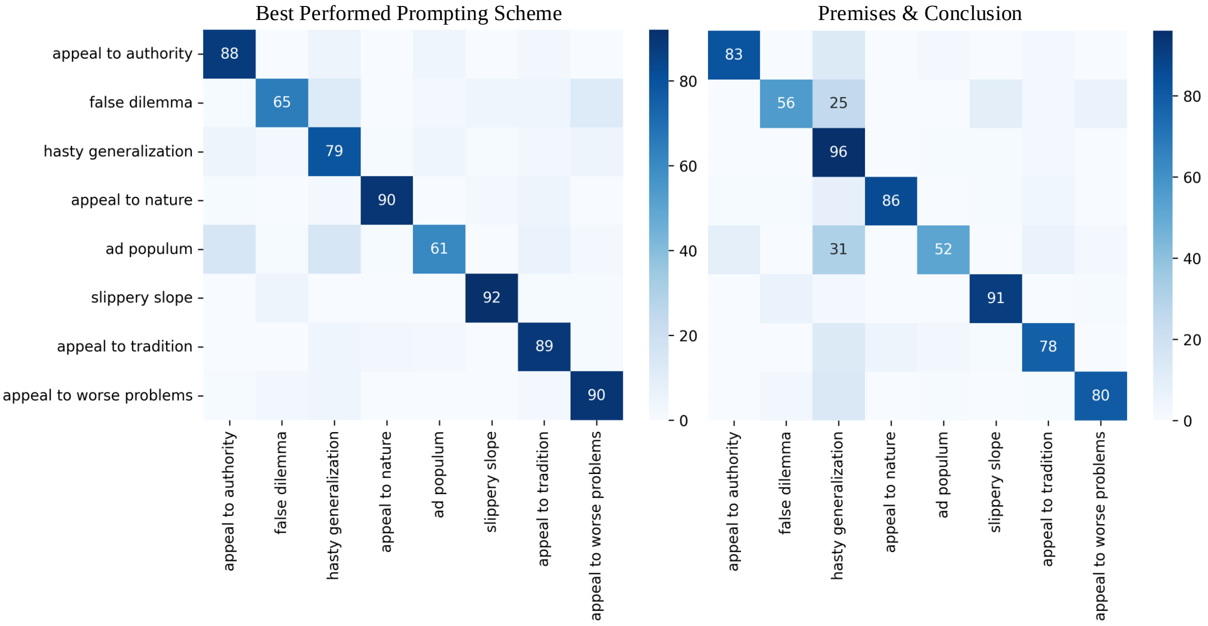
D Error Analysis

D.1 Performance Analysis of Premises & Conclusion

We compare the confusion matrix of GPT-4 under the top three best performed prompting schemes and the confusion matrix of GPT-4 under the multi-round prompting scheme of Premises & Conclusion on each dataset. We find that, under the Premises & Conclusion scheme, GPT-4 is significantly biased to over predict “No Fallacy” when this label is given or fallacy types that contain words related to the semantics of “causality”, such as“causal” and “generalization”. We speculate that the prompt formulation of Premises & Conclusion by offering the formal definition of the term fallacy “*whether the premises entail the conclusion*” has a misleading implication that renders GPT-4 to place excessive attention on fallacy conceptions with semantics that overlap with the word “entail”. This is why Premises & Conclusion prompting scheme underperforms the other prompting methods.

