AUTOMATING FALLACY DETECTION: ADVANCEMENTS WITH OPEN-SOURCE LARGE LANGUAGE MODELS

Diogo Pedrosa
AI Flow Solutions
Pombal, Portugal
diogofranciscop@hotmail.com

Duarte Gomes
AI Flow Solutions
Leiria, Portugal
duarte.gcgcomes@gmail.com

ABSTRACT

Argumentation is a fundamental element of human discourse, crucial in shaping societies and influencing decision-making processes. However, logical fallacies can compromise the integrity of arguments, leading to the spread of misinformation and harmful ideologies. This study explores the application of machine learning, particularly Large Language Models (LLMs), in detecting fallacies within natural language text. Using LLaMA-3.1 70B, an open-source LLM, we developed a methodology that leverages prompt engineering to achieve high accuracy in fallacy detection. Extensive evaluation with the LOGIC dataset resulted in an F1 score of 92.9%, significantly outperforming previous models designed for fallacy detection. This research highlights the potential of advanced LLMs to enhance critical reasoning through real-time fallacy detection, contributing to social networking moderation, misinformation mitigation, and political discourse analysis

***K*eywords** Artificial Intelligence, Large Language Models, Fallacy, Logical Fallacies, Open-source, Fallacy Detection, Natural Language Processing, Real-Time Analysis, Speech Analysis, Machine Learning

1 Introduction

Argumentation is a fundamental aspect of human existence, serving as the primary means through which ideas are exchanged, decisions are made, and societies are shaped. Whether logical or fallacious, argumentation significantly influences the trajectory of human thought and societal development. However, the efficacy of an argument depends crucially on its logical structure; flawed reasoning compromises the integrity of the entire discourse.

The occurrence of fallacies in speech, writing, and other forms of communication is a common phenomenon. As noted by Walton [1], while the presence of fallacies does not inherently invalidate the ideas presented, it does provide valuable insights into the argumentation process and the underlying logical structure.

Fallacies can lead to misguided beliefs, poor decision-making, and the spread of misinformation, all of which may have serious societal consequences. Many structural issues in society, such as chauvinism, racism, and misogyny, lack a logical foundation and are instead rooted in emotional biases. These prejudiced beliefs are not supported by logical principles, which strive to be universally absolute.

The implications of fallacious reasoning extend to any domain where information is disseminated, including the realms of politics and education, where the influence on society can be profound, either for better or for worse.

While argumentation is a universally accessible tool, its impact is determined by how individuals choose to wield it. It can be a force for progress or a vehicle for harm, spreading both accurate information and misinformation, and potentially undermining the very fabric of society.

Given the dual-edged nature of argumentation, the question arises: how can we prevent its misuse and guide individuals toward constructive discourse?

Traditionally, detecting fallacies has been the domain of philosophers, often rooted in Aristotelian principles. Aristotle's 'Organon,' which encompasses works like 'Prior Analytics' and 'Posterior Analytics,' laid the foundational framework

for formal logic and dialectical reasoning [2]. Aristotle emphasized that logic is fundamental to all forms of rational inquiry and essential to reasoning. His contributions have profoundly influenced argumentation theory over centuries.

However, the vast volume of information in the modern era exceeds the analytical capacity of philosophers or any individual. This challenge necessitates a new approach to identifying logical flaws in reasoning on a large scale.

This is where artificial intelligence (AI) and machine learning (ML) can play a crucial role. ML offers the potential to analyze text and speech at scale, leveraging mathematical models to detect fallacies with precision.

In this paper, we propose the development of a tool capable of reading, analyzing, and detecting fallacies in various forms of text and speech. This tool aims to ensure that critical decisions, particularly those with far-reaching consequences, are based on sound logical reasoning.

1.1 Expected Impact and Contribution

The research presented in this paper is anticipated to make significant contributions to the field of Fallacy Detection and Prediction. By evaluating the impact of fallacies on the progression of conversations, arguments, and, consequently, the development of society, this study aims to underscore the importance of identifying fallacious content across various sources. Moreover, it seeks to establish a foundational framework upon which future researchers can build.

Should the findings of this paper prove favorable, they could have profound implications for specific sectors such as Social Networking Moderation, Misinformation Mitigation Strategies, and Journalism. Any domain that values transparency and truth as core principles stands to benefit substantially from the outcomes of this research. Conversely, entities that rely on misinformation, fallacious reasoning, or populist rhetoric for profit may face adverse effects.

