Pattern-based Logical Fallacy Detection using Decoder-Only Large Language Models

Anonymous ACL submission

Abstract

In today's fast-paced information era, logical fallacies, defined as defective patterns of reasoning, inevitably contribute to the growth of information disorder. However, often fallacies appear in nuanced forms that complicate automated identification. In this study, we investigate whether the logical structure of arguments proves beneficial for fallacy detection. To address the inherent variability of logical fallacies, we develop an experimental framework that extracts logical patterns from sentences via Large Language Models (LLMs) from the LOGIC dataset. We evaluate the impact of these patterns across different LLMs and experimental zero- and one-shot configurations and we test their robustness on different datasets. Our generated patterns achieve a significant performance increase on LOGIC, validating the effectiveness of this structural approach.

1 Introduction

A logical fallacy is an error in reasoning that renders an argument invalid or unsound. These arguments often appear rational and logically coherent on the surface, but deeper analysis reveals they are not (Copi et al., 1953). Fallacies are traditionally classified into formal and informal types: formal fallacies violate the rules of logical structure regardless of content, while informal fallacies are patterns of mistakes that are made in the everyday uses of language and are related to contextual meaning (Hamblin, 1970; Bacon et al., 1999; Copi et al., 1953).

To evaluate the quality of an argument, it is helpful to reconstruct it into what is known as logical form, the structure that emerges when the specific content of a statement is replaced by variables (Johnson and Blair, 1977). For example, the argument *If it rains, then the ground will be wet. It is raining. Therefore, the ground is wet* has the logical form *If P, then Q. P. Therefore, Q.* Building on this formalization framework, Jin et al. (2022) developed a structure-aware model for fallacy detection on the LOGIC dataset that compares arguments' and fallacies' logical forms. However, this approach may not work effectively for informal fallacies, where reasoning is often more nuanced and context-dependent than abstract representations suggest. Fundamentally, a single logical scheme frequently fails to capture the full spectrum of ways a particular fallacy can manifest in natural discourse.

This gap between theory and practice raises a key question: **Is the logical structure of arguments valuable for automated fallacy detection?** To answer this question, we investigate whether Large Language Models (LLMs) can successfully extract fallacies' logical patterns from fallacious examples and their explanations, attempting to capture both logical forms and context-aware logical schemes that reveal the underlying mechanisms of deception in fallacious arguments. Hereafter, we refer to these extracted structures collectively as *patterns*. While existing supervised approaches require extensive labeled datasets and computational resources for fine-tuning (Lei and Huang, 2024; Vijayaraghavan and Vosoughi, 2022; Sourati et al., 2023), to our knowledge, no prior work has explored fallacy detection from a structural perspective within an unsupervised setting.

We evaluate multiple prompting configurations to determine which informational components enhance performance and examine the impact of demonstrations on detection capabilities. Our approach, incorporating the generated patterns, achieves state-of-the-art results among unsupervised methods on the dataset LOGIC. Finally, to validate the robustness and transferability of our patterns, we assess their performance across three different datasets spanning diverse domains and argumentative styles.

In summary, our contributions are threefold:

- We examine whether Large Language Models (LLMs) can successfully extract reasoning templates for different fallacy types and use them for fallacy classification.
- We evaluate decoder-only LLMs' capabilities in fallacy detection with different prompt designs.
- We validate the generated patterns across three different datasets, testing their generalizability across different domains and structures.

2 Related Work

Recent advances in fallacy detection have increasingly turned to LLMs, though few studies have relied exclusively on prompting-based techniques. Several works have employed fallacy detection to probe LLMs' logical reasoning abilities (Teo et al., 2025; Hong et al., 2024; Li et al., 2024; Xu et al., 2025). Among these, Hong et al. (2024) investigated self-verification capabilities and showed that LLMs face more challenges with structure-based (formal) fallacies with respect to content-based (informal) ones, and that fallacy definitions provide minimal improvements. It is worth noting that only one study has explored fallacy classification using reasoning models, demonstrating superior performance compared to non-reasoning ones (Xu et al., 2025). Among studies relying exclusively on prompting techniques, Pan et al. (2024) designed single-round and multi-round prompting schemes for zero-shot detection, while Jeong et al. (2025) introduced contextual prompting incorporating counterarguments, explanations, and goals with confidence-based ranking, showing that explanations particularly enhance performance. Lim and Perrault (2024) assessed detection abilities on the LOGIC dataset using few-shot prompting, though their different taxonomy limits direct comparison with our work. Other research combines generative LLMs with fine-tuned models, such as Alhindi et al. (2024) who employed LLMs to generate synthetic training examples for fine-tuning classification models. Most notably, Jin et al. (2022) developed a structure-aware model based on Electra that distills arguments into logical forms and compares them against fallacy patterns sourced from logicallyfallacious.com. Their approach significantly outperformed zero-shot experiments, demonstrating the proven importance of structural information in fallacy detection systems. Against this scenario, our work represents the first systematic attempt to exploit logical structure through inference-only methods.

3 Datasets

The LOGIC dataset is a collection of 2449 examples across 13 fallacy types. Instances are sourced from educational platforms about fallacies such as Quizziz and study.com. The dataset consists of brief dialogues and short statements. Given the educational intent behind these examples, sentences tend to have relatively straightforward syntactic structures, making the dataset particularly well-suited for pattern recognition and alignment with logical forms.

Although it contains 13 distinct fallacy types, a thorough analysis revealed that some of the classes actually contain instances of different fallacies that were grouped together. For instance, the class *Hasty Generalization* contains examples of actual *Hasty Generalization* as well as *Slippery Slope* (Table 1). While these grouped fallacies share common logical flaws and thus belong to the same conceptual group, they manifest through different structural patterns, which complicates the attempt to match them all to a single logical scheme.

Class	***Fallacies included***
Intentional Fallacy	Intentional Fallacy Shifting the Burden of Proof Moving the Goalposts No True Scotsman
False Cause	Post Hoc False Cause
Hasty Generalization	Hasty Generalization Slippery Slope

Table 1: Examples of classes in LOGIC containing instances of different fallacy types. While logically coherent, these groupings comprise fallacies with distinct structural patterns. A detailed breakdown of all classes' subtypes is provided in the appendix of Jin et al. (2022).

We use LOGIC to extract logical patterns. We evaluated the generated patterns on LOGIC and on three other datasets: ARGOTARIO (Habernal et al., 2017), a general-domain collection of fallacious arguments annotated with five fallacy classes and a *No Fallacy* class; COVID (Musi et al., 2022), a dataset of COVID-19-related articles annotated with ten fallacy classes and a *No Fallacy* category; CLIMATE (Alhindi et al., 2023), which contains text segments from climate change articles and adopts the same taxonomy as COVID. We preserved the original data splits provided by the authors, with the exception of ARGOTARIO, which we partitioned into a 75/25 train-test split. Datasets' summary is reported in Table 2 and a description of each taxonomy is provided in Appendix D.

Data	# Examples	# Classes	Genre	Domain
LOGIC	2449	13	Dialogue	Education
ARGOTARIO	1344	6‡	Dialogue	General
COVID-19	154	11‡	News	Covid-19
CLIMATE	685	11‡	News	Climate

Table 2: Statistics of the four datasets. ‡ indicates that the *No Fallacy* class is included.