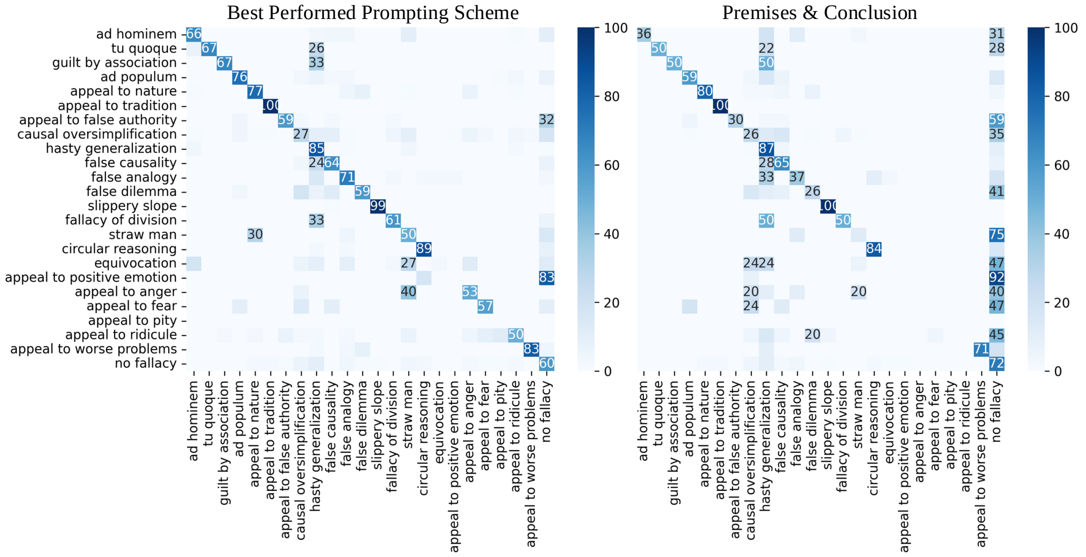


(a) Argotario

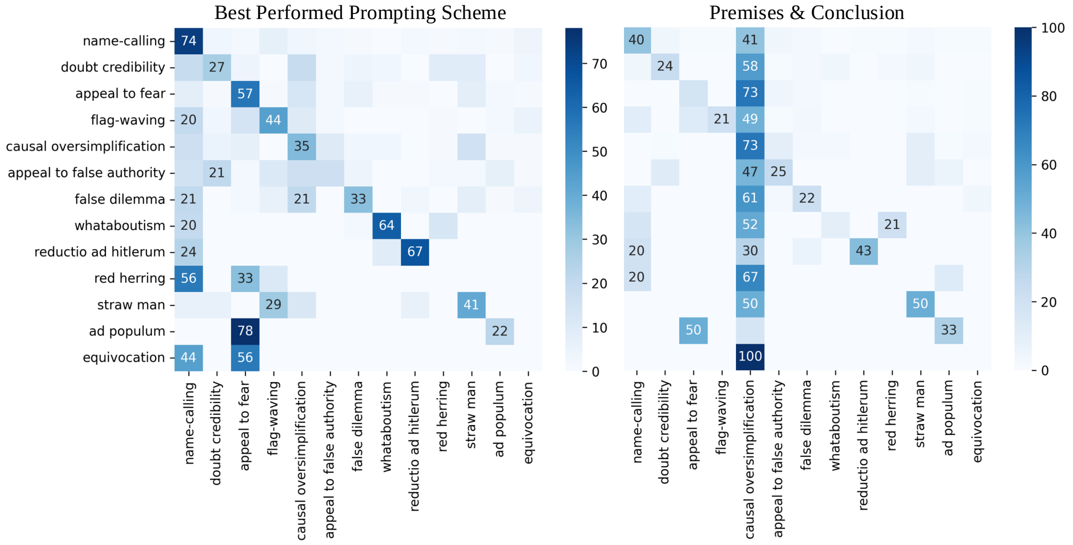


(b) Reddit

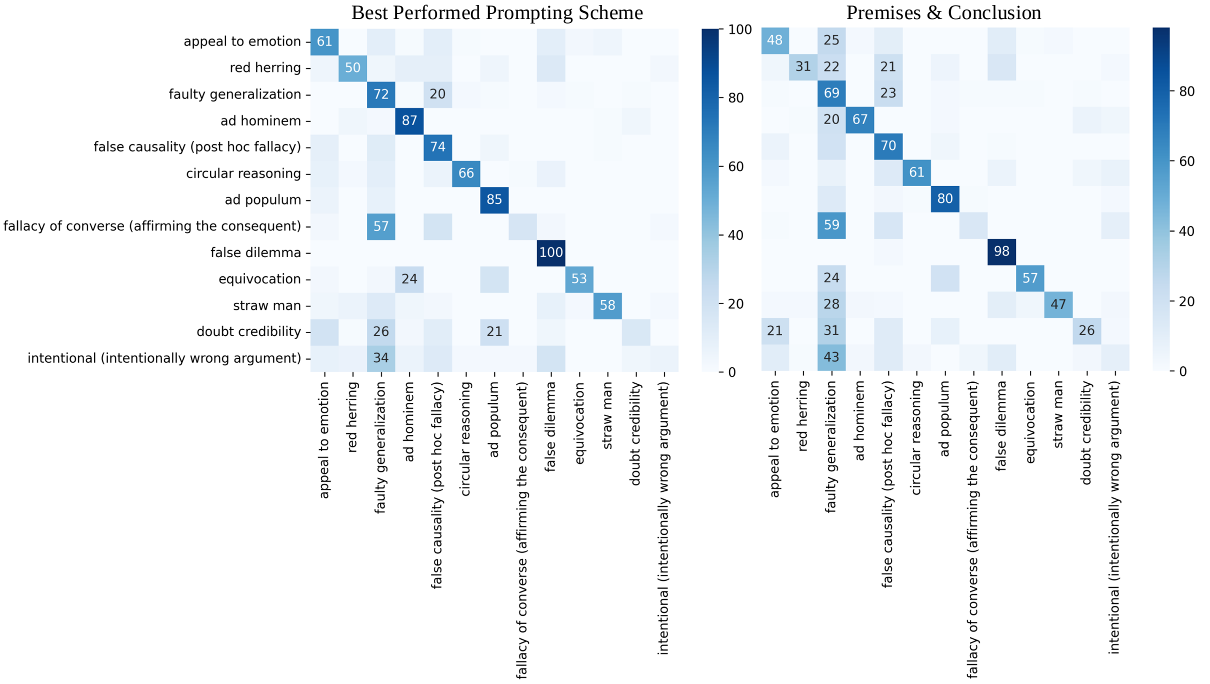
Figure 4: Comparison of GPT-4’s results under the best performed prompting scheme and the multi-round Premise & Conclusion on ARGOTARIO and REDDIT.



