# "The Facts Speak for Themselves": GPT and Fallacy Classification

**Erisa Bytyqi** **Annette Hautli-Janisz**
Faculty of Computer Science and Mathematics
University of Passau
firstname.lastname@uni-passau.de

## Abstract

Fallacies are not only part and parcel of human communication, they are also important for generative models in that fallacies can be tailored to self-verify the output they generate. Previous work has shown that fallacy detection and classification is tricky, but the question that still remains is whether the inclusion of argumentation theory in prompting Large Language Models (LLMs) on the task enhances the performance of those models. In this paper we show that this is not the case: Using the pragma-dialectics approach to fallacies (Van Eemeren and Grootendorst, 1987), we show that three GPT models struggle with the task. Based on our own PD-oriented dataset of fallacies and a carefully curated extension of an existing fallacy dataset from Jin et al. (2022a), we show that this is not only the case for fallacies "in the wild", but also for textbook examples of fallacious arguments. Our paper also supports the claim that LLMs generally lag behind in fallacy classification in comparison to smaller-scale neural or even statistical models.

## 1 Introduction

Fallacies are part and parcel of human argumentation, they are woven into our conversations and with the rise of misinformation, fallacies point to communication components that are crucial to identify in order to differentiate between valid and invalid arguments. But fallacies are also crucial for Large Language Models (LLMs) in that the models should be tailored to self-verify the output they generate, an area that will gain significance with the increasing ubiquity of those models in everyday communication. Even though fallacies have attracted millenia of work in argumentation theory, they have proven to be a tricky feat in argument mining (Jin et al., 2022a; Ruiz-Dolz and Lawrence, 2023; Alhindi et al., 2023, inter alia) – they are hard to detect and even harder to classify. This also holds true when LLMs are put to the task, models, which have shown impressive capabilities in a number of other NLP tasks.

In the present paper, we use the theory of Pragma-Dialectics (PD) (Van Eemeren and Grootendorst, 1987) to guide the models with a solid theoretical foundation of fallacies. The ten rules in PD that, if violated, create a fallacy, give direction to a successful discussion, the argument stage in which these rules are applicable, and the interlocutor who can break the rules (the antagonist and/or the protagonist). This level of detail allows us to craft the prompts in a controlled manner steered by the extent to which we include the aforementioned elements in them. The paper shows that even with significant manual effort in prompt design, both in terms theoretically-driven explanation in the form of pragma-dialectic rules and reasoning chains for large sets of examples, the task of fallacy detection remains prone to errors. This is illustrated based on a new dataset of manually curated fallacies from a PD textbook (Van Eemeren and Grootendorst, 1987) and an enhanced version of a larger-scale, general-purpose fallacy dataset from Jin et al. (2022a), which we manually enhance with the violated PD rule and their reasoning chains.

Overall, 'the facts speak for themselves': The three GPT models GPT-3.5, GPT-4 and GPT-4o struggle even when they are prompted carefully with (a) a solid theoretical foundation of what constitutes a fallacy and (b) manually crafted reasoning chains as examples in the prompt. The models improve to some extent with prompt engineering, but there is no evidence that later versions of GPT (which are significantly larger) generally perform better on the task. We do find that GPT-3.5 and GPT-4o benefit from chain-of-thought prompting, which surprisingly is not the case for GPT-4.

The paper proceeds as follows: Section 2 summarizes previous work on fallacy classification, with 3 describing the data collection and enhance-



*Proceedings of the 12th Argument Mining Workshop*, pages 1–10

ment. Section 4 details the prompt engineering process, the results of which are presented in Section 5. Section 6 discusses the findings and concludes the paper.

## 2 Background

Even though the field of argument mining has attracted significant attention in the last 10 years, also with the rapid progress of deep learning, the automatic identification and classification of fallacies is still one of the main open issues. Overall, much of the previous literature relies on the 'innate' capabilities of the LLMs, i.e., the model(s) are queried outright for the task of fallacy detection without fine-tuning or with little to no prompt engineering. One exception is Ruiz-Dolz and Lawrence (2023), who use the argument schemes by Walton et al. (2008) to guide the model in capturing the fallacious nature of natural language arguments. The authors present a classification task where four of the classes contain fallacious arguments ('Appeal to Authority', 'Appeal to Majority', 'Slippery Slope' and 'Ad Hominem') which are related to seven argumentation schemes. The fifth class contains non-fallacious arguments.