4 Pattern generation

Natural arguments appear in several different forms. Such variability manifests itself in LOGIC dataset as well as many others (Habernal et al., 2018; Da San Martino et al., 2019). For this reason, we address our research question by modeling patterns inductively from the observed text instances. Our pattern generation procedure features two steps:

Step 1: Explanation Generation Explanations have been shown to be instrumental in identifying and discrediting fallacious reasoning, as they make the logical structure of arguments explicit and open to scrutiny (Storer, 1949). Furthermore, Jeong et al. (2025) has demonstrated that providing explanations constitutes valuable contextual information in zero-shot settings.

Given a sentence from the training set and its fallacy label, we used `LLama 3.3-70B` to generate an explanation that justifies why that sentence contains the specified fallacy.

Step 2: Pattern Extraction For each fallacy class, we used OpenAI's reasoning model `o4-mini` (OpenAI, 2025) to extract patterns from the collected sentences and their explanations, requiring the model to preserve logical connectors such as prepositions or adverbs and to abstract away from content words by using placeholders while keeping the original reasoning form.

You can find the used prompts in Appendix F.1. In the initial phase of our research, we aimed to cover two distinct logical aspects from our arguments and explanations, specific to formal and informal fallacies, respectively:

- arguments' **logical forms** as defined by formal logic theory;
- recurring **reasoning schemes** that frequently appear in both sentences and explanations, capturing specific information about the reasoning behind the fallacy, including frequent syntactic particles, phrases, and examples that convey the fallacious intent.

Table 3 shows the patterns relative to the fallacy class *Intention Fallacy*. The full list of patterns is available in Appendix E (Table 17).

The process resulted in approximately 3-6 patterns per fallacy class. Final patterns were obtained after selecting the best performing subset on the validation set, in the attempt to retain only useful information and avoid redundancy.

In some cases, the model is able to detect some specific types of fallacy that deviate from canonical schemes. For example, the model automatically generates the pattern relative to *Tu quoque* (a specific case of *Ad Hominem*). However, the model sometimes fails to capture some frequent fallacy types within mixed classes. This is expected because we include the fallacy class name in the prompt, which likely biases the model toward patterns that match its internal knowledge of that particular class name. To ensure a broader coverage of fallacies listed in Table 1, we manually isolated instances of frequent and undetected fallacies (such as *Shifting the Burden of Proof*) and repeated the procedure.

Intentional Fallacy Patterns
1. The argument assumes that because X (e.g., someone's intention, belief, or lack of counter-evidence), therefore Y is true.
2. Asserting P is true because it has not been disproven.
3. Because the creator intended [interpretation], the work should be understood as [interpretation].
4. Questions framed to presuppose guilt or a specific intention (e.g., "Have you stopped X?"), thus assuming what is to be proven.
5. If A does not have trait X, and X is (allegedly) typical of group G, then A is not a member of G.

Table 3: Patterns for *Intentional Fallacy* combining logical forms (5) and reasoning schemes (4) that encode structure and intent.

5 Experiments

This section describes our experiments for fallacy detection, including our patterns produced by the procedure introduced in Section 4 and several competing prompting strategies. Additional experiments are reported in Appendix B.

We used the following LLMs for our experiments: `o4-mini`, `GPT-4.1-mini`, `LLama-3.3-70B` and `Gemma-3-27B-it` for a total cost of 45 USD. Our intent was to test LLMs from different providers and with different sizes and to compare reasoning and non-reasoning models.

5.1 Prompt Design

Baselines We compared our approach against several baselines that vary in the type and amount of information provided to the model. The simplest baseline (**ZERO-SHOT**) provides only the list of fallacy names in the dataset as a reference, establishing a minimal information condition. Our second baseline incorporates fallacy definitions to provide more comprehensive background knowledge (**DEF**). These definitions were initially sourced from Lei and Huang (2024) and subsequently refined based on our analysis to ensure clarity and consistency. Finally, we tested a baseline using standard logical forms, following the approach of Jin et al. (2022) and sourcing these forms from logicallyfallacious.com. This final baseline (**LOGICAL FORMS**) allows us to assess the effectiveness of expert-made logical representations compared to our generated pattern-based approach.

LLM-derived Patterns and Definitions Beyond generating logical patterns, we leveraged the explanations from Section 4 to automatically create new fallacy definitions based on LOGIC training samples. We then replicated experiment **DEF** with these new definitions (**NEW DEF**). We also exploited the patterns extracted by adding them to the prompt (**PATTERNS**) and by implementing a two-step approach where we first ask the LLM to identify the pattern and then to output the corresponding fallacy (**PATTERN MATCHING**).

One-shot Prompting We further investigated the impact of providing examples to the model through several experimental configurations (Brown et al., 2020). Initially, we tested a static approach where one example per fallacy was randomly selected and shown to all test sentences (**ONE-SHOT**), establishing a baseline for example-based learning. To enhance this approach, we augmented the same examples with manually crafted explanations following our previously established definitions as guidelines (**ONE-SHOT + DEF**). We sampled 5 different example sets and performance across all configurations was assessed over 5 runs to ensure statistical reliability.

Original argument	Every time I wear this necklace, I pass my exams. Therefore, wearing this necklace causes me to pass my exams.
Masked argument	Every time MSK<0> MSK<2>, MSK<0> MSK<4>. Therefore, MSK<2> causes MSK<0> to MSK<4>.

Table 4: Example of a masked argument in LOGIC. The distillation algorithm is explained in Jin et al. (2022). The masked version of the dataset was publicly released by the authors and was not created by us.

More sophisticated was our dynamic one-shot prompting approach (**DYNAMIC ONE-SHOT**), which computes embeddings for both training and test sentences to retrieve the most similar example per class for each test sentence. We used `sentence-transformers/all-MiniLM-L6-v2` model and `cross-encoder/stsb-roberta-base` cross-encoder from SentenceTransformers (Reimers and Gurevych, 2019) to compute embeddings and employed cosine similarity to evaluate similarity. We included the previously generated explanations of examples in the prompt as well (**DYNAMIC ONE-SHOT + EXP**).

Furthermore, we explored structure-focused similarity. Since Jin et al. (2022) released a version of LOGIC with masked arguments (with content words replaced by placeholders), we conducted the same similarity-based procedure using these masked sentences (example in Table 4) in an attempt to force the embedding model to focus on structural rather than lexical similarities. For this configuration (**SYNTAX-BASED DYNAMIC ONE-SHOT**), we used `sentence-transformers/all-MiniLM-L6-v2` from SentenceTransformers alongside a syntax-augmented version of RoBERTa-large extracted from Sachan et al. (2021) (see Appendix C).

Finally, we incorporated the generated patterns into our dynamically retrieved examples and their explanations (**ONE-SHOT + EXP + PATTERNS**).

Multi-step Classification An alternative approach involves decomposing the classification task into three sequential steps within a single model call (**MULTISTEP**) using chain-of-thought prompting (Wei et al., 2023). In the first step, the model is required to generate a logical form representation of the argument according to predefined structural rules (prompt in Appendix F.2). Subsequently, the model should match the generated logical form to one from the ones provided and, as a result, classify the argument.