(c) Mafalda

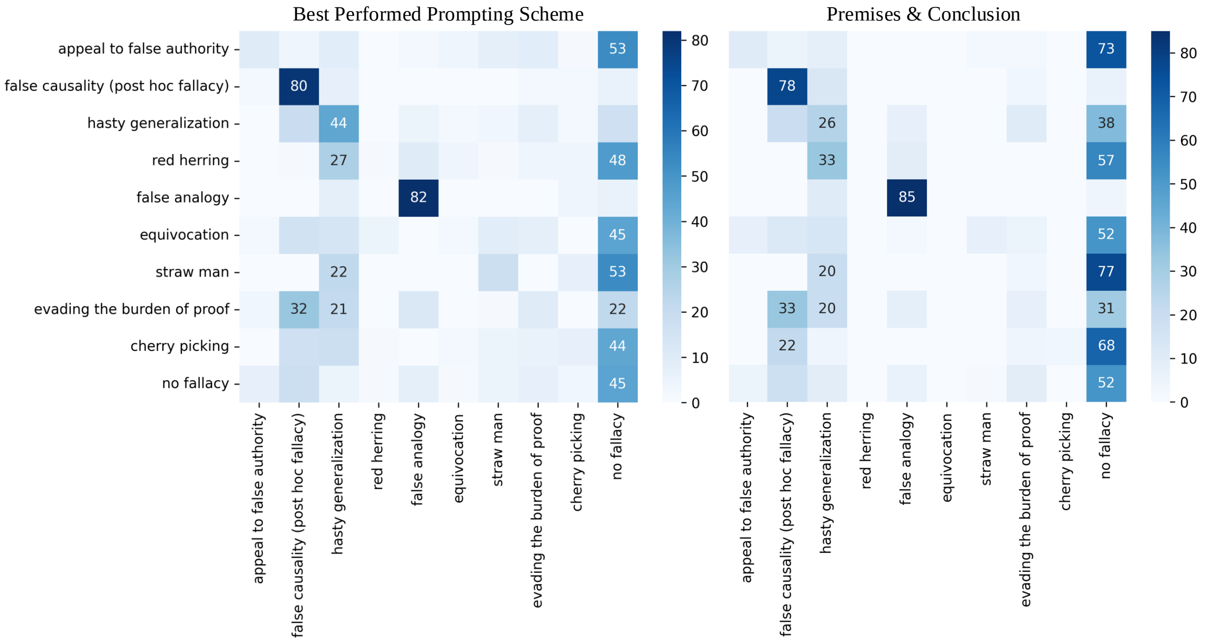


(d) Propaganda

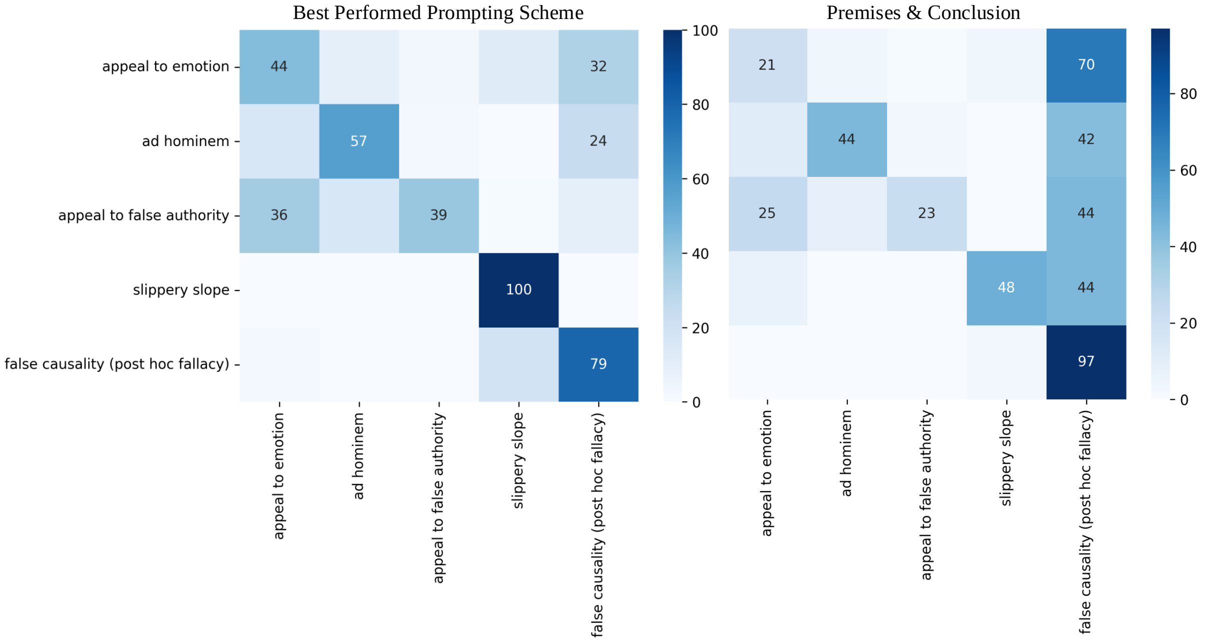


(e) Logic

Figure 5: Comparison of GPT-4's results under the best performed prompting scheme and the multi-round Premise & Conclusion on MAFALDA, PROPAGANDA and LOGIC.



(f) Covid-19



(g) ElecDeb

Figure 6: Comparison of GPT-4's results under the best performed prompting scheme and the multi-round Premise & Conclusion on COVID-19 and ELECDEB.