Goffredo et al. (2023), along the lines of the majority of other work, remain theory-agnostic and assume six fallacy types (they partly overlap with those of Ruiz-Dolz and Lawrence (2023)). Jin et al. (2022a) construct two datasets and test 12 different LLMs for their fallacy detection abilities. In the end, a structurally aware classifier (of significantly smaller size) outperforms the LLMs for the same task. Hong et al. (2024) split fallacies into two main groups, namely formal and informal fallacies. The results show that LLMs have a harder time with fallacies that are present in the logical structure of the argument and perform somewhat better for fallacies that are related to the actual content. GPT-3 performs well on the Argotario dataset (Habernal et al., 2017), but is outperformed by the T5 model on the the other four datasets.

## 3 Data

Our capability assessment builds on an integrative approach to data collection, i.e., we consult two different data sources, combine the data points and analyze the performance of the models on the individual as well as the combined dataset. Both datasets comprise of informal fallacies and are described more closely in the following.[1]

### 3.1 The PD dataset

The PD dataset (henceforth, 'PD-data') draws on textbook examples in (Van Eemeren and Grootendorst, 1987) and comprises of clear-cut instances of fallacious arguments that violate one of the ten pragma-dialectical rules. Each of the arguments in PD-data is accompanied by the rule that is violated plus the reasoning behind its fallacious nature. For instance, Example (1) (Van Eemeren and Grootendorst, 1987, p. 285) violates the 'Freedom Rule' because a personal attack is performed on an opponent by casting suspicion on his motives. PD-data contains these three pieces of information: the actual example, the rule that is violated and the explanation given in the textbook.

(1) *He just says so because he wants to be elected.*

This is an extensive manual effort, illustrated furtherby Example (2), which also violates the 'Freedom Rule', but the reasoning as to why the rule was violated differs from the earlier example. Here, the personal attack takes the form of trying to depict the opponent as stupid, bad, unreliable and so forth (instead of casting suspicion on the motives as in (1)). Therefore there is no one-to-one mapping between the rule and the explanation, instead the latter one is solely based on the content of the example.

(2) *Don't listen to this moron, crook, liar, etc.*

The manual data collation in (Van Eemeren and Grootendorst, 1987) yields an initial seed set of 43 items. While this is already a significant increase in data points compared to (Ruiz-Dolz and Lawrence, 2023), we still construct an additional 43 fallacious arguments that are minimal pairs to the instances in PD-data: They mirror the violated rule and the reasoning behind the violation, but contain a slight variation in the linguistic surface. An example of the hand-crafted minimal pair of (1) is shown in Example (3). For the additional examples we also record the example, the violated rule and the reasoning.

(3) *She only agrees with that because she wants to win their approval.*

[1] Both datasets with PD rules and reasoning chains are available at https://github.com/Erisa-Bytyqi/PD-data.

Overall we end up with 86 data instances that are in violation of the 10 pragma-dialectic rules, a substantial increase from the 14 natural language arguments that constitute the dataset of (Ruiz-Dolz and Lawrence, 2023). Table 1 shows the distribution of data instances across the violated rules. The instances are not evenly distributed across the classes of violated rules, with 'Freedom Rule' and 'Argument Scheme Rule' having seven instances each, as opposed to 'Relevance Rule' and 'Unexpressed Premise Rule' which contain two instances each. However, given that we are not interested in training a fallacy classifier but evaluate a pretrained model on its performance, this does not have an effect on the evaluation.

| Violated PD rule | #Orig | #Added |
|---|---|---|
| Freedom Rule | 7 | 7 |
| Obligation To Defend Rule | 4 | 4 |
| Standpoint Rule | 3 | 3 |
| Relevance Rule | 2 | 2 |
| Unexpressed Premise Rule | 2 | 2 |
| Starting Point Rule | 4 | 4 |
| Validity Rule | 7 | 7 |
| Argument Scheme Rule | 5 | 5 |
| Concluding Rule | 3 | 3 |
| Language Use Rule | 6 | 6 |
| Overall | 43 | 43 |