5.2 Results and discussion

Table 5 shows a consistent improvement when the model leverages information about the underlying logic extracted through the LLMs, suggesting that models were effectively able to capture the necessary information to detect fallacies, especially with `o4-mini`. When using the reasoning model, the model-generated definitions yield a 4.7% improvement over our manually corrected definitions. In the same way, including our generated patterns causes a 9.9% increase with respect to the logical forms extracted by the website logicallyfallacious.com and used in Jin et al. (2022). When it comes to non-reasoning models, the different definitions do not really affect the performance, whereas using our patterns improves the accuracy by 5.2% on average.

A notable result is the performance increase achieved through dynamic one-shot prompting. In particular, **DYNAMIC + EXP** approach yields an average 8.1% increase in Micro F_1 compared to **ONE-SHOT + EXP**, despite relying on semantic similarity for example selection. On the other hand, adopting a syntax-oriented example selection strategy (**SYNTAX-BASED DYNAMIC ONE-SHOT**) does not produce any improvement. This may be partially due to inaccuracies in the sentence masking process, which can negatively impact the retrieval of similar examples and the classification, consequently. The **MULTISTEP** approach shows significantly weaker performance than **PATTERN MATCHING** for `o4-mini` and `llama-3.3-70B`, implying that generating logical forms without explicit guidance constitutes the main challenge for the model in the request.

In summary, performance benefits from structure-based information, indicating that incorporating logical reasoning structure into prompts enhances fallacy detection and our best model, **DYNAMIC+EXP+PATTERNS** outperforms the state-of-the-art on LOGIC in an unsupervised setting by 25.5% (Table 7).

5.3 Error analysis

Pattern matching Requesting the model to identify the closest pattern for each argument provides insight into the association process between sentences and patterns. For our analysis, we have split our fallacies into two groups in Table 8: i) group 1, consisting of informal fallacies whose patterns include logical forms while still including additional contextual cues; ii) group 2, consisting of informal fallacies that lack highly structured patterns and rely more on contextual and semantic features of the sentence.

Figure 1 shows consistently superior F_1 scores for Group 1, whose classes maintain relatively high performance across all experimental settings. The class *Circular Reasoning* emerges as the most accurately predicted class across all models. For what concerns Group 2, the overall F_1 is, on average, 22.25% lower with respect to Group 1, though this gap is least pronounced for `o4-mini`. The classes *Emotional Language*, *Red Herring* and *Extension Fallacy* achieve moderate prediction accuracy, whereas only *Evading the Burden of Proof*'s patterns within the *Intentional Fallacy* category are correctly classified, and *Equivocation* remains entirely undetected by `GPT-4.1-mini`. In summary, the models achieve better performance on logical fallacies that exhibit clearer structural characteristics but face difficulties with fallacies requiring more nuanced semantic understanding and contextual analysis.

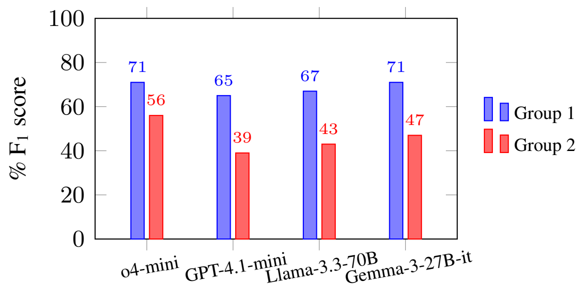

Figure 1: Group-wise F_1 scores for each model, relative to the **PATTERN MATCHING** prompt setting.

Furthermore, matching patterns allows us to see that some instances can be deemed as fitting from a structural point of view, thus partially explaining the inherent difficulty of the classification task. While providing guidance through logical structure proves beneficial for fallacy detection, this approach does not eliminate all sources of ambiguity, as some sentences may conform to multiple structural patterns. The critical point lies in context-aware pattern application: models must not only identify matching logical forms but also evaluate whether the specific content and contextual factors make those patterns valid in each specific instance.

To quantify the degree of ambiguity inher-

Method	o4-mini		gpt-4.1-mini		llama-3.3-70B		gemma-3-27b-it	
	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1
Baselines								
ZERO-SHOT	61.7	62.5	**57.8**	**59.3**	55.8	57.1	60.5	62.6
DEF	62.1	62.8	57.8	57.7	59.1	59.3	**63.5**	**63.8**
LOGICAL FORMS	**63.2**	**65.0**	57.8	57.9	**60.2**	**60.3**	62.8	63.0
LLM-derived Patterns and Definitions								
NEW DEF	66.8	67.3	57.5	57.9	58.8	59.0	64.8	64.9
PATTERNS	**72.2**	**72.7**	63.5	63.5	**64.5**	**64.6**	**68.5**	**68.5**
PATTERN MATCHING	70.1	70.3	**65.2**	**65.2**	66.2	66.3	67.2	67.2
One-shot prompting								
ONE-SHOT	63.6	63.9	56.2	56.8	56.1	56.3	60.0	60.4
ONE-SHOT + EXP	65.2	65.1	56.8	57.3	56.3	56.5	59.2	59.7
DYNAMIC ONE-SHOT								
all-MiniLM-L6-v2	70.2	70.5	65.8	66.1	65.5	65.6	68.5	68.9
roberta-base	69.5	70.1	65.5	66.1	64.8	64.9	66.5	66.8
SYNTAX-BASED DYNAMIC ONE-SHOT								
all-MiniLM-L6-v2	68.2	68.5	63.2	63.4	62.8	62.8	64.5	64.6
syntax-augmented roberta-large	65.5	68.9	64.5	65.2	64.2	64.2	63.5	67.2
DYNAMIC + EXP	71.2	72.4	**67.8**	**68.2**	**67.5**	**67.6**	68.2	68.5
DYNAMIC + EXP + PATTERNS	**74.2**	**75.9**	66.8	67.2	67.2	67.3	**70.5**	**71.0**
Multi-step classification								
MULTISTEP	65.4	64.9	65.8	66.1	62.5	62.6	66.8	67.2

Table 5: Logical fallacy classification performance. **Bold**: best approach in section per model, **Bold**: best approach overall per model. F_1 score denotes Micro F_1 score, which accounts for the significant class imbalance in the dataset.

ent in pattern matching, we instructed the best-performing model `o4-mini` to return the five most similar patterns for each argument. This multi-candidate approach enables us to analyze whether lower-ranked patterns might also represent valid interpretations of the same argument. By examining the distribution of pattern similarities and evaluating classification accuracy when considering alternative matches, we can better understand the boundaries of pattern-based classification and identify instances where structural ambiguity genuinely complicates fallacy detection.

Table 9 shows that, when the model is prompted to return multiple matching patterns rather than a single best match, its confidence in the initial prediction decreases, resulting in a 3.4% drop in accuracy. However, this apparent degradation is misleading when viewed in isolation. By incorporating the second-ranked pattern choice into our evaluation, performance recovers to 75.1%, and continues to improve as we expand our candidate pool to include progressively lower-ranked options. Table 10 illustrates a representative case where the model successfully identifies the correct pattern as its second choice, while its first-ranked selection remains structurally plausible: the model likely assigns one of the *Ad Populum* patterns because it closely matches the argument's logic, while the *Irrelevant Authority* pattern doesn't fit the sentence since it requires discussion of an unrelated topic, which is not present in the sentence. These subtle distinctions likely make pattern matching more challenging than direct classification because it requires strict structural alignment as well as capturing broader content-related features.