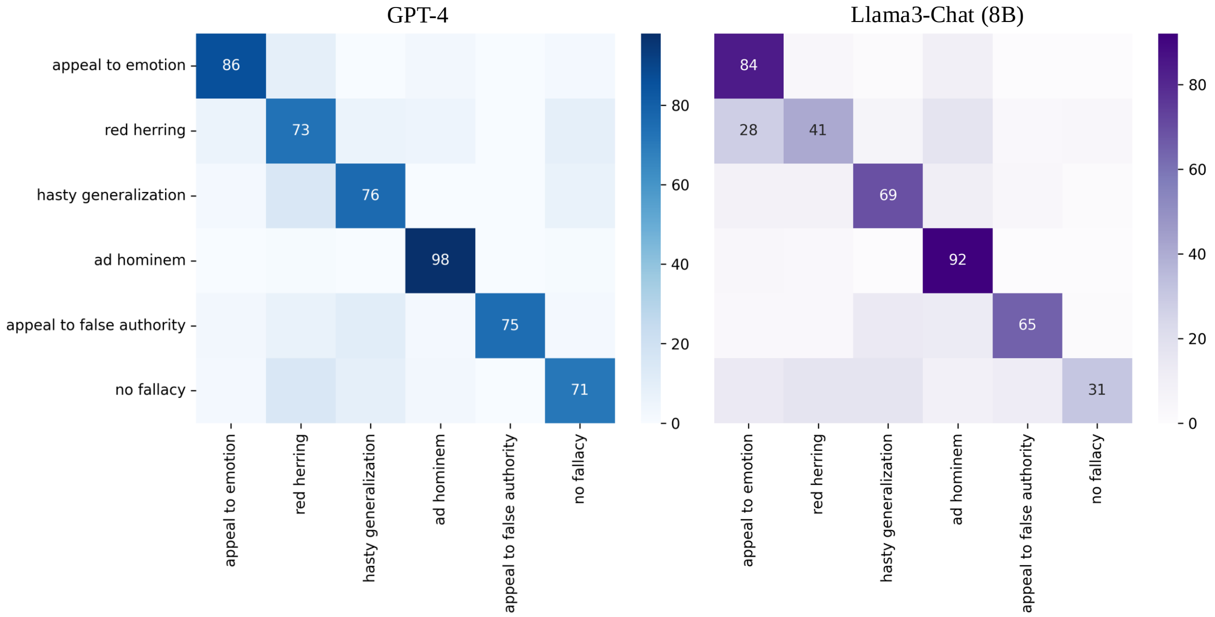
D.2 Error Cases

Scheme	Model		
DG	GPT-4	**Input**	A: Do we need a global environment task force? B: The global taskforce for human rights is doing a good job. That's why we need one for the environment, too.
		Output	Hasty Generalization: A conclusion based on insufficient or biased evidence; rushing to a conclusion before you have all relevant facts.
		Pred.	No Fallacy
		GT	Hasty Generalization
	Mistral	**Input**	<YouTube conspiracy theorist> Jake Morphonios said that for all he knows, Doug Haig didn't play a role in Paddock's massacre.
		Output	Red Herring: Introducing a distracting or irrelevant topic, so as to lead others away from the main subject at hand, and avoid addressing the actual argument. Name-calling Using derogatory or offensive terms to describe someone or their arguments, instead of presenting a counterargument.
		Pred.	Red Herring
		GT	Name-calling
GFA	GPT-3.5	**Input**	something that kills is deadly.
		Output	The focal argument of the discourse is logically reasonable. The statement "something that kills is deadly" is a tautology, as it simply states that if something has the capability to kill, it is indeed deadly. This argument is logically sound and does not contain any fallacies.
		Pred.	No Fallacy
		GT	Circular Reasoning
	Llama-2	**Input**	If you are outraged by this, stop eating meat and dairy. Otherwise, you are a hypocrite
		Output	The comment's argument is potentially fallacious. The argument is an example of the "hasty generalization" fallacy, where a conclusion is drawn based on insufficient or incomplete evidence. The comment assumes that anyone who is outraged by the topic must be a hypocrite if they consume meat and dairy, without considering the possibility that they may have valid reasons for their outrage that have nothing to do with their dietary choices. Additionally, the comment does not address the possibility that some people may be vegetarian or vegan and still be outraged by the topic. Therefore, the comment's argument is not logically reasonable and may be considered fallacious.
		Pred.	Hasty Generalization
		GT	False Dilemma
GFA-W	GPT-3.5	**Input**	"<Are the 2,000 U.S. troops in Syria really up to that task,> should the Turks decide to cleanse the Syrian border of Kurds, or should the Syrian regime decide to take back territory occupied by the Kurds?"
		Output	The fragment highlighted by '<>' is potentially fallacious, specifically exhibiting a form of hasty generalization. It questions whether 2,000 U.S. troops are sufficient to handle major conflicts without providing detailed evidence or context about the capabilities of these troops, their operational support, or the nature of the potential military engagements they might face. This can lead to an oversimplified conclusion about the effectiveness of U.S. military presence based on troop numbers alone, without considering other strategic factors.
		Pred.	Causal Oversimplification
		GT	Doubt Credibility
	Mistral	**Input**	Paige wants to play a guitar solo on our next recording. Her father is a horrible musician, though, so I doubt that Paula's any good either.
		Output	The argument presented is potentially fallacious, known as "Negative Ad Hominem Fallacy", which is an argument that attacks the person rather than the argument, and incorrectly assumes that because the father is a "horrible musician", his daughter (Paige) must be equally "horrible" as well. It is not fair to judge the skills of a person (Paige) based on the skills of another person (her father) without considering Paige's own talents or skills.