Table 1: Distribution of fallacy types in PD-data

### 3.2 The enhanced LOGIC dataset

The second dataset comes from (Jin et al., 2022a)[2], a dataset that encompasses a range of general logical fallacies, split across 13 different classes (henceforth, 'LOGIC') ('Faulty Generalization', 'Ad Hominem', 'Ad Populum', 'False Causality', 'Circular Reasoning', 'Appeal to Emotion', 'Fallacy of Relevance', 'Deductive Fallacy', 'Intentional Fallacy', Fallacy of Extension', 'False Dilemma', 'Fallacy of Credibility', 'Equivocation'). There are a total of 2449 instances in the dataset sourced mainly from student quiz websites.

Our study considers only a subset (300 out of 2449) of LOGIC, a set of fallacies that violate one of the ten pragma-dialectic rules. To that end, six of the thirteen fallacy types in LOGIC are mapped to their corresponding rule violation in pragma-dialectics, thereby harmonizing the LOGIC and PD-data labels. The dataset contains instances such as Example (4), a textbook case of an 'Ad Hominem' fallacy:

(4) *You're too ugly to be class president!*

In order to use LOGIC for the study in this paper, we manually map the LOGIC fallacy types to the rule violation stipulations in pragma-dialectics. To illustrate this, the personal attack in Example (4) is treated as a 'Freedom Rule' violation in PD, because it *"attacks the other party's person"* (Frans H. van Eemeren, 2020). We also map the 'Appeal to Emotion' fallacies in LOGIC to the 'Freedom Rule' in PD, because they a) unambiguously violate the rule and b) cannot be attributed to any other pragma-dialectic rule. A LOGIC instance of the 'Appeal to Emotion' fallacy is given in (5).

(5) *If you love your family, you'll buy this new stealth security system.*

Another mapping holds between the 'Circular Reasoning' fallacy in LOGIC and the 'Starting Point Rule' in PD. As illustration, see Example (6): The argument ('she is better than anyone else') merely restates the standpoint ('she is the best'), and as such violates the 'Starting Point Rule' of pragma-dialectics, thereby validating its classification under this rule.

(6) *She is the best because she is better than anyone else.*

The fallacy types 'Faulty Generalisation' and 'False Causality' are mapped to the 'Argument Scheme Rule' in PD, and 'Equivocation' is mapped onto the 'Language Rule' in PD. The LOGIC fallacies of 'Ad Populum', 'Fallacy of Relevance', 'Deductive Fallacy', 'Intentional Fallacy', Fallacy of Extension', 'False Dilemma', and 'Fallacy of Credibility' cannot be mapped reliably onto the PD-data rules. This is a result of the pragma-dialectic postulations, i.e., the stage of the argument, the interlocutor 'allowed' to violate a rule, and the argument from a conflict resolution perspective. We briefly illustrate this by way of the 'Ad Populum' fallacies in LOGIC which violate both the 'Relevance Rule' and 'Argument Scheme Rule' of pragma-dialectics, which are considered as *"variants of a fallacy which are not the same kind of fallacy when viewed from the perspective of resolving differences of opinion ´ ´* (Frans H. van Eemeren, 2020). Examples (7) and (8) are of the 'Ad Populum' fallacy type. However, Example (7) violates

[2]Code and dataset available at https://github.com/causalNLP/logical-fallacy

| LOGIC fallacy type | PD rule | #Instances |
|---|---|---|
| Faulty Generalisation | Argument Scheme Rule | 61 |
| Ad Hominem | Freedom Rule | 41 |
| Appeal to Emotion | Freedom Rule | 23 |
| Circular Reasoning | Starting Point Rule | 19 |
| False Causality | Argument Scheme Rule | 18 |
| Equivocation | Language Use Rule | 5 |
| Overall | | 197 |

Table 2: Fallacy Type distribution in LOGIC and PD

the 'Relevance Rule' as the audience's feelings or prejudices are exploited to defend the standpoint; this constitutes a non-argument, hence the violation. In Example (8), the protagonist, by referring to a kind of authority (here the majority) wants to push forward the truth or acceptability of a standpoint. For this particular case, the use of an unsuitable argumentation scheme results in the violation of the 'Argument Scheme' rule. For simplicity, LOGIC fallacy types that can be attributed to several pragma-dialectic rules are omitted in our study.