Multistep classification The MULTISTEP approach fails to produce significant results. We conduct this experiment in a single passage to force the model to reason using both semantic and syntactic information. However, classification performance depends critically on the quality of the extracted logical forms, which proves inconsistent and model-dependent. For instance, `o4-mini` embeds classification-relevant contextual informa-

	ARGOTARIO		CLIMATE		COVID	
	Accuracy	**Micro F_1**	**Accuracy**	**Micro F_1**	**Accuracy**	**Micro F_1**
ZERO-SHOT	56.6	56.6	26.4	26.4	29.1	29.1
DEF	57.2	57.3	32.1	32.1	33.3	33.4
LOGICAL FORMS	56.3	56.3	35.7	35.7	**45.8**	**45.9**
PATTERNS	56.3	56.3	35.7	35.7	41.6	41.7
PATTERN MATCHING	58.4	58.4	**38.4**	**38.4**	29.1	29.1
DYNAMIC ONE-SHOT	61.2	61.2	32.1	32.1	33.3	33.3
DYNAMIC + EXP	**62.4**	**62.4**	36.6	36.6	41.6	41.7
DYNAMIC + EXP + PATTERNS	61.5	61.5	36.6	36.6	37.5	37.5

Table 6: Logical fallacy classification performance using `o4-mini` on ARGOTARIO, CLIMATE and COVID, using patterns generated for LOGIC dataset. *No Fallacy* class is included. **Bold**: best approach per dataset.

tion directly into its generated logical forms (Table 11). Furthermore, models demonstrate substantially weaker performance on Group 2 sentences compared to Group 1, showing an average decrease of 25% in F_1 score. Additionally, models frequently bypass the pattern matching phase entirely, arbitrarily assigning matches and corresponding classes despite clear misalignment with the extracted logical forms. For example, given the argument *People nowadays only vote with their emotions instead of their brains* (an instance of *Hasty Generalization*), the model `o4-mini` first extracts the logical form *All A only do B instead of C*. The model then matches this form to the pattern *Generalizing from a small sample or single event to an entire group or population*, which correctly belongs to *Hasty Generalization*. While this produces an accurate classification, the assigned pattern does not precisely correspond to the extracted logical form. In summary, while humans naturally decompose pattern matching into multiple cognitive steps, this multi-stage process proves to be challenging for current LLMs. Models struggle to bridge the gap between abstract logical patterns and their content-dependent instantiations, often failing to identify the implicit premises and unstated logical connections that underlie the reasoning chain.

Method	**Acc.**	**F_1**
`(Jeong et al., 2025)`	49.0	37.0
`o4-mini`	**74.2**	**75.9**
`gpt-4.1-mini`	67.8	68.2
`llama-3.3-70B`	67.5	67.6
`gemma-3-27b-it`	70.5	71.0

Table 7: Comparison of best results per model against the baseline provided by Jeong et al. (2025) (described in Appendix A). **Bold**: best result throughout the whole experimental framework.

Group 1	**Group 2**
• Ad Hominem • Ad Populum • Circular Reasoning • Irrelevant Authority • False Cause • Hasty Generalization • Deductive Fallacy • Black-and-White Fallacy	• Red Herring • Equivocation • Emotional Language • Extension Fallacy • Intentional Fallacy

Table 8: Grouped fallacy classes based on pattern features for analytical purposes.

	Top 1	**Top 2**	**Top 3**	**Top 4**	**Top 5**
F_1	66.7	75.1	81.8	86.5	88.5

Table 9: Performance analysis with expanded solution pool: classification results including top 5 predictions as correct.

6 Evaluation on Further Datasets

In order to further assess the effectiveness of our approach, we conducted a subset of the experiments on additional datasets with varying characteristics and taxonomy of fallacies. Taxonomy alignment is reported in Appendix D.4. Although COVID and CLIMATE are not directly comparable because not all their fallacies are present in LOGIC, the differences are small enough to allow for a qualitative comparison. We have used the model `o4-mini` after it demonstrated the highest performance across all experiments conducted on LOGIC.

Consistent with previous findings, logical pattern incorporation produced superior results for both COVID and CLIMATE datasets, with the model achieving optimal results on CLIMATE specifically through pattern matching (Table 6). However, for ARGOTARIO, there was no observable differ-

I have no intention of stopping the use of somatostatin on patients suffering from acute pancreatitis. I consider it to be a very reasonable choice. After all, it has been standard practice in our department for many years and we've been quite satisfied with the results we've had. ***Irrelevant Authority***	
Top 1	Because many people [do/ believe/support] X, X must be true/good/right/best/valid. (*Ad Populum*)
Top 2	Using [personal trait, experience, past action] as implicit proof of authority on a distinct or unrelated subject. (*Irrelevant Authority*)

Table 10: Sentence accurately classified by `o4-mini` with 2^{nd} ranked pattern. Although the model fails initially, the matching can still be deemed reasonable.

Teaching assistant: the homework assignment was much harder than we thought, so I think we should give a few extra points to students who completed it. Professor: that's a terrible idea. If we give everyone a perfect score for no reason, students won't bother working hard in the future. ***Extension Fallacy***	
`o4-mini`	• **Form:** A proposes: "Because B, we should do X_1." Opponent replies: "If we do X_2 (an exaggerated version of X_1) for no reason, then C will happen; therefore, we must not do X_1." • **Matched pattern:** X proposes Y; response exaggerates Y to an extreme version Y. • **Fallacy:** Extension Fallacy
`gpt-4.1-mini`	• **Form:** A says B was harder than expected, so A suggests giving extra points to those who completed it. B replies: If we give everyone a perfect score for no reason, then C won't work hard; therefore, we should not give extra points. • **Matched Pattern:** Arguing that taking a minor action will cause a chain of escalating events leading to a disastrous outcome, even though no strong evidence supports the inevitability of that chain. • **Fallacy:** Hasty Generalization
`llama-3.3-70B`	• **Form:** If we do A (give everyone a perfect score for no reason), then B (students won't bother working hard in the future). • **Matched Pattern:** Presenting a coincidental or correlated event as proof of causation, often ignoring other plausible causal factors or explanations. • **Fallacy:** False Cause
`gemma-3-27b-it`	• **Form:** If we do A (give extra points), then B (students won't work hard). Therefore, we should not do A. • **Matched Pattern:** If P then Q; Q is true; therefore, P is true. • **Fallacy:** Deductive Fallacy

Table 11: Comparison of outputs of the four models evaluated in **MULTISTEP** configuration.

Hasty Generalization examples	
LOGIC	I know five people from Kentucky. They are all racists. Therefore, Kentuckians are racist.
ARGOTARIO	Is Turkey ready to join the European Union? just do better just goooooooooooooooooooooooooooooooooooooooooo
CLIMATE	Temperatures in the Arctic have soared recently, and scientists are struggling to explain exactly why.
COVID	The Biden administration had to "start from scratch" with a comprehensive COVID-19 vaccine distribution plan because the Trump administration had no working plan.