		Pred.	Red Herring
		GT	Hasty Generalization
CoT	GPT-4	**Input**	"I have contacted the school's principal, Dr.", "Branch, and I trust that he will take this matter seriously and deal with it appropriately." "Is the principal acting on this?"
		Output	The fallacy present in the segment highlighted by '<>' is "Red Herring." This is because the question about the principal's action diverts attention from the main issue discussed in the article (Islamizing the schools) to the actions of the school principal, which is a separate issue.
		Pred.	Red Herring
		GT	Doubt Credibility
	Llama-3	**Input**	I don't think you can just rely on U.N. sanctions, Randee. But you're absolutely correct, it is a threat, it's a huge threat.
		Output	After analyzing the speech, I found that the fallacy present in the highlighted argument is the False Causality (Post Hoc Fallacy). The speaker is implying that the threat from North Korea grew while the President was preoccupied with Iraq, suggesting that the President's focus on Iraq caused the threat to grow. However, this is a post hoc fallacy because it's an assumption that the President's focus on Iraq caused the threat to grow, without providing any evidence or logical connection between the two.
		Pred.	False Causality

		GT	Appeal to Emotion
P&C	GPT-4	**Input**	I bought a ticket to win a new car at the mall, since I have never won anything like that in the past."",
		Output	The premise that the person has never won anything like a new car in the past does not logically entail the conclusion that they bought a ticket to win a new car at the mall. The premise provides a context or a possible motivation for why they might have decided to buy a ticket, perhaps hoping to change their luck, but it does not necessarily lead to the conclusion that buying a ticket was a rational or necessary action. There is no causal or logical necessity connecting the lack of previous winnings to the decision to buy a ticket; the conclusion does not directly follow from the premise.
		Pred.	No Fallacy
		GT	False Causality
	Llama-3	**Input**	A: Is Turkey ready to join the European Union? B: Turkey helped us with the Refugees. They were beneficial for the EU there, they will be beneficial for the EU in general.
		Output	The premise “Putin has nothing to say regarding the EU. It is our choice who joins and who does not” does not entail the conclusion “Turkey is ready to join the European Union”. This premise only establishes that the decision to admit Turkey into the EU is up to the EU member states, not Putin, and does not provide any information about Turkey’s readiness to join. The premise “Turkey and Europe share much culture and many people from Turkey already live in European countries. They are ready” is more relevant to the conclusion, but it is still a weak argument. The fact that Turkey and Europe share culture and have people living in each other’s countries does not necessarily mean that Turkey is ready to join the EU. There are many other factors to consider, such as Turkey’s political system, economic stability, and human rights record, among others. Therefore, the premises do not entail the conclusion, and B’s argument is flawed.
		Pred.	Appeal to Consequences
		GT	Hasty Generalization

