Large Language Models are Better Logical Fallacy Reasoners with Counterargument, Goal, and Explanation-aware Prompt Formulation

Anonymous ACL submission

Abstract

The advancement of Large Language Models (LLMs) like GPT-4 has significantly enhanced our capability to process complex language. However, accurately detecting and classifying logical fallacies—a crucial aspect of reasoning and argumentation—remains a challenging task. This study introduces a simple but powerful prompt formulation approach that can be leveraged for both zero-shot settings and fine-tuned models. Our proposed method formulates an input prompt by enriching the input text in view of counterarguments, explanations, and goals. The formulated prompts are used for providing an answer in the zero-shot setting or integrated into the training of existing Small Language Models (e.g., RoBERTa). Our experiments span diverse datasets, featuring 5 to 13 types of logical fallacies, to assess the method's robustness and adaptability with *GPT-3.5-turbo* and *GPT-4.0*, placing a particular emphasis on the impact of various query types. The findings reveal significant improvements across the board: for zero-shot settings, the method increased the Macro F1-score by up to 0.20 in detection tasks, while in multiclass classification tasks involving fine-tuned models, the Macro F1-score saw enhancements of up to 0.56.

1 Introduction

In recent years, the field of Natural Language Processing (NLP) has witnessed remarkable growth, propelled by the advent of Large Language Models (LLMs) like GPT-4 (Brown et al., 2020; Achiam et al., 2023). Despite these strides, accurately identifying and classifying logical fallacies—a prevalent challenge across various forms of discourse—remains a significant hurdle. These fallacies, common in casual conversations, formal debates, and educational texts, underscore the complexities of human thought and language in the realm of automated reasoning and analysis (Habernal et al., 2017; Petric, 2020; Goffredo et al., 2022; Jin et al., 2022; Goffredo et al., 2023) Recent efforts to detect and classify logical fallacies have incorporated prompting techniques (Hong et al., 2023) and methodologies like Case-Based Reasoning (CBR) (Sourati et al., 2023). While (Hong et al., 2023) has sought to refine logical fallacy analysis by providing definitions for various fallacy types through prompts, such strategies primarily leverage direct textual interactions with models. Concurrently, (Sourati et al., 2023) enhances logical fallacy analysis using Case-Based Reasoning (CBR) and enriched representations. However, it primarily focuses on matching within a structured framework, which may underutilize the generative capabilities of Large Language Models (LLMs) for a deeper, dissected examination of fallacies. As shown in Figure 1, our proposed method obtains enriched representations of an input text to formulate prompts for LLMs, such as gpt-3.5-turbo-instruct, to gather contextualized information potentially useful for analyzing logical fallacies. Subsequently,

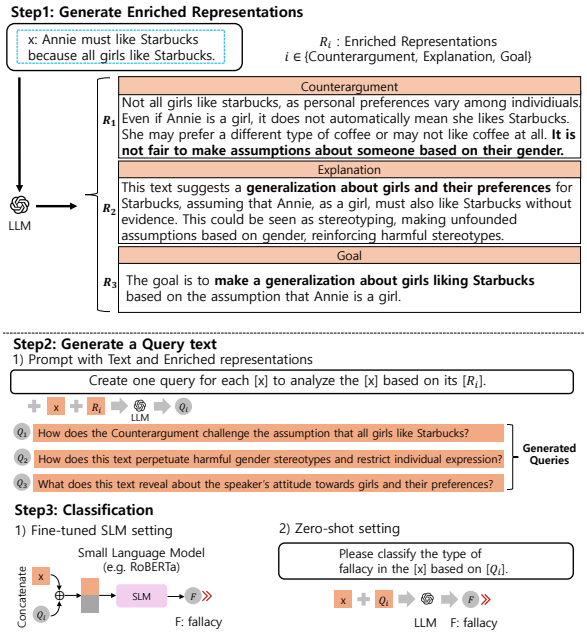


Figure 1: Example of prompt formulation (x: Input text to classify)

our model reformulates the representations for getting a query. Additionally, we classify labels of logical fallacies in 1) a supervised learning or 2) zero-shot setting using a RoBERTa-base model. The approach demonstrates significant performance improvements across diverse datasets.

Our contributions are twofold:

1. **Advanced Prompt Formulation**: Our approach utilizes contextualized representation to enrich the analysis of logical fallacies, guiding LLMs like *GPT-3.5-turbo* and *GPT-4.0* through a sophisticated querying strategy. This method enhances the LLMs' engagement with texts, allowing them to uncover and evaluate logical fallacies more accurately.
2. **Comprehensive Fallacy Analysis**: Our method enhances logical fallacy detection and classification across diverse datasets, proving effective in both zero-shot settings and with fine-tuned models. This demonstrates the adaptability and effectiveness of our querying approach in improving analysis precision.

2 Related Work

Logical fallacies, pervasive across various forms of discourse, compromise argumentative quality and reasoning accuracy. Historical and computational studies within informal logic traditions and critical discussion rules (Hansen, 1996; Van Eemeren et al., 2002; Tindale, 2007; Damer, 2008) stress the importance of recognizing these fallacies in domains such as public policy, legal reasoning, and scientific discourse (Bailin and Battersby, 2016). Computational research has ventured into fallacy detection in dialogues (Habernal et al., 2017), argument sufficiency (Stab and Gurevych, 2017; Wachsmuth et al., 2017), Reddit discussions (Sahai et al., 2021), and misinformation contexts (Musi et al., 2022), often with a focus on singular datasets, limiting generalizability. Conversely, broader methodologies, like employing T5 for multiple fallacy datasets (Alhindi et al., 2023) and exploring Case-Based Reasoning (CBR) for new classification scenarios (Sourati et al., 2023), seek expansive applicability. Such methodologies seek to overcome the limitations of previous approaches and offer new directions for understanding and managing complex logical fallacies. (Hong et al., 2023) leveraged prompts to define various fallacy types, seeking to deepen models' analytical capabilities. Yet, such methods, primarily relying on direct, definitional prompts, may not fully engage the model in nuanced understanding. Furthermore, (Sourati et al., 2023)'s focus on textual modifications or architectural changes through CBR might not adequately address the complexities of logical fallacies. Building on this foundation, our study employs enriched case representations to generate queries for deeper interaction with logical fallacies across multiple fallacy datasets, aiming to bridge understanding gaps highlighted by (Field, 1977) and enhance LLMs' reasoning capabilities. Addressing these gaps, our research introduces a simple yet powerful methodology that goes beyond conventional frameworks by leveraging prompt formation based on enriched case representations.