Additionally, this study aspires to contribute meaningfully to sectors often regarded as more critical, such as the Political domain. As noted by numerous scholars, one of the most significant political developments of the 21st century is the rise of populism [3]. Populism is frequently supported by two key pillars: an appeal to emotion rather than logic, and the dissemination of misinformation. While populism is a tool employed by many political parties, it is particularly prevalent in extreme right- and left-wing factions.

Therefore, if a Non-Governmental Organization (NGO) or international entity (such as the European Union) had access to an open-source, transparent tool designed to identify logical contradictions (fallacies) or misinformation propagated by these extreme parties, it could enable the public to better discern truth, thereby fostering societal progress.

2 Related Work

Detecting fallacies in arguments is an emerging and increasingly important topic within the field of Natural Language Processing (NLP). This area has garnered attention from several researchers, who, like us, are focused on addressing the pervasive issue of misinformation.

Zhijing Jin et al. [4] have made notable contributions by providing two new datasets to evaluate the performance of language models in detecting fallacies—one dataset focuses on logical fallacies commonly found in text, while the other targets fallacies related to climate change claims. Their findings indicate that existing pretrained large language models (LLMs) under-perform when compared to their structure-aware classifier. However, their study does not fully explore the potential of LLMs when optimized through careful prompt engineering. Given the rapid advancements in LLMs since their publication, these models are likely to achieve improved performance with proper methodological adjustments.

A related study by Pierpaolo Goffredo et al. [5] focused on detecting fallacies in political speeches. Their research introduced a twofold approach: they first extended the ElecDeb60To16 dataset, which includes annotated fallacious arguments from U.S. presidential debates, by incorporating data from the recent Trump-Biden debate. Subsequently, they developed neural network architectures based on Transformer models to perform the dual task of detecting and classifying fallacious arguments, integrating textual data, argumentative features, and engineered features.

Zhivar Sourati et al. [6] contributed to the field by formalizing a prior theoretical framework on logical fallacies into a comprehensive three-stage evaluation approach encompassing detection, coarse-grained, and fine-grained assessment. They combined language models with background knowledge and explainable mechanisms, achieving an accuracy of 0.631 using the Electra model and the Instance-based Reasoning (IBR) method on the same dataset as Jin et al. [4].

In another related work, Ramon Ruiz-Dolz and John Lawrence [7] investigated the practical applicability of automatic fallacy identification in natural language text in real-world scenarios. They developed a validation corpus consisting of natural language argumentation schemes and provided new empirical results that underscore the challenges of

identifying fallacies in the wild. Their findings suggest that relying solely on LLMs for such a complex task may not be the most effective approach.

Francisco Zanartu et al. [8] focused on detecting fallacies specifically within the domain of climate misinformation. In their paper, "Detecting Fallacies in Climate Misinformation: A Technocognitive Approach to Identifying Misleading Argumentation," they present a novel methodology that integrates both technical aspects of NLP and cognitive science theories to enhance the identification of misleading arguments in climate-related discourse. Their custom NLP pipeline, which includes fallacy detection models trained on annotated climate misinformation datasets, demonstrates improved performance over baseline models. This work significantly contributes to the growing field of automated misinformation detection by emphasizing the importance of combining linguistic analysis with cognitive insights.

K. Nieto-Benitez et al. [9] explored the application of machine learning techniques to the automatic detection of "appeal to emotion" fallacies. They employed Support Vector Machine (SVM) and Multilayer Perceptron (MLP) models, with the latter achieving an F1 score of 0.60 in fallacy identification. Their study highlights the necessity of systematic approaches to understanding the linguistic complexities inherent in fallacy detection, particularly in the context of political discourse. The use of affective lexical dictionaries and advanced neural models are proposed as crucial elements for improving detection efficacy.

Abhinav Lalwani et al. [10] introduced a novel approach for the automatic detection of logical fallacies by translating natural language into First-Order Logic (FOL) using a step-by-step process with LLMs. Their methodology incorporates Satisfiability Modulo Theory (SMT) solvers to evaluate the validity of logical formulas and classify statements as either fallacious or valid. Their approach is notable for its robustness, interpretability, and independence from training data or fine-tuning. Evaluation on the LOGIC dataset resulted in an F1-score of 71%, with even better performance on the LOGICCLIMATE challenge set, achieving an F1-score of 73%, surpassing state-of-the-art models by 21%.