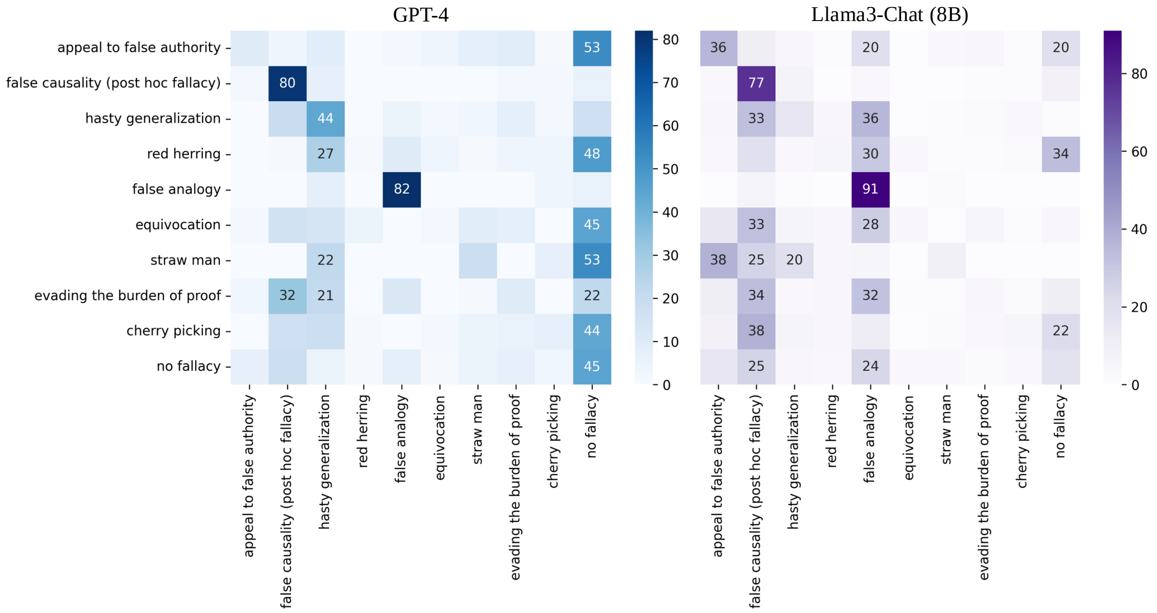
Table 30: Error cases of GPT-4 and Llama3-Chat (8B) for each multi-round prompting scheme.