Table 12: Instances of *Hasty Generalization* across the four datasets.

ence between these logical forms and our LLM-generated patterns. In fact, providing illustrative examples emerged as the most effective approach overall. Unlike LOGIC, these other datasets contain a *No Fallacy* class. Our analysis shows that, when using explicit patterns (especially when the prompt instructs the model to match them) the model tends to incorrectly classify non-fallacious instances as fallacious more frequently compared to the zero-shot setting. This suggests that the presence of explicit patterns may bias the model toward over-identifying fallacies, highlighting its limitations in accurately matching patterns to sentences with a different structure. Our results across all datasets are in line with fine-tuned baselines (Alhindi et al., 2023, 2024; Lei and Huang, 2024), with the notable exception of Jeong et al. (2025), which demonstrates superior performance in an unsupervised way. However, for ARGOTARIO, our pattern-based approach remains below said supervised baselines. It is important to remind that our patterns were specifically tailored to the data distribution of LOGIC and the four datasets exhibit markedly different linguistic and structural characteristics (Table 12). This confirms that LLMs represent a potentially valuable tool for customizing logical forms and reasoning schemes to accommodate diverse data distributions and domain-specific requirements.

7 Conclusions and Future Work

Fallacy detection is an important and complex task to solve. We showed that incorporating logical patterns enhances fallacy detection, marking the first successful application of this approach in an unsupervised setting. We developed an experimental framework that captures both the logical form of fallacies and broader reasoning schemes extracted from fallacious arguments and their explanations.

Our findings consistently show that incorporating information about the underlying logical structure, together with contextual examples, results in state-of-the-art performance for the unsupervised setting on LOGIC (F_1=75.9).

As future work, since the improvement we observe on LOGIC does not directly translate to the other datasets, we plan to implement an adaptive procedure to extract the patterns and we plan to combine our approach with Jeong et al. (2025).

8 Limitations

While this work demonstrates the efficacy of large language models in detecting logical fallacies by exploiting the underlying logical structure of sentences, it has several limitations. First, we intentionally generated patterns exclusively from the LOGIC dataset due to the quality and straightforward structure of its sentences. We are aware, however, that it does not fully cover the complex and multi-faceted spectrum of fallacies. Furthermore, our work is based on a small sample of LLMs. Nevertheless, we selected a diverse and representative subset, including models from different providers, with varying sizes and reasoning capabilities.

9 Ethics Statement

Logical fallacies can reinforce societal bias and facilitate the spread of misinformation, leading to harmful consequences for society. This work focuses on leveraging LLMs for detecting logical fallacies in argumentation and should not be employed to manipulate discourse by exploiting identified reasoning patterns. Furthermore, this approach risks amplifying existing LLM biases, potentially causing unfair detection. We acknowledge these limitations and encourage future bias mitigation research. We are aware of the environmental impact of large-scale LLMs usage. However, this study exclusively employs inference-only methods, significantly reducing computational requirements compared to training approaches. All datasets are used in accordance with their license and they have been checked for personally identifying and offensive content.

References

Tariq Alhindi, Tuhin Chakrabarty, Elena Musi, and Smaranda Muresan. 2023. Multitask instruction-based prompting for fallacy recognition.

Tariq Alhindi, Smaranda Muresan, and Preslav Nakov. 2024. Large language models are few-shot training example generators: A case study in fallacy recognition.

John B. Bacon, Michael Detlefsen, and David Charles McCarty. 1999. *Logic from A to Z: The Routledge Encyclopedia of Philosophy Glossary of Logical and Mathematical Terms*. Routledge.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Irving Marmer Copi, Carl Cohen, and Kenneth McMahon. 1953. *Introduction to Logic*. Macmillan, New York, NY, USA.

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)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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.

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

C.L. Hamblin. 1970. *Fallacies*. University paperbacks. Methuen.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2018. Representation learning on graphs: Methods and applications.

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.

Jiwon Jeong, Hyeju Jang, and Hogun Park. 2025. Large language models are better logical fallacy reasoners with counterargument, explanation, and goal-aware prompt formulation.

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

Ralph Henry Johnson and J. Anthony Blair. 1977. *Logical Self-Defense*. Toronto, Canada.

Yuanyuan Lei and Ruihong Huang. 2024. Boosting logical fallacy reasoning in llms via logical structure tree.

Yanda Li, Dixuan Wang, Jiaqing Liang, Guochao Jiang, Qianyu He, Yanghua Xiao, and Deqing Yang. 2024. Reason from fallacy: Enhancing large language models' logical reasoning through logical fallacy understanding.

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.

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.

Elena Musi, Myrto Aloumpi, Elinor Carmi, Simeon Yates, and Kay O'Halloran. 2022. Developing fake news immunity: Fallacies as misinformation triggers during the pandemic. *Online Journal of Communication and Media Technologies*, 12:e202217.

OpenAI. 2025. Openai o3 and o4-mini system card.

Fengjun Pan, Xiaobao Wu, Zongrui Li, and Anh Tuan Luu. 2024. Are llms good zero-shot fallacy classifiers?

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Devendra Singh Sachan, Yuhao Zhang, Peng Qi, and William Hamilton. 2021. Do syntax trees help pre-trained transformers extract information?

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

Thomas Storer. 1949. Carl g. hempel and paul oppenheim. studies in the logic of explanation. philosophy of science, vol. 15 (1948), pp. 135–175. *Journal of Symbolic Logic*, 14(2):133–133.

Nicole Teo, Donghao Huang, Erik Cambria, and Zhaoxia Wang. 2025. Large language models for logical fallacy detection. In *Trends and Applications in Knowledge Discovery and Data Mining*, pages 387–398, Singapore. Springer Nature Singapore.

Prashanth Vijayaraghavan and Soroush Vosoughi. 2022. TWEETSPIN: Fine-grained propaganda detection in social media using multi-view representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3433–3448, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Zihao Xu, Junchen Ding, Yiling Lou, Kun Zhang, Dong Gong, and Yuekang Li. 2025. Socrates or smartypants: Testing logic reasoning capabilities of large language models with logic programming-based test oracles.

A Baseline

We compare our results against the most recent study that employs unsupervised approaches on LOGIC and provides micro F_1 for comparability:

- Jeong et al. (2025): it provides implicit contextual information such as counterarguments, explanations and goals in the prompt and rank such queries based on confidence score. The reported results was achieved with GPT-4.

B Additional Experiments

We are going to report some other experimental setups that have been explored, including some basic baselines that we have not included in Section 5.

B.1 Prompt design

- **EXP**: to investigate whether explicit reasoning improves performance, we implemented a baseline that not only provides fallacy names but also requests the model to generate a two-sentence explanation for its classification decision. This approach tests whether forcing the model to articulate its reasoning leads to better outcomes and assess the model's actual comprehension of fallacious arguments.