Table 1: Summary of four fallacy datasets. N represents the number of samples, C represents the number of classes. † indicates the numbers include the No Fallacy class.

Data	N	C	Genre	Domain
ARGOTARIO	1338	6†	Dialogue	General
LOGIC	2449	13	Dialogue	Education
COVID-19	154	11†	SocMed/News	Covid-19
CLIMATE	685	11†	News	Climate

3 Approach

3.1 Data

Our study analyzes four distinct datasets, excluding the Propaganda dataset due to its unique annotation approach. These datasets are: (1) **ARGOTARIO**, highlighting six types of fallacies in QA pairs. (2) **LOGIC**, identifying 13 logical fallacies from educational content. (3) **CLIMATE**, analyzing fallacies in 778 climate change article segments. (4) **COVID-19**, focusing on 11 fallacies in pandemic-related fact-checked content. Each dataset, detailed in (Alhindi et al., 2023), spans 18 fallacy types across various domains and genres, offering a comprehensive overview for our logical fallacy analysis (see Table 1).

3.2 Methods

3.2.1 Step1: Generating Enriched Representations

Our approach utilizes Large Language Models (LLMs) to enrich the analysis of texts containing logical fallacies (x) through three distinct perspectives: Counterargument, Explanation, and Goal. Each perspective is denoted by an index i, where i represents the specific instruction applied ($\mathcal{I}^1$ for Counterargument, $\mathcal{I}^2$ for Explanation, and $\mathcal{I}^3$ for Goal). These instructions guide the LLM to focus

on various aspects of the argument, leading to the generation of enriched representations $\mathcal{R}_i(x)$. By applying specific instructions $\mathcal{I}^i$ to *GPT-3.5-turbo-instruct*, we generate these enriched representations that delve into the fallacy's multifaceted nature:

$$\mathcal{R}_i(x) = \text{LLM}(x, \mathcal{I}^i), \quad (1)$$

Consider, for instance, a statement x = *"Annie must like Starbucks because all girls like Starbucks."* In analyzing this statement through the **Goal** perspective, we prompt the LLM with *"Express the goal of the text"*, leading to $\mathcal{R}_i(x)$ = *"The goal of this text is to make a generalization about girls liking Starbucks based on the assumption that Annie is a girl."* This enriched representation is used for generating a query.

3.2.2 Step2: Generating Queries

To provide a context-driven query, we design a tailored query generation method $\mathcal{Q}_i(\mathcal{R}_i(x), x)$ using the LLM instruction: *"Create one query for each text to analyze the text based on its goal."* This process yields $\mathcal{Q}_i(\mathcal{R}_i(x), x)$ = *"What does this text reveal about the speaker's attitude towards girls and their preferences?"* Such queries enable a nuanced understanding of the underlying assumptions and biases in arguments, demonstrating the model's capability to engage critically with the content:

$$Q_i(R_i(x), x) = \text{LLM}(\mathcal{R}_i(x), \mathcal{Q}^i) \quad (2)$$

3.2.3 Step3: Detecting and Classifying Logical Fallacy

The comprehensive analysis facilitated by these tailored queries allows for getting prediction probability p_{LLM} of the fallacy's class label (l) with each enriched representation i for an input text x:

$$\textit{Fallacy}(i) = \arg\max_{l \in L} p_{\text{LLM}}(l|x, Q_i(\mathcal{R}_i(x), x)), \quad (3)$$

where L denotes a set of logical fallacy labels. By embedding enriched representations and tailored reformulated queries, our approach enables classification. Additional examples and details are available in Table 9 in the Appendix. For detailed instructions on our prompt formulation and application, refer to the Appendix C.1 C.2 C.3.

4 Experiments

Overview In this study, we evaluated the capabilities of the *GPT-3.5-turbo* and *GPT-4.0* models in detecting and classifying logical fallacies across various datasets. We utilized a zero-shot learning framework for detection tasks against a *No Fallacy* baseline and engaged both zero-shot settings and fine-tuning techniques for multiclass fallacy classification. For full classification results, please refer to the appendix A.

Table 2: Accuracy and Macro F1 score of the Binary class fallacy detection on all datasets with different formulated prompts: CG for *Counterargument*, EX for *Explanation*, and GO for *Goal*. **Bold** scores indicate the highest score for each metric.

Model (Parameters)	Argotario		COVID-19		CLIMATE	
	Acc	F1	Acc	F1	Acc	F1
GPT-3.5-turbo	0.69	0.52	0.46	0.32	0.52	0.49
GPT-3.5-turbo + CG Formulation	0.70 (+0.01)	0.50 (-0.02)	0.54 (+0.08)	0.35 (+0.03)	0.38 (-0.14)	0.35 (-0.14)
GPT-3.5-turbo + EX Formulation	0.69 (+0.00)	0.47 (-0.05)	0.58 (+0.12)	0.48 (+0.16)	0.35 (-0.17)	0.32 (-0.17)
GPT-3.5-turbo + GO Formulation	0.69 (+0.00)	0.49 (-0.03)	0.54 (+0.08)	0.41 (+0.09)	0.44 (-0.08)	0.42 (-0.07)
GPT-4.0	0.76	0.69	0.62	0.55	0.48	0.48
GPT-4.0 + CG Reformulation	0.73 (-0.03)	0.65 (-0.04)	**0.77** (+0.05)	**0.75** (+0.20)	0.59 (+0.11)	0.59 (+0.11)
GPT-4.0 + EX Formulation	0.74 (-0.02)	0.65 (-0.04)	0.73 (+0.11)	0.71 (+0.16)	0.59 (+0.11)	0.59 (+0.11)
GPT-4.0 + GO Formulation	**0.81** (+0.05)	**0.76** (+0.07)	0.65 (+0.03)	0.63 (+0.08)	**0.60** (+0.12)	**0.60** (+0.12)