D.3 Misclassification Confusion Matrices of GPT-4 and Llama3

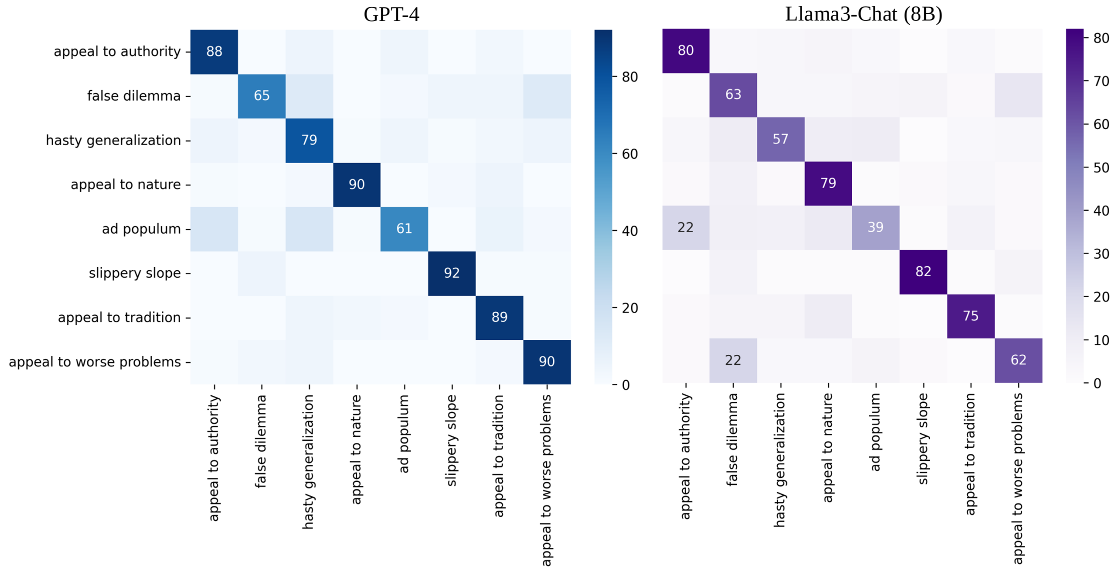


(a) Argotario

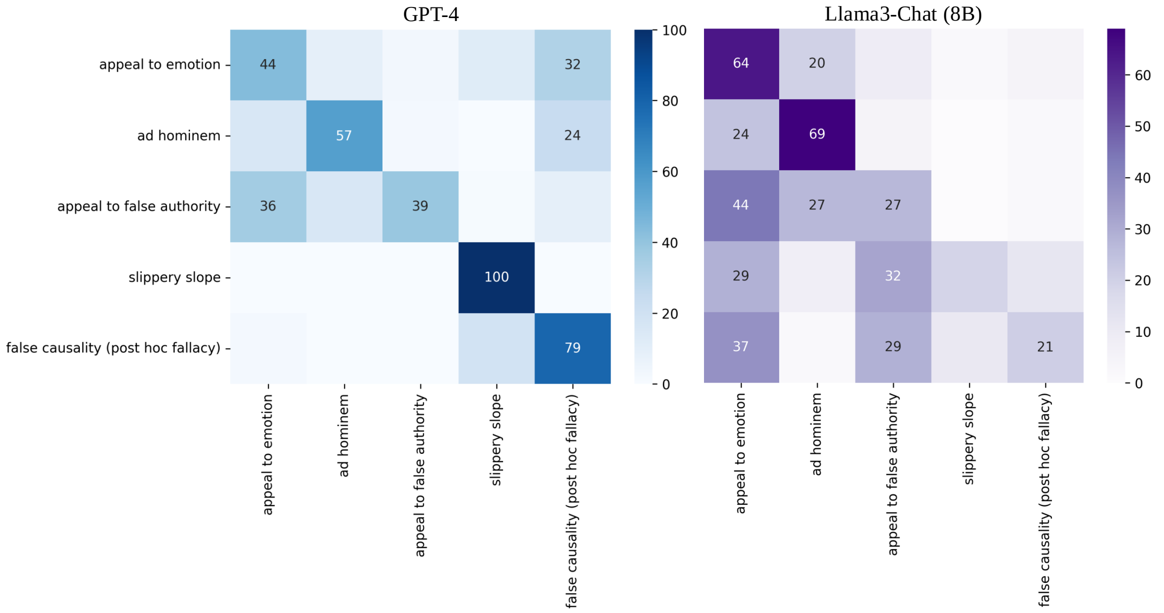
Figure 7: Misclassification confusion matrix of GPT-4 and Llama3-Chat (8B) on ARGOTARIO. Cell values are the percentages of row fallacy types that are misclassified as column fallacy types.