(7) *You do want your children to be safe in your own neighbourhood, don't you?*

(8) *Everybody says so, so it must be true.*

Overall, four of the ten pragma-dialectic rules have counterparts in the subset of LOGIC employed for the present study. Table 2 shows the mapping from LOGIC fallacy type to PD rule and the resulting number of data points per PD rule. Again, we see a class imbalance which closely mirrors that of PD-data, with the 'Argument Scheme Rule' and the 'Freedom Rule' instances surpassing those of the 'Starting Point' and 'Language Use Rule'. Aggregated, the 'Argument Scheme Rule' And 'Freedom Rule' categories contain 143 instances, whereas the two remaining rules have only 24 data points.

In summary, our investigation builds upon two datasets of fallacies, both labeled with PD rules, where one dataset (PD-data) contains the example, the rule and the reasoning behind the violation and the second dataset (LOGIC) which contains the example and the violated rule. This has an impact on setting up the prompt to get the responses from the model, which will be detailed in the following.

## 4 Probing the GPT models

For probing the models, we use zero-shot prompting (§4.1), chain-of-thought prompting (§4.2) and two-shot chain-of-thought prompting (§4.3) – methods that have been used in previous work on fallacy classification.

### 4.1 Zero-shot prompting

For the fallacy study in the present paper, the prompt contains the following elements (see Figure 1 in the Appendix for the full rendering): (a) the persona that we ask the model to adopt ('You are the world's leading expert in Pragma-Dialectics...'), (b) more information on pragma-dialectics ('an argumentation theory created by ...), (c) the details regarding the instructions ('You are specifically concerned with fallacies [...]'), (d) more detailed instructions regarding the output format to help minimize redundant information ('Be as concise as possible, name the rule, and give a very brief explanation.'), (e) the actual query to the model ('Given the pragma-dialectic approach to fallacies, ...') and (f) the fallacious argument under investigation, separated with a colon from the preceding material. The response *R* is generated by the model without further interaction.

The zero-shot prompt has two variations, the one just described which we will refer to with 'No Rules' (NR) in Section 5 on the results, plus a second one where we include the ten rules (classes) of fallacies and their definitions from (Van Eemeren and Grootendorst, 1987) in the prompt immediately before the fallacious argument (the 'With Rules' (WR) varation). The reason for this is that (OpenAI, 2024) hints at the fact that the inclusion of additional relevant information might help in obtaining better responses.

### 4.2 Chain-of-thought (CoT) prompting

*Chain-of-thought* (CoT) prompting has been shown to outperform zero-shot prompting for a multitude of reasoning tasks (Wei et al., 2022). Given that reasoning goes hand in hand with fallacy detection, we assume that prompting GPT-3.5 and GPT-4 with CoT prompts yields better responses that with zero-shot prompts. The reasoning chain in the CoT prompt is the same as the original reasoning chain given in the book (an example of a CoT prompt is shown in Figure 2 in the Appendix).

This prompt setting is a bit more taxing than zero-shot prompting since the prompt contains at least one more piece of information which has to be assembled manually for each prompt. The same is true if we escalate prompt sophistication and include two examples in the reasoning chain, as illustrated in the following.

### 4.3 Two-shot CoT

Two-shot CoT prompting increases the possibility that the model correctly interprets the task in the prompt ('the more exemplars the merrier'), a property suggested in previous work. Wei et al. (2022) use 7-shot CoT for commonsense tasks, OpenAI use a 10-shot prompt for their GPT4 evaluation on commonsense reasoning tasks (Achiam et al., 2023). We assume that with this extension, the model 'grasps' the characteristics of the fallacy more easily and is not just triggered by how similar the arguments are on the basis of the words contained in them.