Detection and Classification Results As in Table 2, specialized prompt formulation strategies substantially improve the detection and classification of fallacies across diverse datasets. Particularly, employing the *Goal Formulation* method with the GPT-4.0 model on the **ARGOTARIO** yields the best results for binary class fallacy detection, achieving an accuracy of 0.81 and a Macro F1 score of 0.76. The same approach also leads to identical scores on the **CLIMATE**. For the **COVID-19**, the *Counterargument Formulation* with GPT-4.0 was most effective, showing accuracy and Macro F1 scores of 0.77 and 0.75, respectively. Table 3 shows the results of multi-class fallacy classification. In the table, *Explanation Formulation* combined with the RoBERTa model demonstrated superior performance, especially on the **ARGOTARIO** with an accuracy of 0.81 and a Macro F1 of 0.80, and the **CLIMATE**, achieving 0.84 in accuracy and 0.74 in Macro F1. These findings underscore the pivotal role of tailored query formulations in enhancing model efficacy for fallacy detection and classification. Notably, recent models, such as *Electra-StructureAware* and ELECTRA(CBR) are worse in the LOGIC dataset despite a similar number of parameters.

Table 3: Accuracy and Macro F1 score of the Multi class fallacy classification on all datasets with different reformulated texts: CG for *Counterargument*, EX for *Explanation*, and GO for *Goal*. **Bold** scores indicate the highest score for each metric. We utilized both fine-tuned models and prompt-based models, with the prompt-based models being evaluated in a zero-shot manner. Results for Electra-Structure-Aware and ELECTRA are from the original papers, and the same test splits are used for comparison. A dash ('-') indicates the absence of known comparison results.

Model (Parameters)	ARGOTARIO		LOGIC		COVID-19		CLIMATE	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Learning-based model								
Electra-Structure-Aware (110M)	-	-	0.48	0.59	-	-	-	-
ELECTRA (CBR) (110M)	-	-	-	0.66	-	-	-	-
RoBERTa (125M)	0.59	0.60	0.65	0.62	0.36	0.14	0.26	0.18
RoBERTa + CG (125M)	0.56 (-0.03)	0.57 (-0.03)	0.53 (-0.12)	0.42 (-0.20)	0.36 (-0.07)	0.16 (+0.02)	0.39 (+0.13)	0.30 (+0.12)
RoBERTa + EX (125M)	0.81 (+0.22)	**0.80** (+0.20)	**0.80** (+0.15)	**0.79** (+0.27)	0.36 (+0.00)	0.17 (+0.03)	**0.84** (+0.58)	**0.74** (+0.56)
RoBERTa + GO (125M)	0.79 (+0.20)	0.78 (+0.18)	0.71 (+0.06)	0.71 (+0.09)	0.43 (+0.00)	0.28 (+0.14)	0.61 (+0.35)	0.5 (+0.32)
Zero-shot model								
GPT-3.5-turbo (20B)	0.50	0.41	0.39	0.25	0.29	0.14	0.18	0.13
GPT-3.5-turbo + CG (20B)	0.5 (+0.00)	0.41 (+0.00)	0.4 (+0.01)	0.25 (+0.00)	0.43 (+0.14)	0.29 (+0.15)	0.63 (+0.45)	0.54 (+0.41)
GPT-3.5-turbo + EX (20B)	0.70 (+0.20)	0.67 (+0.26)	0.46 (+0.07)	0.32 (+0.07)	0.57 (+0.28)	0.44 (+0.30)	0.7 (+0.52)	0.65 (+0.52)
GPT-3.5-turbo + GO (20B)	0.79 (+0.29)	0.78 (+0.37)	0.71 (+0.32)	0.71 (+0.46)	0.43 (+0.14)	0.28 (+0.14)	0.61 (+0.43)	0.5 (+0.37)
GPT-4.0	0.60	0.50	0.40	0.28	0.43	0.33	0.18	0.11
GPT-4.0 + CG	0.61 (+0.01)	0.51 (+0.01)	0.4 (+0.00)	0.27 (-0.01)	0.57 (+0.14)	0.40 (+0.07)	0.33 (+0.15)	0.26 (+0.15)
GPT-4.0 + EX	**0.84** (+0.24)	0.70 (+0.20)	0.48 (+0.08)	0.35 (+0.07)	**0.71** (+0.28)	**0.68** (+0.35)	0.67 (+0.45)	0.66 (+0.55)
GPT-4.0 + GO	0.79 (+0.19)	0.66 (+0.16)	0.46 (+0.06)	0.33 (+0.05)	0.64 (+0.21)	0.43 (+0.10)	0.57 (+0.39)	0.45 (+0.34)

Experimental Validation of Formulated Prompts To validate the effectiveness of our prompt formulation-driven approach, we conducted an experiment focusing on the adequacy of analysis contained within formulated prompts generated by LLMs for sentences presenting logical fallacies. Specifically, we assessed whether the queries—encompassing counterarguments, explanations, and goals—provided by LLMs were capable of facilitating a comprehensive analysis of logical fallacies. This evaluation was performed using a Small Language Model (SLM), specifically *roberta-base*, to measure the model's certainty in its predictions through confidence scores. These scores, defined as the highest softmax likelihood (p_{SLM}) for a particular class label l given a sentence x, served as an indicator of the model's confidence in identifying and classifying logical fallacies based on the enriched formulated texts:

$$\text{Confidence}(x) = \max_{l \in L} p_{SLM}(l|x; X), \quad (4)$$

where L denotes a set of logical fallacy labels within the test dataset X. This methodology aims to critically examine the extent to which LLM-generated queries could enrich the model's understanding and analytical depth, thereby enhancing its performance in the classification of logical fallacies. Our findings, illustrated in Figure 2, 3, indicate that, in most cases, the inclusion of queries generally leads to higher performance at lower confidence scores across all datasets, affirming the efficacy of our enriched formulation methodology. For more detailed insights into the experimental setup, please refer to the appendix A.