From our review of the literature, it is evident that the capabilities of prompt-based systems are frequently underestimated. The potential of LLMs, when properly prompted, is immense. Thus, this study aims to address this gap by demonstrating that precise prompt engineering is one of the most effective methods for detecting fallacies.

2.1 Practical Works

Several practical implementations have aimed to achieve automatic fallacy detection. Notably, various GitHub repositories demonstrate promising results with high apparent accuracy. Although these repositories have not been subject to formal scientific studies, they have been instrumental in guiding our experiments.

In particular, our current system's prompt engineering was inspired by the approach used in the GitHub repository [11], which provided valuable insights and practical methodologies that influenced our work.

3 Methodology

In this section we'll explain in detail our methodology for achieving a detecting fallacy detector.

3.1 Large Language Model

Although there are models with greater capabilities, our goal is to democratize the ability to detect fallacies in real-time, making this technology accessible to a wider audience. To achieve this, we believe that an open-source model is the most suitable solution. Recent advancements by Meta in this field have substantially narrowed the performance gap between open-source models and proprietary counterparts, such as those developed by OpenAI and Anthropic.

For this study, we have selected LLaMA-3.1 70B as our open-source model of choice. While this model is capable of running locally on sufficiently powerful hardware, we have opted to utilize Groq Cloud for faster and more efficient inference. Groq Cloud's API, which is currently available free of charge, offers a convenient and cost-effective solution for deploying our model in real-world scenarios.

Given that the primary objective of this study is to develop a robust methodology for detecting fallacies, we have not conducted comparative testing with other models. However, our codebase is fully open-source [12], allowing researchers and practitioners to easily experiment with alternative models. This flexibility ensures that the broader research community can contribute to and benefit from improvements in fallacy detection methodologies.

We are committed to fostering an open and collaborative environment, where advancements in detecting fallacies are shared and accessible to all. The availability of our code and the use of an open-source model are integral to this vision, enabling further experimentation and potential improvements by the wider AI and NLP communities.

3.2 Prompt Engineering

Our initial approach involved testing a Large Language Model (LLM) using only prompt engineering techniques. Although our method can be implemented with local tools, for this study, we utilized cloud resources—specifically Groq—to achieve faster inference

We provided the system with a prompt instructing the LLM that it was "an expert in critical reasoning and logical fallacy detection. Your task is to analyze the given text and accurately identify any logical fallacies present."

Additionally, within the system prompt, we supplied the LLM with examples of how to respond when a phrase contains a fallacy. In this initial phase, we drew inspiration from a repository on GitHub [11]

3.3 Evaluation

To evaluate the effectiveness of the LLM in detecting fallacies, we used a dataset, LOGIC, that was previously cleaned and made available by another study, which involved training a machine learning model from scratch [4]. The dataset is also available on our GitHub repository.

Since this task can be framed as a classification problem, we used accuracy, precision, recall, and F1 score as metrics to assess the effectiveness of the LLM in classifying phrases

As a baseline, we will compare the results achieved by the Llama model with those previously obtained in the cited study.

To check what each measure is we define them the the sub sections bellow. The labels of each measure are:

- TP - True Positive
- FP - False Positive
- FN - False Negative
- TN - True Negative

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

3.4 Experimental Setup

The experimental setup for detecting fallacies using a Large Language Model (LLM) is structured as illustrated in the diagram bellow 1. The process begins with an input phrase, which is fed into the LLM for analysis. The LLM evaluates the phrase and generates a JSON output that includes various details, such as the detected fallacy type, a fallacy explanation, and potentially additional metadata like a GIF query to visually represent the identified fallacy.

The JSON output is compared against a predefined true label for the phrase, indicating the actual fallacy type present in the input. This comparison is facilitated by a dictionary that matches the true label with the fallacy detected by the LLM. If the detected fallacy corresponds to the true label, it validates the LLM's performance in accurately identifying the logical flaw.

This setup allows for an automated and systematic evaluation of the LLM's ability to detect and classify fallacies, providing a clear framework for testing the model's accuracy against known benchmarks. The integration of a dictionary for label comparison ensures that the assessment remains precise and consistent, making it easier to quantify the model's effectiveness in fallacy detection.