Two-shot CoT prompting is used for the LOGIC dataset for which we do not have reasoning chains and where we use examples from the PD literature for correctly predicting the fallacy. The setup for this task is as follows: each of the instances from the LOGIC dataset has a ground-truth label (the pragma-dialectical rule violated by the instance). The first example in the two-shot CoT prompt is the same that is used in the CoT prompt for PD-data to classify the fallacious arguments from (Van Eemeren and Grootendorst, 1987) of that same class. The second example of the two-shot CoT prompt mirrors the pragma-dialectical fallacious argument from (Van Eemeren and Grootendorst, 1987), incorporating reasoning steps from the textbook that explain the argument's fallacious nature in case GPT did not identify the violated rule correctly in the CoT PD-data study. For those instances where GPT responded with the correct rule and the correct reasoning, GPT's response is turned into the second example in the two-shot CoT prompt, this is the case shown in Figure 3. For the argument 'Don't listen to this moron, crook, liar, etc.', the class and the reasoning steps generated by the model were both correct, making GPT's response a valid chain-of-thought and were therefore included in the two-shot CoT.

### 4.4 General prompting parameters

We restrict the length of the generated responses to 128 and 256 tokens for the LOGIC and PD-data instances, respectively, doing justice to the fact that the textual content of the prompt for LOGIC is longer than that of PD-data (more details in §4.2). The temperature is set to zero and the seed parameter is set to a random number.

## 5 Results

We apply a strict evaluation criterion on the generated responses, namely that both the violated rule and the provided reasoning need to be correct in order for the response to be judged correctly. If only one of these is correct, the response is treated as incorrect. This provides a realistic assessment of the capabilities, because we want to establish how reliable the models are without additional human interference, such as needing to determine the correct and incorrect portions of the GPT response. The metric we use to report the performance is accuracy, i.e., the fraction of correct predictions made by the three models.

**Zero-shot prompting** Table 3 provides an overview of the accuracy of the models for the zero-shot prompt setting on PD-data. Overall we can conclude that the performance is low, GPT-4o without rules only achieves an accuracy of .13, GPT-3.5 is at .3 and GPT-4 is slightly better with an accuracy of .49. Those results are comparable to those reported in Jin et al. (2022a) for GPT-3, but worse than those reported by Ruiz-Dolz and Lawrence (2023) who use GPT-3.5 and 4 and Walton's argument schemes. Adding the rules of PD to the prompt decreases the performance for GPT-3.5 and 4 (to .12 and .43 respectively), whereas it slightly helps GPT-4o (which still underperforms with an accuracy of .39). This suggests that the additional information rather confuses than helps the model.

If we dive into the performance regarding individual fallacies, we see significant differences.

| Violated PD rule | #Instances | Zero-shot | | | | | | CoT | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | GPT-3.5 | | GPT-4 | | GPT-4o | | GPT-3.5 | GPT-4 | GPT-4o |
| | | NR | WR | NR | WR | NR | WR | | | |
| Freedom Rule | 14 | 0.43 | 0.14 | 1 | 0.86 | 0.57 | 0.85 | 0.71 | 1 | 1 |
| Obligation To Defend Rule | 8 | 0.75 | 0.25 | 0.75 | 0.25 | 0 | 0.75 | 1 | 0.75 | 1 |
| Standpoint Rule | 6 | 0 | 0.33 | 0 | 0.67 | 0 | 0.33 | 0.67 | 0 | 0.33 |
| Relevance Rule | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0.5 | 1 | 0 |
| Unexpressed Premise Rule | 4 | 0 | 0 | 0 | 0.5 | 0 | 0 | 0 | 0.5 | 1 |
| Starting Point Rule | 8 | 0.25 | 0 | 0.25 | 0 | 0 | 0 | 0.75 | 0.5 | 0.75 |
| Validity Rule | 14 | 0 | 0 | 0 | 0.6 | 0 | 0 | 0.2 | 0.2 | 0.8 |
| Argument Scheme Rule | 10 | 0.43 | 0 | 0.71 | 0 | 0.28 | 0.14 | 0.71 | 1 | 1 |
| Concluding Rule | 6 | 0 | 0 | 0 | 0.33 | 0 | 0.33 | 0.67 | 0.67 | 1 |
| Language Use Rule | 12 | 0.17 | 0 | 0.83 | 0.67 | 0 | 0.66 | 0.33 | 0.67 | 1 |
| Overall | | 0.3 | 0.12 | **0.49** | 0.42 | 0.13 | 0.39 | 0.6 | 0.63 | **0.86** |

Table 3: Accuracy of the GPT models for the two zero-shot prompt settings with no rules (NR) and with the rules (WR) and the chain-of-thought prompt for PD-data.