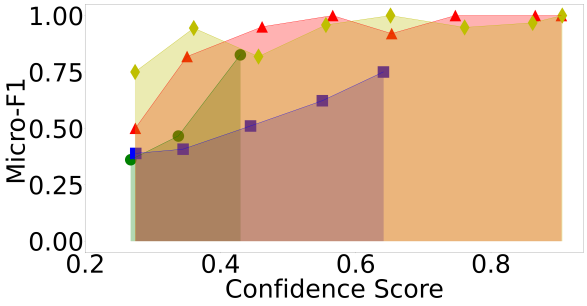


Figure 2: Relationship between confidence scores and performance with/without formulated texts for the ARGOTARIO dataset. Symbols represent different approaches: (○) Base, (□) Counterargument, (△) Explanation, and (◇) Goal. Breaks in the lines indicate the absence of corresponding confidence scores. Additional dataset results are available in the Figure 3.

5 Ethics Statement

This research into the application of Large Language Models (LLMs) for the detection and classification of logical fallacies carries profound ethical considerations that we have addressed throughout the study's design, execution, and analysis phases. In harnessing the capabilities of LLMs such as GPT-3.5-turbo and GPT-4.0, we acknowledge the responsibility to ensure that our methodologies and findings contribute positively to society and do not exacerbate existing disparities or introduce new forms of bias.

6 Conclusion

This study introduced a novel approach leveraging Large Language Models (LLMs) like *GPT-3.5* and *GPT-4.0* to enhance the detection and classification of logical fallacies. By reformulating enriched representations from sentences containing logical fallacies, we evaluated the models' capabilities across various datasets. Our methodology, focusing on strategic reformulating based on enriched case representations like counterarguments, explanations, and goals, demonstrated significant success in improving model precision.

7 Limitation

Despite the promising advancements demonstrated in leveraging Large Language Models (LLMs) like GPT-4 for detecting and classifying logical fallacies through a novel prompt formulation approach, this study acknowledges several limitations. The generalizability of our findings across more diverse or complex datasets remains untested, particularly outside the 5 to 13 types of logical fallacies explored. Additionally, our method's reliance on specific models (GPT-3.5-turbo and GPT-4.0) may limit its applicability with other LLMs or future iterations, potentially impacting its scalability and cost-effectiveness due to the increased complexity and computational requirements of prompt reformulation. Moreover, the improvements in Macro F1-scores, while significant, highlight the need for further research to enhance the interpretability and transparency of the reformulation process, ensuring the method's broader applicability and effectiveness in practical settings.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 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 preprint arXiv:2301.09992.

Sharon Bailin and Mark Battersby. 2016. Reason in the balance: An inquiry approach to critical thinking. Hackett Publishing.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

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

Hartry H Field. 1977. Logic, meaning, and conceptual role. The Journal of Philosophy, 74(7):379–409.

Pierpaolo Goffredo, Mariana Espinoza, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 11101–11112. Association for Computational Linguistics.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In IJCAI, pages 4143–4149.

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

Hans V Hansen. 1996. Aristotle, whately, and the taxonomy of fallacies. In International Conference on Formal and Applied Practical Reasoning, pages 318–330. Springer.

Carl G Hempel and Paul Oppenheim. 1948. Studies in the logic of explanation. Philosophy of science, 15(2):135–175.

Ruixin Hong, Hongming Zhang, Xinyu Pang, Dong Yu, and Changshui Zhang. 2023. A closer look at the self-verification abilities of large language models in logical reasoning. arXiv preprint arXiv:2311.07954.

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

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(3).

E Michael Nussbaum, CarolAnne M Kardash, and Steve Ed Graham. 2005. The effects of goal instructions and text on the generation of counterarguments during writing. Journal of educational psychology, 97(2):157.

Domina Petric. 2020. Logical fallacies. On-line Article (preprint), doi, 10.

Saumya Sahai, Oana Balalau, and Roxana Horincar. 2021. Breaking down the invisible wall of informal fallacies in online discussions. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 644–657.

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

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

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

Karen Tracy. 2013. Understanding face-to-face interaction: Issues linking goals and discourse. Routledge.

Frans H Van Eemeren, A Francisca Sn Henkemans, and Rob Grootendorst. 2002. Argumentation: Analysis, evaluation, presentation. Routledge.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 176–187.

A Experimental Details

Fine-tuning setup To evaluate the effectiveness of formulation in a fine-tuning setup, we fine-tuned a *roberta-base* model. Given the limitations on input sentence length during training, we did not combine generated queries but instead concatenated each generated query with the logical fallacy sentences individually. The hyper-parameters for fine-tuning were set as follows: learning rates within {1e-5, 2e-5}, training and evaluation batch sizes among {4, 8, 16, 32, 64}, and a maximum sequence length of 512. This approach allowed us to comprehensively assess the model's performance across different configurations, as detailed in Table 3. Our codes are available at `https://anonymous.4open.science/r/Logical_Fallacy-5019/`

Fallacy Class To ensure coherence in our analysis, we consolidated similar fallacy categories across the datasets: ***Hasty Generalization*** and ***Faulty Generalization*** into ***Faulty Generalization***; ***Fallacy of Credibility*** and ***False Authority*** into ***Irrelevant Authority***; and ***False Cause***, ***False Causality***, and ***Post Hoc*** into ***False Causality***.