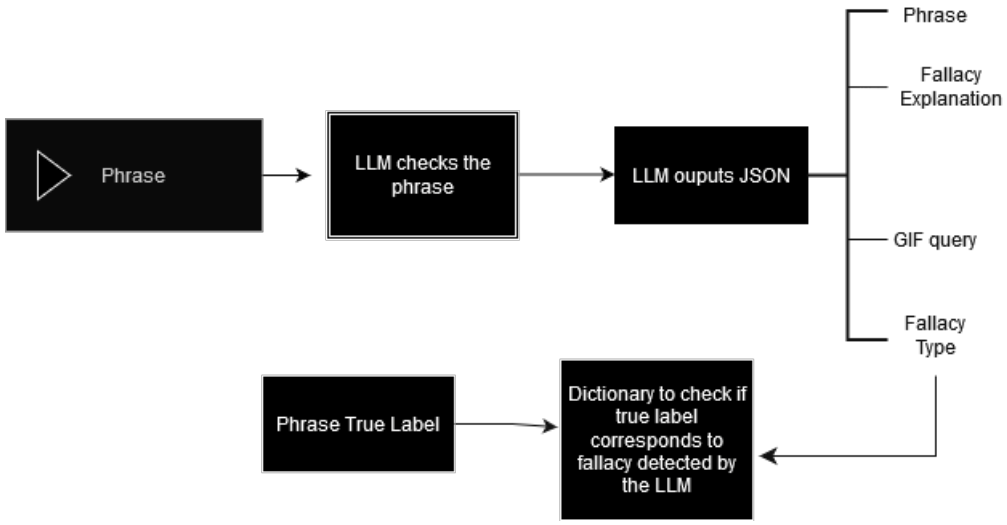

Figure 1: Squeme on how we performed the experiences to evaluate if the LLM detected fallacies correctly

4 Feature to fetch text

Fallacies can be present everywhere where information is transmitted, whether in newspapers, journals, books, interviews, or even casual conversations. They always arise in various forms of communication, therefore we also must fetch information from various sources. The most common Information Retrieval (IR) tools that this paper will use are the rather straightforward. In this papers, the following are included, but not limited to:

- Direct Input
- Speech-to-Text with Diarization and streaming Technology

More IR tools can be applied to this concept, such as Web Scraping/Crawling, Video Transcription or Live Data Sourcing from On-Screen Information, however to keep the readers engaged and allow them to fully understand the concepts mentioned here, only tools used in the bullet list above will be specifically mentioned.

4.1 Direct Input

The Direct Input Feature was by far the easiest implementation in this paper. However, despite its simplicity, it was extremely efficient detecting fallacies in short texts. Basically, the text that needed to be detected was directly inserted into the source-code (no GUI or any other method of data storage) and the Model would detect their respective fallacies and an explanation of how it was logically incongruent.

4.2 Speech-to-Text feature

For this feature, the Google Speech-to-Text API (STT)(Version 1) was implemented. Although the V2 is available, it cannot be used directly from the microphone, but only rather from a local file, which in the paper's specific case, was not necessary. Nevertheless, the API proved very accurate with simple conversations, therefore, the Google API suited perfectly for our case. We have also added Diarization to the STT code, which means that the tool can understand and differentiate different voices, therefore it can detect conversations between people. For instance, if a conversation on a microphone was being made (such as podcasts, or remote interviews), the text transcribed from the speech will be directly sent into the Fallacy detection feed, and the LLM would interpret it as different people having a conversation. This tool was also fine-tuned to specific conditions (based on software and hardware conditions), however the gains were statistically insignificant.