For instance, the 'Freedom Rule' appears to be the class that all models have the least difficulty with in predicting correctly. But we cannot draw general conclusions, except that the inclusions of rules (WR) seems to trigger lower performance (except for 'Unexpressed Premise Rule' and 'Concluding Rule' in GPT-4). An interesting observation is that the incorporation of additional information in the prompt, namely the ten pragma-dialectic rules and their definitions (WR), degrades the performance of GPT-3.5 and GPT-4 with respect to the classification of 'Argument Scheme Rule' violations. This particular prompt setting leads to all PD-data instances of this class being misclassified as violations of either the 'Standpoint Rule' or the 'Validity Rule'.

**CoT prompting** The last two columns in Table 3 report the results when using CoT prompting on PD-data. Overall, CoT prompting significantly increases the performance of GPT-3.5 (overall accuracy of .6), to the extent that it is comparable to GPT-4 in the same setting (overall accuracy of .63), despite being much smaller in the number of parameters ('size'). GPT-4o shows the strongest results here (accuracy of .86), which leads us to conclude that the extensive manual effort in extracting textbook reasoning chains that are then used for prompting the model pays out.

A more detailed manual analysis of the misclassified instances reveals that the arguments in violation of the 'Unexpressed Premise Rule' are erroneously classified as violations of the 'Standpoint Rule' in all but two instances. This phenomenon can be attributed to what these two rules entail as violations: In case of the 'Standpoint Rule', the distortion of the co-interlocutors standpoints by either means of oversimplification (of their qualifications) or exaggeration (of their statements) is a direct violation of the rule (Van Eemeren and Grootendorst, 1987). This is also known as a straw man fallacy. The 'Unexpressed Premise Rule', on the other hand, is violated when an unexpressed premise is either exaggerated or not correctly reconstructed by the antagonist and then denied by the protagonist, which is a special case of the straw man fallacy. Regarding the *PD-data* dataset, the straw-man fallacy (regardless of its nature) is overwhelmingly associated with a violation of the 'Standpoint Rule' by the models.

Diverging from the assumption that CoT prompting improves the classification of fallacious arguments, the case of the 'Language Use Rule' shows that in fact zero-shot prompting performs slightly better than CoT for GPT-3.5 and 4, but not so for GPT-4o (accuracy of 1). In sum it is difficult to establish general conclusions as to which prompt design leads to consistently better results for fallacy detection. This issue of drawing meaningful insights is supported in the following where we include fallacies from a larger dataset and use PD to identify their fallacy type.

**Two-shot CoT prompting** Two-shot CoT prompting is performed on the LOGIC dataset for which no reasoning chains are available, under the assumption that if we include two examples of PD-data fallacies and their violated rule in the prompt the model is better able to classify those examples with relatively high performance.

The overview of the results in Table 4 paints a different picture, however. All models struggle to correctly classify the majority of the LOGIC arguments (accuracies of 0.23 for GPT-3.5 and GPT-4o

| LOGIC type | PD rule | 2s-CoT | | |
|---|---|---|---|---|
| | | GPT-3.5 | GPT-4 | GPT-4o |
| Faulty Generalisation | Argument Scheme Rule | 0.44 | 0.51 | 0.57 |
| Ad Hominem | Freedom Rule | 0.1 | 0.02 | 0.02 |
| Appeal to Emotion | Freedom Rule | 0.07 | 0.07 | 0.08 |
| False Causality | Argument Scheme Rule | 0.28 | 0.78 | 0.44 |
| Equivocation | Language Use Rule | 0 | 0 | 0 |
| Overall | | 0.23 | **0.29** | 0.23 |

Table 4: Results of 2-shot CoT prompting for a subset of the LOGIC datset.

and 0.29 for GPT-4). Striking is the difference in performance for the 'Freedom Rule', where all models was much better in identifying this type of fallacy in PD-data. In the case of the 'Ad Hominem' fallacy type, this inconsistency may be attributed to the difference in explicitness of the character attack in the two datasets. Arguments from LOGIC are to some extent more subtle in comparison to the text examples from PD. An example is given in (9) below: 'Don't listen to this moron, crook, liar, etc.' and 'Anyone who says that about me [that he's a racist bigot] is a Nazi' are much more stronger in terms of nature and wording than the LOGIC ad hominem 'Students who want cell phones in school have no idea what they're talking about'.