Data Selection The experiments were conducted on test subsets comprising 20% of each dataset for the initial evaluation, except for the LOGIC dataset, where we used the test set predefined by (Jin et al., 2022). This selection was based on the need to ensure a consistent evaluation framework across different datasets and to utilize existing benchmarks where available. Specifically, for the task of multi-class fallacy classification, where fine-tuning was employed, the datasets were split into training, validation, and test sets in proportions of 65%, 15%, and 20%, respectively. This partitioning was aimed at providing a robust structure for model training and evaluation, allowing for an effective balance between learning complex fallacy patterns and ensuring generalization across unseen data.

Generated Queries' Impact The analysis of generated query impact is specifically focused on the task of logical fallacy multiclass classification, which is recognized as a challenging task for the Small Language Model (SLM). Due to the complexity of accurately classifying multiple types of logical fallacies, a fine-tuning approach was adopted to achieve the results(See Figure 2, 3). Breaks in the data lines within Figure 2, 3 indicate the absence of corresponding confidence scores.

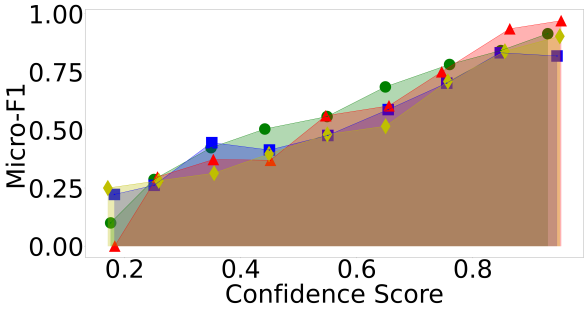
This process was applied across test datasets. Notably, for the **COVID-19** dataset, given its smaller size, cross-validation methods were employed to expand the test sample size, ensuring a comprehensive assessment. Our findings, illustrated in Figure 2, 3, indicate that: (1) With few exceptions (notably the LOGIC dataset), the inclusion of queries generally leads to higher performance at lower confidence scores across all datasets, affirming the efficacy of our enriched query generation methodology. (2) Uniformly, the presence of queries at any given confidence level enhances model performance, validating the positive influence of our query generation approach on the model's effectiveness. However, the performance on the **LOGIC** dataset was comparatively lower, likely due to the complexity introduced by its 13 distinct classes.

This additional step highlights our commitment to rigorously evaluating the influence of prompt Promptformulation, particularly in complex classification scenarios, thereby underlining the efficacy and necessity of our methodology in enhancing model performance.

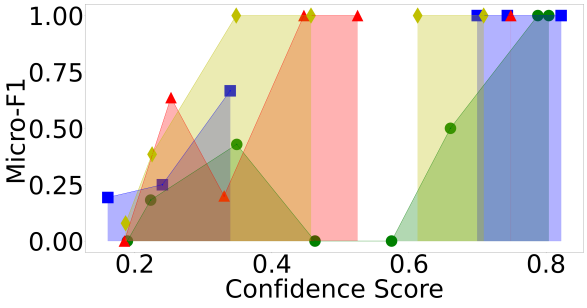
B Enriched Representation text

In this section, we delve into the core reasoning behind our selection of **counterargument**, **explanation**, and **goal** as the pivotal elements for enriching the representation texts. Our approach is inspired by the foundational work of (Sourati et al., 2023), who first demonstrated the effectiveness of using enriched case representations—focusing on counterarguments, goals, explanations, and structure—to analyze logical fallacies. This pioneering study laid the groundwork for our methodological choices, emphasizing the intrinsic value of these aspects in enhancing the analysis and understanding of logical fallacies within textual content.

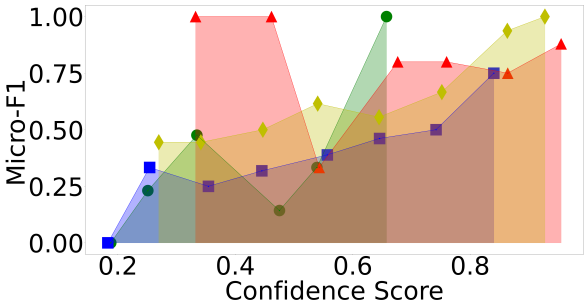
It is important to note that while (Sourati et al., 2023) included **structure** as one of the elements for enriched case representations, we have chosen to focus specifically on **counterargument**, **explanation**, and **goal** in our study. The rationale behind this decision is based on the recognized variability of **structure** across different datasets and domains. This variability can pose challenges in consistently applying and interpreting structural aspects of arguments across diverse contexts. By concentrating on **counterargument**, **explanation**, and **goal**, we aim to utilize elements that offer consistent analytical value and applicability regardless of the domain or dataset. This focus ensures that our methodology remains robust and adaptable, enabling precise and meaningful analysis of logical fallacies across a wide range of textual content.



(a) LOGIC



(b) COVID-19



(c) CLIMATE

Figure 3: Relationship between confidence scores and performance with/without generated queries. Symbols represent different approaches: (○) Base, (□) Counterargument, (△) Explanation, and (◊) Goal.

Counterargument serves as a crucial tool for testing the strength and resilience of the original argument, offering alternative viewpoints that challenge the initial stance(Nussbaum et al., 2005). This dynamic interaction is essential for unveiling hidden biases or weaknesses within the argument, thus providing a richer, more comprehensive analysis.

Explanation goes beyond the surface to explore the underlying rationale of the argument, dissecting the logic and reasoning that supports the fallacious statement. By understanding the "why" behind an argument(Hempel and Oppenheim, 1948), we

gain invaluable insights into its construction and the fallacies that may pervade it.

Goal focuses on the intended outcome of the argument, examining whether the argument's structure logically supports its conclusion(Tracy, 2013). This perspective is pivotal in assessing the effectiveness and coherence of the argumentative strategy employed.