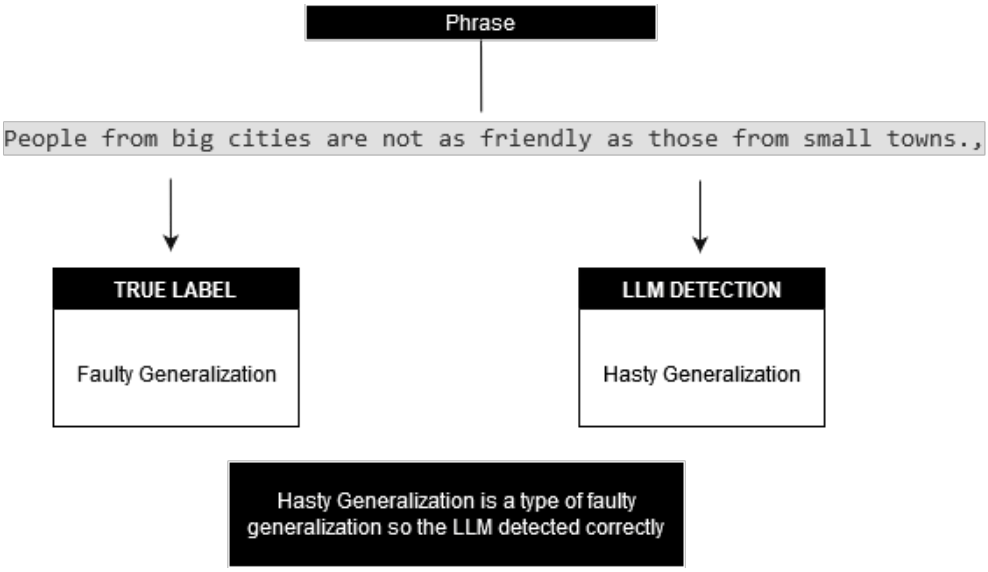

Figure 2: Example of a True label from the LLM although it is a different name

5 Results and Analysis

To assess whether the language model (LLM) can detect logical fallacies, we employed a dataset comprising 2,209 phrases. Initial experiments revealed that the LLM's outputs were related to the true label of the fallacy present in each phrase, but did not always correspond to the exact type of fallacy.

To address this issue, we developed a fallacy dictionary. This dictionary maps the true labels present in the dataset and allows us to compare these with the labels predicted by the LLM. We evaluated whether the predicted labels fell within the same category as the true labels. If a predicted label was within the same spectrum, it was included in the dictionary. In the figure 2 you can find one example.

The results, summarized in Table 1, show that the LLM accurately identified 85% of the fallacies. Considering that detecting false positives is less critical in this context, we estimate that the LLM successfully detected fallacies in 90.49% of the phrases, with a failure rate of 10.51%. The dataset did not include any non-fallacious arguments, resulting in no true negatives.

		Actual Class	
		Positive	Negative
Predicted Class	Positive	1878	121
	Negative	157	0

Table 1: Confusion matrix of the results of the LLM in the dataset

To more precisely analyse the results we have created a classification report where we use the formulas defines in the section 3.3.

$$Accuracy = \frac{1878 + 0}{1878 + 121 + 157 + 0} \approx 0.872 \tag{5}$$

$$Precision = \frac{1878}{1878 + 121} \approx 0.939 \tag{6}$$

$$Recall = \frac{1878}{1878 + 157} \approx 0.923 \tag{7}$$

$$F1Score = 2 \times \frac{0.939 \times 0.923}{0.939 + 0.923} \approx 0.929 \tag{8}$$

The high precision (93.91%) indicates that when the LLM predicts a fallacy, it is very likely to be correct. The recall (92.3%) reflects that the LLM successfully identifies most of the actual fallacies, though a small proportion remains undetected. As expected, the F1 score also shows a good balance between precision and recall. With a score of 92.9%, it signifies that the LLM performed well both in identifying positive cases and in avoiding the misclassification of negatives as positives.

5.1 Comparing Results

The results presented in Table 2 provide a comparative evaluation of the performance of prior models designed for detecting fallacies in text against the LLaMA-3.1 model using a simple system prompt.

Previous studies have extensively explored the capabilities of earlier models in detecting fallacies using large language models (LLMs). However, with the recent revolutionary advancements in LLMs, such as LLaMA-3.1, these models have demonstrated superior performance compared to earlier models specifically developed for fallacy detection. Notably, the LLaMA-3.1 model significantly outperforms the Electra-StructAware model, which was explicitly designed for fallacy detection. Additionally, when comparing LLaMA-3.1 with one of the earliest prominent LLMs, GPT-3, which yielded sub-optimal results as of December 2022, the advancements are even more evident.

Our results with the LLaMA-3.1 70B model indicate that comparable or even superior performance can likely be achieved with contemporary closed-source open models. This analysis demonstrates that modern LLMs possess substantial capabilities for fallacy detection, rendering the need for specialized NLP models to identify fallacies increasingly unnecessary.

	Precision	Recall	F-Score	Accuracy
Electra-StructAware	55.25	63.67	58.77	47.67
Electra	51.59	72.33	53.31	35.66
GPT3	12.00	12.00	12.00	12.00
LLaMA-3.1 70B	**93.9**	**92.3**	**92.9**	**87.2**

Table 2: Table comparing the results of previous study with the one's obtained on this one

5.2 Analysis of Incorrect Detections

The table 3reveals considerable variability in the performance of the fallacy detection system across different types of fallacies. Fallacies such as Faulty Generalization and Circular Reasoning exhibit high rates of both undetected instances and incorrect detections, suggesting a need for significant improvement in the detection algorithm's ability to handle these complex types. Conversely, Ad Hominem and False Dilemma show high accuracy rates when detected, although the total instances are relatively low, indicating that the system performs well with these fallacies when they are identified.