- **GUIDELINES**: to leverage the model's classification errors for improvement, we develop guidelines derived from observed mistakes. We conduct pattern matching evaluation on the validation set, then systematically collect misclassified instances. For each class, we provide the model with its incorrectly classified examples and prompt it to generate

Fallacy	Irrelevant Authority
Core definition	A fallacy that treats an individual's status, title, or popularity as proof of a claim when their expertise or relevance to the topic is absent or insufficient.
Key indicators	Argument rests on "X says so" without independent support. Authority cited has no recognized expertise in the claim's domain. No substantive evidence beyond the authority's endorsement.
Typical confusion patterns	Ad Populum: group popularity vs. single authority endorsement. Appeal to Tradition: 'has always been done by experts' vs. citing irrelevant experts. Equivocation: shifting word senses vs. relying on irrelevant credentials.

Table 13: Guideline relative to *Irrelevant Authority* fallacy generated by `o4-mini`.

Method	o4-mini		gpt-4.1-mini		llama-3.3-70B		gemma-3-27b-it	
	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1
EXP	61.3	61.5	57.5	57.9	56.1	56.5	59.1	60.8
GUIDELINES	65.5	65.7	60.5	60.6	52.8	53.3	58.8	59.4

Table 14: Logical fallacy classification performance. F_1 means Micro F_1.

comprehensive detection guidelines, given our generated pattern as reference. These guidelines include a core definition, key identifying features, common confusion patterns with similar fallacies, and a practical checklist to aid in detecting the specific type of fallacious reasoning, as you can see from table 13. These guidelines are then adopted to evaluate the test set. Notably, while all guidelines were generated from misclassified patterns, only those produced by `o4-mini` and partially by `gpt-4.1-mini` incorporate a little structural and logical information such as common connectors or logical forms. The majority of guideline content across models focuses primarily on semantic characteristics rather than structural patterns.

B.2 Results

EXP's results show that requesting the model to articulate the reasoning does not really cause any improvement. Specifically, certain classes such as *Intentional Fallacy* and *Extension Fallacy* exhibit extremely low F_1 scores under the non-reasoning models (0.04 and 0.081 respectively on average), indicating performance deterioration compared to the **ZERO-SHOT** baseline. This proves that models process surface-level semantic patterns without being able to access the multi-layered intentional structures behind reasoning.

Including **GUIDELINES** yields only modest results. While these guidelines are designed to provide comprehensive fallacy knowledge, they appear to lack the appropriate type of information from which models can benefit. Indeed, providing explicit information about the underlying logical structure proves significantly more beneficial for model performance.

C Syntax-augmented roBERTa

Sachan et al. (2021) introduces a syntax-augmented model that incorporates dependency tree information into pre-trained BERT-based (Devlin et al., 2019) transformers through specialized Graph Neural Networks (GNNs) (Hamilton et al., 2018) that process dependency trees. The authors introduce two distinct fusion strategies to integrate syntactic structure into BERT representation. We adopted specifically roBERT-large (Liu et al., 2019) in the attempt to perform a syntax-driven examples selection. Further details about the implementation are available in Sachan et al. (2021).

D Fallacy Datasets

D.1 Logic

The dataset LOGIC (Jin et al., 2022) contains the following 13 fallacy classes: *Faulty Generalization (Hasty Generalization)*, *Ad Hominem*, *Ad Populum*, *Circular Claim (Circular Reasoning)*, *False Cause (False Causality)*, *Appeal to Emotion (Emotional Language)*, *Fallacy of Relevance (Red Herring)*, *Deductive Fallacy*, *Intentional Fallacy*, *Fallacy of Extension (Extension Fallacy)*, *False Dilemma (Black-and-White Fallacy)*, *Fallacy of Credibility (Irrelevant Authority)* and *Equivocation*. The names in the parentheses are the actual names used in our experiments.

D.2 Argotario

The dataset ARGOTARIO contains 5 fallacy classes: *Ad Hominem*, *Appeal to Emotion (Emotional Language)*, *Hasty Generalization*, *Irrelevant Authority* and *Red Herring*. The names in the parentheses are the actual names used in our experiments. It contains the class *No Fallacy* as well.

D.3 Covid and Climate

The datasets COVID (Musi et al., 2022) and CLIMATE (Alhindi et al., 2023) contain the following 10 fallacy classes: *Evading the Burden of Proof*,

Text	Generated explanation	Gold label
The Bible is true because God exists, and God exists because the Bible says so.	The argument uses its conclusion as a premise, claiming the Bible is true because God exists and God exists because the Bible says so. This creates a logical loop without providing independent evidence for either claim. ***Circular Reasoning***	***Circular Reasoning***
My friend said that if you sneeze more than three times, you have the corona virus.	The argument assumes that sneezing more than three times directly indicates having the corona virus based on insufficient evidence. It generalizes a specific symptom without considering other possible causes or medical diagnosis. ***Hasty Generalization***	***Irrelevant Authority***

Table 15: Examples from `GPT-4.1-mini` in **EXP** setting: the first sentence is correctly classified with a well-grounded explanation; the second argument is misclassified because, while its explanation appears coherent in isolation, it fails to capture the underlying fallacious reasoning.

Cherry Picking, *Red Herring*, *Strawman (Extension Fallacy)*, *False Authority (Irrelevant Authority)*, *Hasty Generalization*, *False Cause*, *Post Hoc*, *False Analogy (Deductive Fallacy)* and *Vagueness*. It contains a *No Fallacy* class. The names in the parenthesis are the actual names used in our experiments.

D.4 Taxonomy alignment

The dataset ARGOTARIO contains exactly a subset of LOGIC's fallacy classes, so the learned knowledge is easily transferable. For what concerns COVID and CLIMATE, the class *Cherry Picking* is excluded altogether, whereas *Post Hoc* and *False Cause* are fairly represented by LOGIC's broader *False Cause* category that contains instances of both. Similarly, the class *Evading the Burden of Proof* is included in LOGIC's *Intentional Fallacy* class and has a dedicated logical form within our framework. Same applies for the class *False Analogy* which is included in LOGIC's *Deductive Fallacy*. The class *Vagueness* is associated with LOGIC's *Equivocation*.

E LLM-derived Patterns and Definitions

The logical patterns presented in Table 17 were extracted following the process described in Section 4 from LOGIC. These patterns form the basis of our evaluation of how structural features contribute to LLMs' performance in logical fallacy detection, assessing the role of logical patterns in model reasoning capabilities. Table 18 illustrates the LLM-made fallacies' definitions used in the experimental setting **NEW DEF**.

F Prompts Templates

F.1 Pattern Generation

Step 1 You will be given a fallacious argument and the name of the logical fallacy it contains. Your task is to explain what is happening in the argument and why it is fallacious. Do not include definitions, labels or general commentary: focus only on describing the flaw in reasoning specific to the example in a concise way.

Step 2 You will be given a list of arguments containing a {fallacy_name} fallacy and an explanation of why it is fallacious. Your task is to provide the following information, returning a JSON object with the following fields:
{{
"summary": Write a concise summary (max 2 sentences) that captures the common logical pattern behind these explanations. The summary should start with the name of the fallacy.
"syntactic_patterns": Identify common syntactic or structural patterns in how the arguments are phrased.
- Derive the abstract logical structure following formal logic principles. Use abstract placeholders like A, B, C to replace specific nouns or phrases, but ensure the pattern closely mirrors the logical structure and progression of the original sentence.
- Find recurring sentence structures, phrases, or ways in which the fallacious reasoning is introduced, including typical linguistic markers and illustrative cases that signal the flawed reasoning.
}}
Your explanation should help someone recognize new examples of {fallacy_name} by highlighting the shared reasoning mistake across all cases.