(9) **PD-1:** *Don't listen to this moron, crook, liar, etc.*
**PD-2** *Anyone who says that about me [that he's a racist bigot] is a Nazi."*
**LOGIC:** *Students who want cell phones in school have no idea what they're talking about.*

A better performance, especially for GPT-4, is observed for the identification of 'Argument Scheme Rule' violations: When an argument makes the erroneous assumption that the correlation of two events means they have a cause-effect relationship, this is known as a 'False Causality' fallacy (Jin et al., 2022a) and a violation of the 'Argument Scheme Rule'. For more than half the instances belonging to this class, GPT-4 correctly identifies where the reasoning of the arguments falls apart, i.e., it identifies that correlation does not mean causation and that hence the argument is fallacious and violates the 'Argument Scheme Rule'. As noted previously, the ambiguous nature of arguments is notably difficult for both GPT models. This inability leads to violations of the 'Language Use Rule' not being correctly identified, as is the case for the LOGIC arguments, where this class has the poorest performance with no instances correctly classified.

Overall, the results indicate that both GPT models struggle to apply the pragma-dialectic model to fallacy detection and classification on data in the wild, i.e., data that does not originate in pragma-dialectic textbooks and arguments similar to them. The prompt content, as was our hypothesis, has a significant impact on the the models' performance, however, contrary to OpenAi's reporting (OpenAI, 2024), including fallacy definitions lowered the performance of the models. What emerges from the results reported here is an indication that fallacy detection and classification, which are also important in terms of having the models self-verify the content they generate as sound or not remains quite a challenging task no matter the model used to approach fallacies.

## 6 Discussion and conclusion

The present study explores the capabilities across a set of GPT models for the task of fallacy classification according to the pragma-dialectic theory of argumentation. Building on the success of chain-of-thought (CoT) prompting for several reasoning tasks, the models were subjected to zero-shot and CoT prompting for the task of classifying the fallacious arguments from Van Eemeren and Grootendorst (1987) and a subset of the LOGIC dataset (Jin et al., 2022b) as violations of one of the ten pragma-dialectic rules. In the course of this study we created a novel dataset (PD-data) comprising 86 fallacious arguments plus their reasoning chains that explain their fallacious nature. The prompts use best practices as described in OpenAI (2024) and the reasoning chains are informed by the fallacy definitions in Van Eemeren and Grootendorst (1987).

The finding of this investigation complement those of earlier studies, namely that the LLMs struggle to correctly identify the type of fallacy committed in an argument. It seems that language models with much smaller size such as RoBERTa (Ruiz-Dolz and Lawrence, 2023), Mul-

tiFusion BERT (Goffredo et al., 2023) or ELECTRA*Structaware* (Jin et al., 2022a) perform better, independent of the theoretical framework for classifying the fallacies. Also similar to previous work is the variability in the results across different models. While almost all arguments from PD-data in violation of the 'Freedom Rule' are classified correctly, there is a steep drop in performance for LOGIC arguments that violate the same rule, irrespective of prompt setting. CoT prompting proves successful for GPT-3.5 and GPT-4o, but does not have a great impact on GPT-4's performance for this task.

One can argue that the overall low performance of the models on LOGIC is due to the fact that the theoretical assumptions in PD do not scale beyond the examples that are mentioned in the textbooks supporting the theory, i.e., the ten rules postulated as violations do not hold when looking at fallacious arguments in the wild. However, given that the mapping between the categories in PD-data and LOGIC is possible, the conceptual assumptions seem to be valid, but it might be the naming of the categories and the wording of the PD rules that is confusing to the model. This only provides further support to the fact that generative models lack pragmatic understanding and provide aligned responses only when the wording in the prompt is informative. In sum, there is still substantial work to be done before we arrive at a systematic assessment of black box large-scale language models, not only in argumentation but in linguistic capabilities as whole. This paper is intended as one building block in this endeavor.