The system faces challenges particularly with Appeal to Emotion and Intentional fallacies, where a large number of instances go undetected, but those that are detected tend to be classified correctly. This highlights a discrepancy between detection coverage and accuracy. To enhance the system's performance, future efforts should focus on improving detection rates and reducing incorrect classifications, especially for the more problematic fallacies like Faulty Generalization and Circular Reasoning. Overall, targeted improvements in these areas are essential for increasing the system's reliability and effectiveness.

	Fallacies not detected analys		
	No detection	Detected wrong one	Total
Faulty Generalization	16	20	36
False Casuality	13	7	20
Circular Reasoning	14	22	36
Ad Populum	12	12	24
Ad Hominem	6	0	6
Fallacy of Logic	5	16	21
Appeal to Emotion	25	6	31
False Dillema	6	0	6
Equivocation	7	1	8
Fallacy of extension	5	3	8
Fallacy of Relevance	14	5	19
Fallacy of credibility	9	6	15
Intentional	25	23	48
Miscellaneous	0	0	0

Table 3: Analysis of Incorrect Detections by type of fallacy

5.3 Other Results

In this subsection we'll expose more abstract and simple results, like the results of the LLM explaining the fallacies and the results of our audio to text feature.

5.3.1 LLM explanation

We did not conduct a thorough study on the explanations provided by the LLM when it identifies a fallacy. However, based on our preliminary review, these explanations appear consistently reasonable. Nonetheless, we cannot confirm with certainty that they are always valid, as we did not employ any metrics to assess this. If you wish to validate the explanations provided by the LLM, you are welcome to review our public GitHub repository [12] to evaluate this aspect.

5.3.2 Speech to text

The Google Speech-to-Text API (Version 1) was tested using a standard microphone in a controlled environment. Phrases containing common logical fallacies were spoken and automatically transcribed by the API. These transcriptions were then, automatically, analyzed by our fallacy detection tool to assess its ability to identify fallacious reasoning.

We've selected 50 random samples from the LOGIC dataset and all the phrases were transcribed correctly, resulting in a transcription accuracy of 100%. This performance aligns with the expected capabilities of the Google Speech-to-Text API for handling straightforward spoken language.

The fallacy detection results are not relevant at this point, as we used the same phrases as in the previous subsection.

These results suggest that the Google Speech-to-Text API, combined with our fallacy detection tool, is highly effective in real-time applications, such as podcasts, interviews or live television debates.

6 Limitations

In this section we'll discuss some of the limitations our work had.

6.1 Model Dependency

The study relies heavily on the LLaMA-3.1 model, which, despite being open-source, might not represent the most advanced LLMs available. There could be other models that, although proprietary, might outperform the LLaMA-3.1 in detecting fallacies. The results are thus contingent on the specific capabilities and limitations of this chosen model.

6.2 Generalization

While the model demonstrated high accuracy in the specific dataset used, it is unclear how well it would perform on more diverse or unstructured texts, such as those found in casual conversations, social media, or less formal settings. The model's effectiveness in detecting fallacies in such varied forms of communication remains uncertain.

6.3 Limited Focus on Non-English Texts

The paper does not address how the model performs on non-English texts or whether the approach can be generalized across languages. Given that logical fallacies are a universal concept, this limitation could restrict the global applicability of the tool, especially in multilingual societies.

7 Conclusion

This study has demonstrated the efficacy of using Large Language Models, specifically LLaMA-3.1 70B, in detecting logical fallacies within natural language text. Our findings reveal that contemporary LLMs, when properly prompt-engineered, can surpass the performance of specialized models traditionally used for fallacy detection. With a precision of 93.9% and an F1 score of 92.9%, our model has shown a strong capability to accurately identify fallacies while minimizing false positives. These results underscore the potential of modern LLMs as powerful tools for combating misinformation and enhancing the quality of discourse across various domains, including social media, journalism, and political communication. The ability to automate and scale fallacy detection opens new avenues for fostering informed and rational public discourse, thereby contributing to a more logical and truthful society

8 Future Work

While this essay explored general topics for using fallacy detection tools, its possibilities are almost limitless. Although the findings in this paper were very promising and significant, they are only a starting point because this tool can be applied to very different contexts. In other words, every situation, whether based on communication, critical thinking, or decision-making, is within the reach of Fallacy Detection Tools.