F.2 Classification

`system_prompt`: You are a logical reasoning expert. Your task is to carefully examine the given argument and classify it into one of the following classes: {fallacies }.

- **ZERO-SHOT**: Given an argument, classify the fallacy it contains. Choose one of the following labels: {fallacies}. Respond only with the

Fallacy	Definition	Logical Form
Ad Hominem	The text attacks a person instead of arguing against the claims.	Person 1 is claiming Y. Person 1 is a moron. Therefore, Y is not true.
Ad Populum	The text affirms something is true because the majority thinks so.	A lot of people believe X.Therefore, X must be true.
Black-and-White Fallacy	The text presents two alternative options as the only possibilities yet more exist.	Either X or Y is true.
False Cause	The text assumes two correlated events must also have a causal relation.	X occurred after Y. Therefore, Y caused X (although X was also a result of A,B,C,... etc)
Circular Reasoning	The text tries to prove a point by simply repeating the point in different words.	X is true because of Y. Y is true because of X.
Deductive Fallacy	The text presents a conclusion that doesn't logically follow from the premises.	If A is true, then B is true. B is true. Therefore, A is true.
Emotional Language	The text arouses non-rational emotions.	Claim X is made without evidence. In place of evidence, emotion is used to convince the interlocutor that X is true.
Equivocation	The text uses a key term in multiple senses, leading to ambiguous conclusions.	Term X is used to mean Y in the premise. Term X is used to mean Z in the conclusion.
Extension Fallacy	The text attacks an exaggerated version of the opponent's claim.	Person 1 makes claim Y. Person 2 restates person 12̆019s claim (in a distorted way). Person 2 attacks the distorted version of the claim. Therefore, claim Y is false.
Hasty Generalization	The text draws a broad conclusion based on a limited sample.	Sample S is taken from population P. Sample S is a very small part of population P. Conclusion C is drawn from sample S and applied to population P.
Intentional Fallacy	The text relies on the author's intent instead of focusing on the meaning within the text itself.	Person 1 knows claim X is incorrect. They still claim that X is correct using an incorrect argument
Irrelevant Authority	The text cites an authority, but the authority lacks relevant expertise.	According to person 1, who is an expert on the issue of Y, Y is true. Therefore, Y is true.
Red Herring	The text introduces an irrelevant topic to divert attention from the main argument.	Argument A is presented by person 1. Person 2 introduces argument B. Argument A is abandoned.

Table 16: LOGIC class taxonomy: class names, definitions and logical form representations used in baseline experiments.

name of the fallacy, with no additional text.
Argument: {text}
Fallacy:

- **EXP**: Given an argument, classify the fallacy it contains. Choose one of the following labels: {fallacies}. Respond in the format:
  Fallacy: [fallacy label]
  Reasoning: [brief explanation justifying the choice]
  The reasoning must be exactly two sentences long.
  Argument: {text}
  Fallacy:
  Reasoning:

- **DEF**, **NEW DEF**: Given an argument, classify the fallacy it contains. Choose one of the following labels: {fallacies}. Use the following definitions to guide your classification: {definitions}. Respond only with the name of the fallacy, with no additional text.
  Argument: {text}
  Fallacy:

- **LOGICAL FORMS**, **PATTERNS**: Given an argument, your task is to classify the type of logical fallacy it contains. Choose one of the following labels: {fallacies}. To assist you in this task, common patterns associated with each fallacy are also provided. {patterns}. Carefully compare the argument to these patterns and select the fallacy that best matches. Provide only the name of the fallacy.
  Argument: {text}
  Fallacy:

- **PATTERN MATCHING**: Given an argument, your task is to classify the type of logical fallacy it contains. Choose one of the following labels: {fallacies}. To assist you in this task, you are provided with common reasoning patterns associated with each fallacy: {patterns}
  Return:
  - The specific pattern that best matches the logical reasoning in the argument.
  - The name of the fallacy.
  Don't add any additional text.
  Argument: {text}

Fallacy:

- **ONE-SHOT, DYNAMIC ONE-SHOT, SYNTAX-BASED DYNAMIC ONE-SHOT**: Given an argument, your task is to classify the type of logical fallacy. Choose the correct fallacy from the following list: {fallacies}. Use the following examples to guide your classification: {examples}
  Argument: {text}
  Fallacy:

- **ONE-SHOT + EXP, DYNAMIC + EXP**: Given an argument, your task is to classify the type of logical fallacy. Choose the correct fallacy from the following list: {fallacies}. Use the following examples to guide your classification. Each example includes both the fallacy and a brief explanation of why it applies: {examples}
  Argument: {text}
  Fallacy:

- **DYNAMIC + EXP + PATTERNS**: Given an argument, your task is to classify the type of logical fallacy in a given text. Choose the correct fallacy from the following list: {fallacies}. Use the examples below to guide your classification — each example includes both the fallacy and a brief explanation of why it applies. {definitions} You will be provided with a list of common patterns associated with each fallacy. {patterns} Carefully compare the argument to these patterns and select the fallacy that best matches. Provide only the name of the fallacy.
  Argument: {text}
  Fallacy:

- **GUIDELINES**: Given an argument, your task is to classify the type of logical fallacy. Choose one of the following: {fallacies}. To assist you in this task, you are provided with useful guidelines. These include typical reasoning patterns for each fallacy, common mistakes that often lead to misclassification and a quick practical checklist for classification: {guidelines}. Return only the name of the fallacy.
  Argument: {text}
  Fallacy:

- **MULTISTEP**: Given an argument, your task is to process it in three steps:
  Step 1 (Form Extraction):
  When given a sentence, extract its underlying syntactic logical form. Use abstract placeholders like A, B, C to replace specific nouns or phrases, but ensure the pattern closely mirrors the logical structure and progression of the original sentence. Preserve logical connectors and adverbs such as ‘therefore’, ‘because’, ‘if’, ‘then’, etc. The goal is to abstract away from surface wording while preserving the sentence’s original reasoning form.
  Step 2 (Pattern Matching): From the list of known fallacy patterns provided below, return the one that most closely matches the pattern you extracted in Step 1.
  Step 3 (Fallacy Classification): Based on the extracted and matched pattern, classify the argument into one of the logical fallacy types from the list: {fallacies}
  Choose the most appropriate category based on the structure of the reasoning.
  Use the following reasoning patterns as reference: {patterns}
  Analyze the following argument: {text}
  Step 1 (Form Extraction):
  Step 2 (Matching Pattern):
  Step 3 (Fallacy):