## Limitations

This line of work is subject to at least two limitations. First, the design of our prompts is manual and, while practices reported by previous research for other reasoning tasks were used, we have yet to identify other prompt alterations that might lead to better performance. Second, we only provide at most two exemplars for *chain-of-thought* prompting, which in some cases is not enough to cover all presentations of a fallacy and results in the fallacious argument not being correctly identified as one. Further research might explore the limitation pertaining to the *chain-of-thought* prompt, by incorporating additional exemplars which encompass a wider range of the fallacy's variations, the performance of both GPT models might see an improvement.

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tariq Alhindi, Tuhin Chakrabarty, Elena Musi, and Smaranda Muresan. 2023. Multitask instruction-based prompting for fallacy recognition. *arXiv:2301.09992 [preprint]*. Available from arXiv: https://arxiv.org/abs/2301.09992.

Bart Verheij Frans H. van Eemeren, Bart Garssen. 2020. *Handbook of Argumentation Theory*. Springer Dordrecht.

Pierpaolo Goffredo, Mariana Chaves, 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*, pages 11101–11112, Singapore. Association for Computational Linguistics.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.

Ruixin Hong, Hongming Zhang, Xinyu Pang, Dong Yu, and Changshui Zhang. 2024. A closer look at the self-verification abilities of large language models in logical reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 900–925, Mexico City, Mexico. Association for Computational Linguistics.

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

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

OpenAI. 2024. Prompt Engineering. Https://platform.openai.com/docs/guides/prompt-engineering/prompt-engineering.

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*. Association for Computational Linguistics.

Frans H Van Eemeren and Rob Grootendorst. 1987. Fallacies in pragma-dialectical perspective. *Argumentation*, 1:283–301.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

# A Appendix

*P:* You are the world's leading expert for Pragma-Dialectics, an argumentation theory by Frans H. van Eemeren and Rob Grootendorst. You are specifically concerned with fallacies in Pragma Dialectics and will correctly classify the text in triple quotes. Be as concise as possible, name the rule, and give a very brief explanation.
Given the pragma-dialectic approach to fallacies, which dialectical rule does the following sentence violate: "'The American medical system doesn't care what happens to the patient. I know of a man who was turned away by a hospital and then died."'

*R:* [MODEL RESPONSE]

Figure 1: First variation of the zero-shot prompt using an example from PD-data and the GPT-4 response.

*Q:* You are the world's leading researcher in Pragma Dialectics and provide answers accompanied by very brief reasoning steps and the name of the rule.

Which rule of Pragma-Dialectics does the sentence in triple single quotes violate: "'**This choice will impact your future prospects within the company.**"'

*RC:* As parties must not prevent each other from advancing standpoints, putting pressure on the opponent (here: his choice will affect his job) is a clear violating of the Freedom Rule. Because of the implications, the other party might not advance their standpoint and this is in violation of the aforementioned rule where parties must not prevent each other from advancing or casting doubt on standpoints.

*Q:* Which rule of Pragma-Dialectics does the sentence in triple single quotes violate: "'**You'll have to face the consequences for our personal relationship** "'.

***R:*** [MODEL RESPONSE]

Figure 2: Chain-of-thought prompt for the PD dataset.

*P:* Given the pragma-dialectic approach to fallacies, which dialectical rule does the following sentence violate: **Anyone who says that about me [that he's a racist bigot] is a Nazi."**

*RC:* Parties must not prevent each other from advancing standpoints. If you perform a personal attack on the opponent by depicting him as stupid, bad, etc you directly violate the Freedom Rule. Here by depicting the opponent as a 'Nazi' we have a clear violation of the Freedom Rule.

*P:* Given the pragma-dialectic approach to fallacies, which dialectical rule does the following sentence violate: **Don't listen to this moron, crook, liar, etc.**

*RC:* ...

*P:* Given the pragma-dialectic approach to fallacies, which dialectical rule does the following sentence violate: **Researchers are frauds who don't earn their salaries.**

---

**R:** [MODEL RESPONSE]

Figure 3: Two-shot chain-of-thought prompt for the LOGIC dataset.