8.1 Education

Using Fallacy Detection tools in Academia, whether it is in lower grades (such as Elementary Schools) or Higher Education (Universities) can be significant and rewarding. Educating children to be more prone to use logical argumentation (therefore applying less fallacies) can significantly impact their future growth, which in turn will significantly impact the general society. However, it can also be used in academic papers, such as essays or scientific journals. If these tools are applied at universities, they can easily detect fallacies in these papers, which could increase the quality of academic output. Although every paper usually has fallacies, some papers may have more severe fallacies than others, thus reducing the quality of the papers.

8.2 Business Decision-Making

Fallacies are often present in the Business Sector. Usually, businesses try to rely on logic (in most cases, in Key Performance Metrics). However, that is not always the case. There are external factors or even blunders that businesses

make. Therefore, businesses would gain a lot by applying fallacy detections. For instance, by applying these tools, they could analyse discussions from business meetings, financial reports, KPIs, or any logically driven document to detect flaws in proposals or strategies. Also, it could be crucial to Marketing and Advertising. Although Marketing "usually" relies on an appeal to emotion and not the true essence of the product, the fallacy detection could review promotional content to ensure that claims are free from misleading information.

8.3 Media and Journalism

Media and Journalism are the sectors in which communication happens the most, whether in a B2B or a B2P situation. Because logical fallacies happen when communication is present, this is where it is more critical to apply fallacy detection tools. For instance, Newspapers or companies that rely on news (in any form) would gain because they could use this tool to maintain journalistic integrity. It is almost Utopian to imagine a world where misinformation is scarce. However, with the tools present nowadays, it could be less Utopian and more realistic.

8.4 Politics and Public Discourse

The idea of this paper started with a specific example of politics. Politics has a massive role in everyday life. Most people are under government regulation. Therefore, an impact on the government impacts society, as it is a symbiotic relationship. If the government is performing poorly, people will be negatively affected. However, if the government performs well in general, people's quality of life usually increases. However, improving a government takes work. Also, all governments would be perfect, as is not the case. So many factors affect a government's efficiency and quality, so it is only possible to determine some factors. However, there are known and empirical factors that are known to affect the government directly. One of the most important are political parties. Parties usually differ by having different ideas (whether economic, social, or demographic, for instance). However, people cannot, with certainty, tell which is right or wrong. Different political parties say different things. However, this is where things get interesting; which party is telling the truth? It could be both of them, or it could be none of them. We must base our analysis on which party's view is right by analysing what they say, write, and inform. Ergo, it is possible to analyse political debates using fallacy detection software. Political debates are usually when more fallacies are present (Fallacies in speeches are usually more present than fallacies in writing; therefore, by analysing political debates and highlighting fallacious arguments, voters can be easily aided in making informed decisions. It can also be applied in policies (Agendas or any proposal), evaluating them by identifying underlying fallacies and supporting more rational policymaking. Political Speech can also be positively affected by this tool. Some fallacies are more severe than others; thus, if politicians make very severe fallacies in their Speech, they can be scrutinised to ensure that next time they base their Speech on sound reasoning.

8.5 Final Remarks

Although there are so many applications for these tools, those mentioned above are(from the perspective of this paper) the most important ones. Many ethical considerations still need to be addressed. However, this paper hopes to be a starting point for people who want to make a difference.

Acknowledgments

We would like to express our deepest gratitude to the AI Flow Solutions team for their invaluable support and contributions to this work. This research has been a significant endeavor for the past year, and it is incredibly rewarding to share our findings with the broader community.

Our sincere thanks go out to all the dedicated researchers who have inspired and guided us along the way. However, we owe the greatest debt of gratitude to the open source community in AI and Machine Learning. Your generosity in sharing research, code implementations, and knowledge has been instrumental in our growth and success. This paper is our humble attempt to contribute back to a community that has given us so much and continues to enrich our work every day.

We deeply appreciate your ongoing support and collaboration

References

[1] D. Walton. *Informal Logic: A Handbook for Critical Argumentation*. Cambridge University Press, 1989. Accessed: 2024-08-19.

[2] Aristotle. *Organon*. Archive.org, 2017. Accessed: 2024-08-19.