C Prompt Details

This section outlines the prompts utilized for generating enriched representation texts and queries concerning Counterargument, Goal, and Explanation, employing *GPT-3.5-turbo-instruct*. These prompts are designed to deeply engage the Large Language Model (LLM) in extracting nuanced insights about the logical fallacies within the text, enabling a thorough analysis and understanding of the argumentative structures. For comprehensive details on the utilization of these prompts, see below:

C.1 Enriched Representation Text Prompts

These prompts are utilized to elicit counterarguments, explanations, and goals directly from the LLM, forming the enriched representation texts ($R_i(x)$) as outlined in Equation 1. This foundational step, represented by $\mathcal{R}_i(x) = \text{LLM}(x, \mathcal{I}^i)$, allows us to delve deeper into the logical structure and validity of the argument presented in the text. Here, x represents the input text, and $\mathcal{I}^i$ signifies the instruction for generating either the counterargument ($\mathcal{I}^1$), explanation ($\mathcal{I}^2$), or goal ($\mathcal{I}^3$) of x. Consider the [TEXT]: *"Annie must like Starbucks because all girls like Starbucks."*

- **Counterargument**
  ► I'll give you some texts. The texts can be question and answer pairs or sentences. Represent the counterargument to the text.
  ► Text: [TEXT]
  ► Counterargument (Example): *"Not all girls like Starbucks, as personal preferences vary among individuals. Even if Annie is a girl, it does not automatically mean she likes Starbucks. She may prefer a different type of coffee or may not like coffee at all. It is not fair to make assumptions about someone based on their gender."*

- **Explanation**
  ► I'll give you some texts. The texts can be question and answer pairs or sentences. Analyze the text.
  ► Text: [TEXT]
  ► Explanation (Example): *"This text suggests a generalization about girls and their preferences for Starbucks, assuming that Annie, as a girl, must also like Starbucks without evidence. This could be seen as stereotyping, making unfounded assumptions based on gender, reinforcing harmful stereotypes."*

- **Goal**
  ► I'll give you some texts. The texts can be question and answer pairs or sentences. Express the goal of the text.
  ► Text: [TEXT]
  ► Goal (Example): *"The goal is to make a generalization about girls liking Starbucks based on the assumption that Annie is a girl."*

C.2 Query Generation

Following the generation of enriched texts, these prompts are employed to craft queries that delve deeper into the logical structure and validity of the arguments presented in the texts, as denoted by the equation $Q_i(R_i(x), x) = \text{LLM}(\mathcal{R}_i(x), \mathcal{Q}^i)$. In this equation, $\mathcal{R}_i(x)$ represents the enriched representation obtained from the input text x through specific instructions $\mathcal{I}^i$, and $\mathcal{Q}^i$ denotes the set of generated query instructions. These generated queries are intended to enhance the depth of analysis on x, improving the detection and classification of logical fallacies within. The indices $i = 1, 2, 3$ correspond to the different aspects of queries: Counterargument (Q^1), Explanation (Q^2), and Goal (Q^3). Below are the detailed prompts for generating these queries based on the enriched texts:

- **Query Generation for Counterargument Text**
  ► I'll give you some texts and their counterarguments. The texts can be question and answer pairs or sentences. Create one query for each text to analyze the text based on its counterarguments.
  ► Text: [TEXT]
  ► Counterargument: [COUNTERARGUMENT]
  ► Query text (Example): *"How does the counterargument challenge the assumption that all girls like Starbucks?"*

- **Query Generation for Explanation Text**

► **I'll give you some texts and their explanations. The texts can be question and answer pairs or sentences. Create one query for each text to analyze the text based on its explanations.**
► **Text:** [TEXT]
► **Explanation:** [EXPLANATION]
► **Query text (Example):** *"How does this text perpetuate harmful gender stereotypes and restrict individual expression?"*

- **Query Generation for Goal Text**
  ► **I'll give you some texts and their goals. The texts can be question and answer pairs or sentences. Create one query for each text to analyze the text based on its goal.**
  ► **Text:** [TEXT]
  ► **Goal:** [GOAL]
  ► **Query text (Example):** *"What does this text reveal about the speaker's attitude towards girls and their preferences?"*

C.3 Detection and Classification Prompts

In our experiments aimed at detecting and classifying logical fallacies, we utilized specific prompts grounded in the approach described by Equation 3. This equation outlines the process by which the Large Language Model (LLM) assesses the presence and type of a logical fallacy within a given text, x, based on the generated queries, Q_i, derived from enriched representations, $\mathcal{R}_i(x)$. The methodology is specifically designed to challenge the LLM's ability to recognize logical fallacies in a zero-shot learning framework, leveraging no prior task-specific training but instead utilizing enriched texts and generated reformulated queries to deeply analyze argumentative structures. Focusing on the ARGOTARIO dataset, which encompasses five distinct types of fallacies, this approach enhances the LLM's precision in both detecting the presence of fallacies and classifying their specific types. Below, we detail the prompts utilized in this investigative process:

- **Logical Fallacy Detection**
  ► **Your task is to detect a fallacy in the Text. The label can be 'Fallacy' or 'None'. Please detect a fallacy in the Text based on Queries.**
  ► **Text:** [TEXT]
  ► **Formulated prompt:** [FORMULATED PROMPT]
  ► **Label:**

Table 4: Accuracy and F1 score for each fallacy type on the ARGOTARIO dataset, indicating the number of data points for each type (N). (B) and (F) indicate Base and Formulated queries. The presented results employ the formuated prompts that achieved the highest Accuracy and Macro F1-score for each model: Formulated Goal Prompt for GPT-3.5-turbo and Formulated Explanation Prompt for GPT-4.0.