Fallacy class	Patterns
Deductive Fallacy	• The argument assumes that because X is true, Y must also be true, without establishing a necessary connection between X and Y. • Because X shares a characteristic with Y, therefore Y must also have characteristic Z (unique to X). • If P then Q; Q is true; therefore, P is true. • All A are B; all B are C; therefore, all C are A. • The argument compares X to Y as if they are equivalent, ignoring relevant differences.
Ad Hominem	• Dismisses someone's argument by accusing the opponent of similar behavior, avoiding the argument itself. • Argues that because X has characteristic Y, X's views or claims must be invalid/false. • Uses a personal insult or irrelevant fact about X to discredit X without addressing the core issue. • Focuses on unrelated personal factors (e.g., age, profession, habits) to attack the person instead of the argument.
Emotional Language	• Appeals that highlight personal circumstances or potential consequences without addressing the core issue, e.g., 'I haven't done X, but... [appeal to emotion]'. • Use of emotionally charged or loaded terms in place of neutral language, e.g., calling something an 'outrage', 'dangerous militants', or using phrases like 'taking our freedom away'. • Rhetorical questions or statements designed to evoke feelings of guilt, sympathy, or fear, e.g., 'If we don't do X, disaster Y will happen'. • Evocation of pity or sympathy to distract from the logical evaluation of claims, e.g., 'I studied during my grandmother's funeral'. • Use of vivid imagery or emotionally provocative examples to bypass critical analysis, e.g., showing suffering animals or invoking dramatic suffering stories.
False Cause	• Inferring causation from correlation expressed as 'X happened when Y happened' or 'Every time X occurs, Y follows,' without evidence of a causal link. • Attributing a complex outcome to a single factor due to temporal proximity or repeated coincidence (e.g., 'Because of X, Y happened,' ignoring other influences). • Using phrases implying causality based on timing, such as 'therefore', 'must have caused', 'is the reason', or 'is the cause of', without supporting evidence. • Presenting a coincidental or correlated event as proof of causation, often ignoring other plausible causal factors or explanations. • Statements that simplify multi-factor phenomena to a single cause (e.g., 'Because of single action/event X, complex result Y occurred').
Irrelevant Authority	• The argument assumes that because [authority figure] says/believes [X], therefore [X] must be true/reliable/effective. • Relying on the opinion or endorsement of [famous/unqualified person] outside their field of expertise to support [claim/conclusion]. • Because [person/role/title] holds a position or is respected, their statement on [irrelevant topic] is presented as evidence. • Using [personal trait, experience, past action] as implicit proof of authority on a distinct or unrelated subject. • The argument presents [authority]'s position as the sole justification without providing independent reasons or evidence.
Extension Fallacy	• X proposes Y; response exaggerates Y to an extreme or total version (e.g., 'So you want to [extreme claim]?') • Assuming that because someone holds a position on A, they must also hold an extreme or unrelated position B (e.g., 'If you believe A, then you must believe B') • Misinterpreting a specific statement as implying a much broader or more negative attitude (e.g., 'Because you said X, you must hate Y') • Responding to a moderate or specific claim by substituting a more extreme or absurd claim that is easier to criticize (e.g., 'You think that... [absurd extension]') • Framing a preference or partial stance as a wholesale endorsement or rejection of a related but distinct issue (e.g., 'Preferring X means you hate Y')
Hasty Generalization	• Because a particular instance or individual showed Z, therefore all instances or individuals must be Z. • Arguing that taking a minor action will cause a chain of escalating events leading to a disastrous outcome, even though no strong evidence supports the inevitability of that chain. • If we allow [event A], then [event B] will happen, then [event C], and eventually [event Z] will occur—so we must not allow [event A]. • Generalizing from a small sample or single event to an entire group or population. • Jumping from some instances to 'everyone' or 'all' without acknowledging exceptions or diversity.
Equivocation	• The reasoning equivocates by shifting from a literal to a figurative meaning (or vice versa) of a term to create a false equivalence. • One premise employs TERM with one definition/context, while another premise or conclusion uses the same TERM but with a distinctly different meaning, unacknowledged by the arguer.
Intentional Fallacy	• The argument assumes that because X (e.g., someone's intention, belief, or lack of counter-evidence), therefore Y is true. • Asserting P is true because it has not been disproven. • Because the creator intended [interpretation], the work should be understood as [interpretation]. • Questions framed to presuppose guilt or a specific intention (e.g., 'Have you stopped X?'), thus assuming what is to be proven. • If A does not have trait X, and X is (allegedly) typical of group G, then A is not a member of G.
Ad Populum	• Because many people [do/ believe/support] X, X must be true/good/right/best/valid. • Everyone/Most people [do/believe] X, so you/one should do X too. • The popularity of X is used as evidence for X's quality, validity, or truth, rather than providing objective reasons. • Appealing to the desire to belong to a group, suggesting that conformity implies correctness or value. • Using phrases like 'everyone knows,' 'the majority thinks,' 'most people do,' to justify a conclusion without addressing actual evidence.
Red Herring	• Instead of addressing [original issue], the argument shifts focus to [irrelevant topic], which distracts from the main discussion. • The argument attempts to justify/explain/defend by referencing [irrelevant detail], ignoring the original issue of [main topic]. • A shift from the initial question or problem to a secondary topic that does not logically follow, e.g., 'You asked about X, but I will tell you about Y.'
Black-and-White Fallacy	• Either [option A] or [option B], with no other alternatives considered. • You are either [extreme position A] or [extreme position B]. • You are either with me or against me. • [Action A] or else [negative consequence], ignoring intermediary options.
Circular Reasoning	• X is true/better/good because X is true/better/good. • X is Y because X has property Y (where property Y is essentially restating X). • "Because" + restatement or synonymous phrasing of the claim as the reason. • The argument claims/assumes X to prove/justify X.

Table 17: List of logical patterns extracted in our LLM-based experimental framework from LOGIC dataset.

Fallacy	LLMs-derived definition
Ad Hominem	Ad Hominem occurs when an argument targets a person's character or traits instead of engaging with the actual issue or evidence presented.
Ad Populum	Ad Populum occurs when a claim is deemed true or good simply because many people believe or endorse it, without examining the actual reasoning.
Black-and-White Fallacy	Black-and-White Fallacy occurs when only two extreme options are presented, ignoring the existence of middle ground or alternative solutions.
False Cause	False Cause occurs when a causal relationship is assumed based on correlation alone, without sufficient evidence or consideration of other factors.
Circular Reasoning	Circular Reasoning occurs when the conclusion is assumed in the premises, creating a loop that provides no independent support for the argument.
Deductive Fallacy	Deductive Fallacy occurs when conclusions do not logically follow from the premises, often due to assuming unsupported relationships, oversimplifying, misapplying analogies, or improperly reversing conditions.
Emotional Language	Emotional Language occurs when persuasion relies on appeals to emotion rather than logical reasoning or factual evidence.
Equivocation	Equivocation occurs when a key term is used ambiguously in an argument, shifting meaning and creating an illusion of logical connection.
Extension Fallacy	Extension Fallacy occurs when an argument exaggerates or distorts an opponent's claim to make it easier to attack, rather than addressing the actual position.
Hasty Generalization	Hasty Generalization occurs when a broad conclusion is drawn from an insufficient or unrepresentative sample of evidence.
Intentional Fallacy	Intentional Fallacy occurs when arguments are judged based on the speaker's intentions or characteristics rather than the content and evidence or when asserting that something is true only because it has not been disproven.
Irrelevant Authority	Irrelevant Authority occurs when an argument cites an authority whose expertise is not relevant to the subject matter being discussed.
Red Herring	Red Herring occurs when attention is diverted from the main issue by introducing irrelevant or emotionally charged distractions.

Table 18: List of definitions extracted in our LLM-based experimental framework from LOGIC dataset.