(b) Covid-19

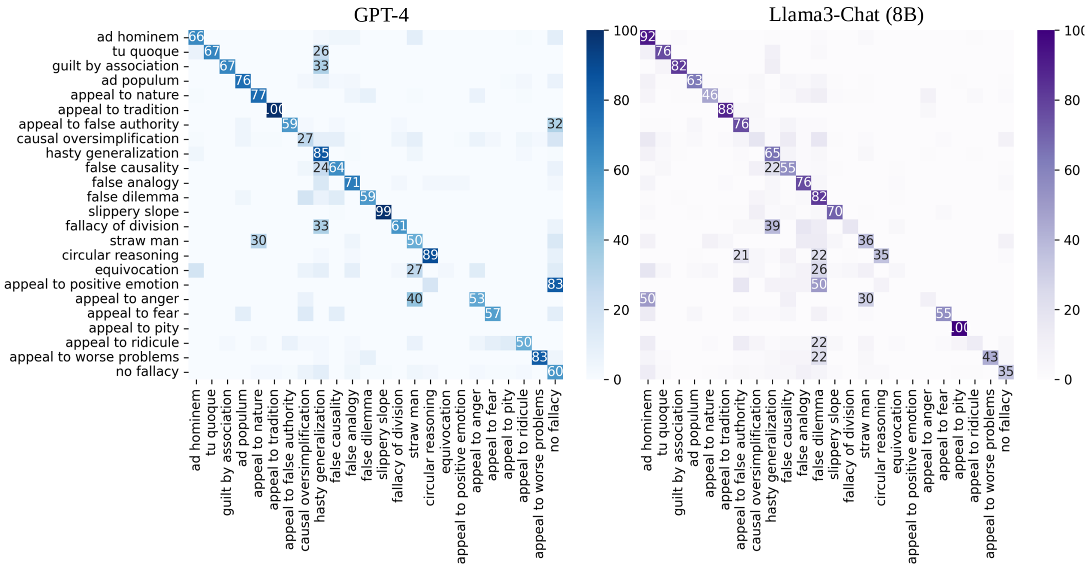


(c) Reddit

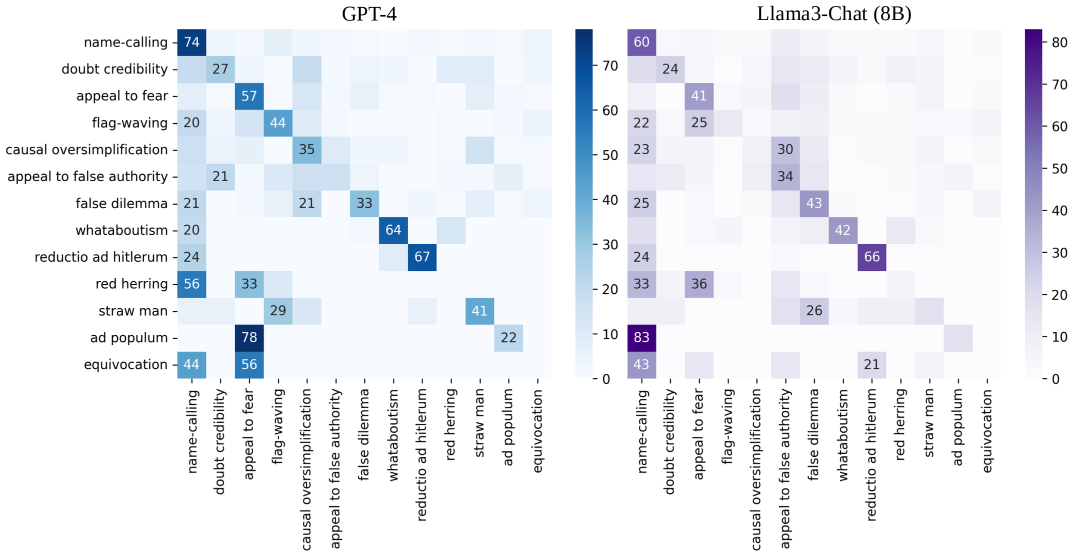


(d) ElecDeb

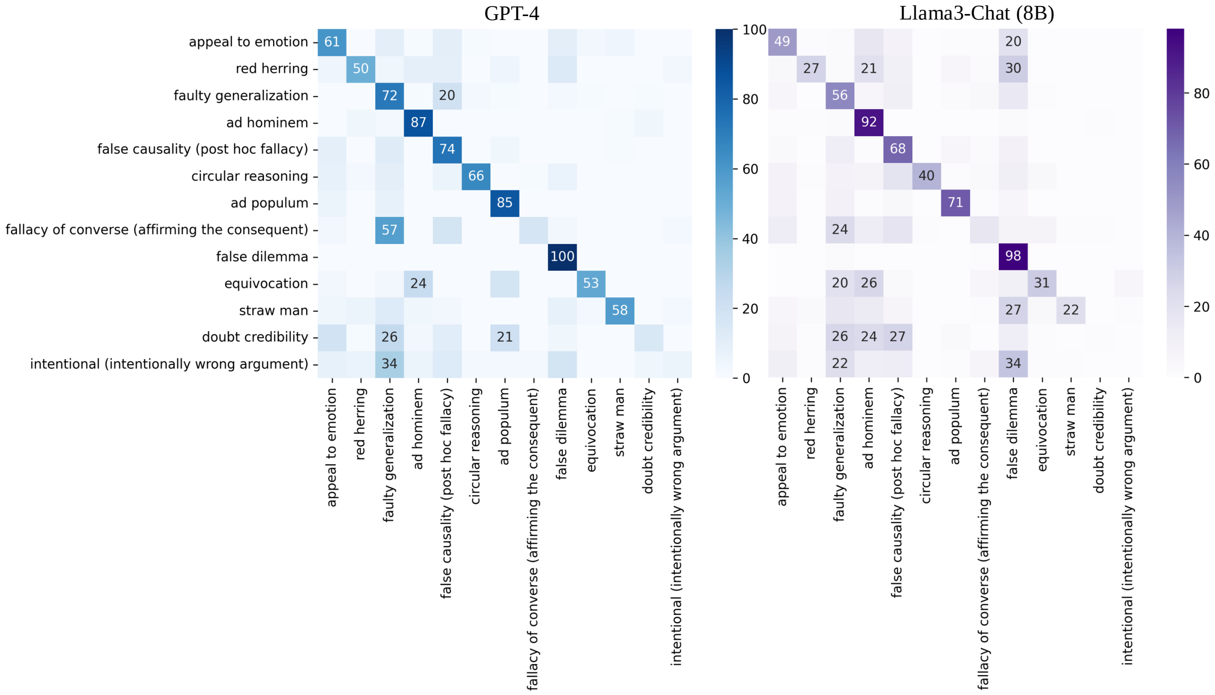
Figure 8: Misclassification confusion matrix of GPT-4 and Llama3-Chat (8B) on COVID-19, REDDIT and ELECDEB. Cell values are the percentages of row fallacy types that are misclassified as column fallacy types.



(e) Mafalda



(f) Propaganda



(g) Logic

Figure 9: Misclassification confusion matrix of GPT-4 and Llama3-Chat (8B) on MAFALDA, PROPAGANDA and LOGIC. Cell values are the percentages of row fallacy types that are misclassified as column fallacy types.