[3] European Centre for Action on Smoking and Health (ECAS). Populism and public health: How political rhetoric affects public health policy. Technical report, European Centre for Action on Smoking and Health, 2019. Accessed: 2024-08-19.

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

[5] Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. Argument-based Detection and Classification of Fallacies in Political Debates. In *EMNLP 2023 - Conference on Empirical Methods in Natural Language Processing*, volume 2023.findings-emnlp.684 of *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11101–11112, Singapore (SG), Singapore, December 2023. Association for Computational Linguistics.

[6] Zhivar Sourati, Vishnu Priya Prasanna Venkatesh, Darshan Deshpande, Himanshu Rawlani, Filip Ilievski, Hông Ân Sandlin, and Alain Mermoud. Robust and explainable identification of logical fallacies in natural language arguments. *Knowledge-Based Systems*, 266:110418, 2023.

[7] Ramon Ruiz-Dolz and John Lawrence. Detecting argumentative fallacies in the wild: Problems and limitations of large language models. In *Proceedings of the 10th Workshop on Argument Mining*. Association for Computational Linguistics, December 2023.

[8] Francisco Zanartu, John Cook, Markus Wagner, and Julian Garcia. Detecting fallacies in climate misinformation: A technocognitive approach to identifying misleading argumentation, 2024.

[9] K. Nieto-Benitez, N. A. Castro-Sanchez, H. Jimenez Salazar, G. Bel-Enguix, D. Mújica-Vargas, J. G. González-Serna, and N. González-Franco. Strategies for automatic detection of fallacious arguments in political speeches during electoral campaigns in mexico. *Proceedings of ISP RAS*, 36(1):259–276, 2024.

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

[11] Latent Variable. Fallacyscoreboard. `https://github.com/latent-variable/FallacyScoreboard`, 2024. Accessed: 2024-08-19.

[12] AiFlowSolutions. Aiflowsolutions github repository. `https://github.com/AiFlowSolutions`, 2024. Accessed: 2024-08-20.

A Appendix

A.0.1 Fallacies in the Evaluation Dataset

The dataset includes some of the most common fallacies found in arguments. Below is an overview of each fallacy present in the dataset:

- Faulty Generalization: This fallacy occurs when a conclusion is drawn from a sample that is not large or representative enough to warrant the conclusion. It often involves making sweeping statements based on limited evidence.
- False Causality: This fallacy involves assuming that because two events occur together or sequentially, one event must have caused the other, without sufficient evidence to support this causal relationship.
- Circular Reasoning: This occurs when the argument's conclusion is used as a premise, essentially reasoning in a circle. The argument assumes what it is trying to prove, offering no actual support for the conclusion.
- Ad Populum: Also known as the "appeal to popularity," this fallacy occurs when something is considered true or good simply because many people believe it or do it, rather than because of solid evidence or reasoning.
- Ad Hominem: This fallacy involves attacking the person making the argument rather than the argument itself. It diverts attention from the merits of the argument by focusing on irrelevant personal characteristics or beliefs of the individual.
- Fallacy of Logic: This is a broad category that includes any errors in logical reasoning, such as contradictions, inconsistencies, or misapplications of logical principles.
- Appeal to Emotion: This fallacy occurs when an argument manipulates emotions rather than using valid reasoning to persuade. It appeals to feelings such as fear, pity, or anger instead of addressing the actual issue.
- False Dilemma: Also known as the "either/or fallacy," this occurs when only two options are presented as the only possible outcomes, ignoring other viable alternatives.
- Equivocation: This fallacy happens when a key term or phrase in an argument is used with different meanings, creating a misleading or unsound argument.
- Fallacy of Extension: This occurs when an argument is misrepresented by extending it to an extreme or absurd form, then attacking the exaggerated version instead of the original argument.
- Fallacy of Relevance: This category includes fallacies where the premises are not logically relevant to the conclusion, often introducing irrelevant information to distract from the real issue.
- Fallacy of Credibility: This fallacy involves dismissing an argument based on the source rather than the content, or overly relying on the authority or credibility of a source without adequate justification.
- Intentional Fallacy: This occurs when the intentions behind an argument or statement are assumed to be the same as the argument's or statement's meaning, rather than analyzing the argument based on its own merits.
- Miscellaneous: This category includes various other fallacies that do not fit neatly into the above categories, encompassing a range of logical errors and rhetorical missteps.