Model	GPT-3.5-turbo				GPT-4.0				N
	B		F		B		F		
Metric	Acc	F1	Acc	F1	Acc	F1	Acc	F1	
Appeal to Emotion	0.78	0.56	0.88	0.78	0.81	0.62	0.94	0.88	48
Faulty Generalization	0.77	0.49	0.90	0.75	0.85	0.63	0.95	0.86	39
Red Herring	0.74	0.45	0.86	0.73	0.81	0.56	0.96	0.89	36
Ad Hominem	0.86	0.59	0.92	0.75	0.85	0.59	0.93	0.81	34
Irrelevant Authority	0.86	0.39	0.91	0.58	0.89	0.62	0.94	0.78	25
Accuracy	0.50		0.74		0.60		0.84		182
Macro F1	0.41		0.72		0.50		0.70		

- **Logical Fallacy Classification**
  ► **Your task is to classify the type of fallacy in the Text. The label can be 'Appeal to Emotion', 'Faulty Generalization', 'Red Herring', 'Ad Hominem', and 'Irrelevant Authority'. Please classify the type of fallacy in the Text based on the Queries.**
  ► **Text:** [TEXT]
  ► **Formulated prompt:** [FORMULATED PROMPT]
  ► **Label:**

The utilization of *GPT-3.5-turbo-instruct* for this endeavor leverages its capability to generate precise and relevant responses, a critical component for an exhaustive exploration of logical fallacies.

Error Analysis This study conducted an error analysis across different prompts for logical fallacy prediction (See Table 5 in Appendix). The analysis categorizes the results into three types: **All Correct Predictions**, **All Incorrect Predictions**, and **Not All Correct Predictions**. In the case of **All Correct Predictions**, All formulated prompts contributed to correctly identifying the answer. For instance, the sentence *"Researchers are frauds who don't earn their salaries."* was accurately classified as an **Ad Hominem** fallacy. Each prompt text provides a unique perspective on examining the claim's credibility: CR questioned the evidence supporting the claim, ER explored underlying assumptions or biases, and GR assessed

the author's views on researchers' credibility and integrity. This multifaceted approach was crucial for accurately identifying logical fallacies. In the **All Incorrect Predictions** scenario, the sentence *"I'm French and I don't like cheese."* was misclassified as a **Faulty Generalization** fallacy, which overlooks the deeper issue of **Intentional Fallacy**. This classification misses the essence of the intentional fallacy, where assertions are made not on the basis of logical evidence or factual support but are driven by the speaker's intent to win an argument. The response *"Then you must not really be French"* highlights an intention to question the speaker's identity without substantive evidence, reflecting an underlying intention to assert dominance in the argument rather than engaging with factual accuracy. In **Not All Correct Predictions**, only the Goal Formulated text (GR) correctly identified the **Cherry Picking** fallacy in *"The bottom line is there's no solid connection between climate change and the major indicators of extreme weather."* The GR provided a clearer perspective for analyzing the text compared to CR or ER. By directly addressing the text's main argument and its clarity, GR facilitated the effective extraction of critical information within complex texts. These findings underscore the importance of diverse perspectives in formulation for the accurate identification and understanding of logical fallacies. The unique approach of each formulation to text analysis supports the effective identification of fallacies, highlighting that formulated query design significantly impacts prediction model performance. This emphasizes the necessity of optimizing queries in future research to enhance model efficacy.

D Analysis of Class-specific Performance on Four Datasets

Analyzing the performance across the **ARGOTARIO**, **LOGIC**, **COVID-19**, and **CLIMATE** datasets provides a comprehensive view of the efficacy of reformulated text types in enhancing the multiclass classification capabilities of GPT-3.5-turbo and GPT-4.0 models for various logical fallacies. In the **ARGOTARIO** dataset, employing Goal Reformulation with GPT-3.5-turbo and Reformulated explanation texts with GPT-4.0 led to marked performance enhancements, particularly for the *Appeal to Emotion* and *Faulty Generalization* fallacies, suggesting that specific reformulated queries can significantly enhance model sensitivity and accuracy in identifying nuanced logical fallacies. The **LOGIC** dataset revealed an overall improvement across all fallacy types when Reformulated explanation texts were applied, highlighting the models' improved capability in discerning fallacies involving circular reasoning and ambiguous language, with *Circular Reasoning* and *Equivocation* fallacies showing strong F1 score improvements. The **COVID-19** dataset, predominantly utilizing Explanation Reformulation, showcased enhanced ability to classify the *False Causality* fallacy, indicating that additional contextual prompts provided by reformulated queries can aid models in better understanding incorrect causal relationships. In the **CLIMATE** dataset, a general performance improvement was observed, with notable advancements in classifying *False Causality*, *Vagueness*, and *Cherry Picking* fallacies through the application of Reformulated explanation texts, highlighting the impact of enriched reformulated queries in deepening models' comprehension of specific fallacies. Across these datasets, the strategic use of reformulated queries, particularly the Reformulated Explanation texts, has been notably effective in not only bolstering models' classification accuracy but also in offering insights into their abilities to process and understand complex logical constructs within textual content. This underscores the nuanced impact of specific reformulated text's types, with Reformulated explanation texts emerging as particularly influential, in enhancing the precision and depth of logical fallacy classification by large language models.

Table 5: Examples of all correct predictions, all incorrect predictions, and not all correct predictions. **All correct predictions** refers to cases where every formulated prompt correctly identifies the answer, **All incorrect predictions** refers to cases where every reformulated text selects an incorrect answer, and **not all correct predictions** indicates that some reformulated queries' types are correct while others are incorrect. **GT** stands for the ground truth label, **B** for base prediction, **C** for counterargument, **E** for explanation, **G** for goal, **CR** for reformulated counterargument text, **ER** for reformulated explanation text, and **GR** for reformulated goal text.

All Correct Predictions

Text: "Researchers are frauds who don't earn their salaries."
G: Ad Hominem
B: Faulty Generalization **C**: Ad Hominem, **E**: Ad Hominem, **G**: Ad Hominem
CR: "What evidence do you have to support your claim that all researchers are frauds and do not earn their salaries?"
ER: "What underlying assumptions or biases might lead someone to make such a claim about researchers?"
GR: "How does this text's statement reflect the author's view on the credibility and integrity of researchers?"

All Incorrect Predictions

Text: A: I'm French and I don't like cheese. B: All French like cheese. A: I don't. B: Then you must not really be French
GT: Intentional Fallacy
B: Faulty Generalization
C: Faulty Generalization, **E**: Faulty Generalization, **G**: Faulty Generalization
CR: "Can an individual's national identity be accurately defined by their personal preferences and tastes?"
ER: "How does this text reinforce the idea of stereotypes and their influence on one's identity?"
GR: "How does the text use the stereotype of French people liking cheese to question the authenticity of the speaker's nationality?"

Not All Correct Predictions

Text: "The bottom line is there's no solid connection between climate change and the major indicators of extreme weather."
GT: Cherry Picking
B: False Causality
C: False Causality, **E**: False Causality, **G**: Cherry Picking
CR: "How do you respond to the argument that there is no solid connection between climate change and extreme weather events?"
ER: "How does the text present the relationship between climate change and extreme weather?"
GR: "What is the main argument of this text and what is it trying to clarify?"

Table 6: Accuracy and F1 score for each fallacy Type on LOGIC dataset. (N) denotes the number of data points for each type. (B) and (F) indicate Base and Formulated prompts. The presented results employ the generated queries that achieved the highest Accuracy and Macro F1-score for each model: Formulated Explanation Prompt for GPT-3.5-turbo and Formulated Explanation Prompt for GPT-4.0.

Model	GPT-3.5-turbo				GPT-4.0				N
	B		F		B		F		
Metric	Acc	F1	Acc	F1	Acc	F1	Acc	F1	
Faulty Generalization	0.83	0.54	0.87	0.67	0.86	0.63	0.87	0.67	61
Ad Hominem	0.96	0.86	0.98	0.92	0.95	0.81	0.99	0.96	41
Ad Populum	0.96	0.75	0.97	0.85	0.97	0.60	0.99	0.97	30
Red Herring	0.92	0	0.90	0	0.92	0	0.91	0	24
Appeal to Emotion	0.92	0	0.92	0	0.92	0	0.92	0	23
Fallacy of Extension	0.93	0.08	0.91	0	0.86	0	0.87	0	21
Circular Reasoning	0.96	0.52	0.98	0.77	0.96	0.60	0.98	0.79	19
False Causality	0.82	0.34	0.82	0.36	0.84	0.41	0.89	0.47	18
Irrelevant Authority	0.93	0	0.92	0.08	0.93	0	0.91	0	17
Intentional Fallacy	0.78	0.11	0.86	0.16	0.84	0.08	0.87	0.14	15
Deductive Reasoning	0.93	0.08	0.89	0	0.91	0	0.89	0	14
False Dilemma	0.95	0	0.92	0	0.92	0	0.89	0	12
Equivocation	0.97	0.18	0.99	0.73	0.99	0.50	1.0	0.89	5
Accuracy	0.39		0.45		0.48		0.48		300
Macro F1	0.25		0.32		0.28		0.35		

Table 7: Accuracy and F1 score for each fallacy Type on CLIMATE dataset. (N) denotes the number of data points for each type. (B) and (F) indicate Base and formulated prompts. The presented results employ the formulated prompts that achieved the highest Accuracy and Macro F1-score for each model: Formulated Explanation Text for GPT-3.5-turbo and Formulated Explanation Text for GPT-4.0.

Model	GPT-3.5-turbo				GPT-4.0				N
	B		F		B		F		
Metric	Acc	F1	Acc	F1	Acc	F1	Acc	F1	
Cherry Picking	0.76	0.27	0.82	0.59	0.73	0.24	0.90	0.74	21
Vagueness	0.73	0.19	0.91	0.73	0.63	0.29	0.93	0.80	14
Red Herring	0.86	0	0.98	0.92	0.83	0	0.95	0.80	13
False Causality	0.60	0.21	0.86	0.58	0.71	0.23	0.89	0.50	11
Irrelevant Authority	0.88	0.35	0.95	0.74	0.85	0.13	0.91	0.64	10
Evading the Burden of Proof	0.90	0	0.95	0.62	0.86	0	0.86	0.52	9
Strawman	0.92	0	0.97	0.77	0.91	0	0.96	0.67	7
False Analogy	0.88	0.27	0.99	0.89	0.93	0.25	0.98	0.80	5
Faulty Generalization	0.90	0	0.98	0	0.98	0	0.98	0	2
Accuracy	0.18		0.70		0.18		0.67		92
Macro F1	0.13		0.65		0.11		0.55		

Table 8: Accuracy and F1 score for each fallacy Type on COVID-19 dataset. (N) denotes the number of data points for each type. (B) and (F) indicate Base and Formulated queries. The presented results employ the formulated queries that achieved the highest Accuracy and Macro F1-score for each model: Explanation Prompt for GPT-3.5-turbo and Explanation Prompt for GPT-4.0.

Model	GPT-3.5-turbo				GPT-4.0				N
	B		F		B		F		
Metric	Acc	F1	Acc	F1	Acc	F1	Acc	F1	
Cherry Picking	0.86	0	0.71	0.50	0.86	0	1.0	1.0	2
Vagueness	0.86	0	0.93	0	0.86	0	0.93	0	1
Red Herring	0.93	0	1.0	1.0	0.93	0	1.0	1.0	1
False Causality	0.43	0.43	0.86	0.75	0.79	0.67	0.86	0.75	3
Irrelevant Authority	0.86	0	0.86	0	0.93	0.67	0.86	0	2
Evading the Burden of Proof	0.86	0	0.93	0.67	0.64	0	0.79	0.40	2
Strawman	0.93	0	0.93	0	1.0	1.0	1.0	1.0	1
False Analogy	1.0	1.0	1.0	1.0	0.93	0.67	1.0	1.0	1
Faulty Generalization	0.93	0	0.93	0	0.93	0	1.0	1.0	1
Accuracy	0.29		0.57		0.43		0.71		14
Macro F1	0.14		0.44		0.33		0.68		

Table 9: One example of different formulation for the text ***Asians make lousy athletes, but do well at the Math Olympiad***

Representation	**Generated Query**
Counterargument	How does the success of Asians in the Math Olympiad contradict the belief that they are not talented athletes?
Explanation	Do you think this text promotes harmful stereotypes about Asians, and why or why not?
Goal	How does the text challenge the idea of making generalizations about a group of people based on limited information?