




Logical Fallacy Detection in Text: Leveraging Large Language Models for Improving Human Discourse

Aravindh Manickavasagam¹ and Danushka Bandara²(✉) 

¹ Syracuse University, Syracuse, NY 13244, USA

² Fairfield University, Fairfield, CT 06824, USA

dbandara@fairfield.edu

Abstract. This study focuses on developing a logical fallacy detector using a fine-tuned large language model (LLM) to classify various logical fallacies in text. We utilize a publicly available logical fallacy dataset (LOGIC) combined with logically sound statements derived from the Stanford Natural Language Inference Corpus in this study. Our methodology involves preprocessing the data, fine-tuning the model, and validating its performance on test data. We demonstrate that an f1-score of 0.79 could be achieved for the 15-class classification task including logically sound statements. Zero-shot chain of thought prompting further improves classification f1-score to 0.81. The logically sound class obtained an f1-score of 0.99, indicating that the method is highly capable of distinguishing non-fallacious statements from fallacious statements. Our results show that the fine-tuned LLM model offers a promising tool for enhancing argument evaluation and promoting critical thinking.

Keywords: logical fallacies · classification · large language models

1 Introduction

Logic is the systematic study of the principles of valid inference and correct reasoning. It involves the examination of arguments, statements, and propositions to determine their truth or falsity based on established rules and principles. Logical fallacies are common mistakes that weaken arguments. They are flaws in the logic or rationality of a particular line of reasoning. Detecting logical fallacies is important because they can lead to incorrect conclusions, even when the argument's premise seems correct. Another reason for detecting fallacies is that they are frequently exploited to persuade the public in certain directions [7]. Here are the definitions of some common logical fallacies used in this study:

- Faulty Generalization: Drawing a conclusion about a population based on a sample that is not representative enough.
- Ad Hominem: Attacking the person making the argument instead of addressing the argument itself.

- Ad Populum: Arguing that a claim must be true because a large number of people believe it.
- False Causality: Assuming that because two events occurred together, one caused the other, without any evidence to support this.
- Circular Reasoning: Using the conclusion of an argument as one of the premises, thereby begging the question.
- Appeal to Emotion: Attempting to manipulate emotions rather than using valid reasoning to make an argument.
- Fallacy of Relevance: Bringing in information that is not logically relevant to the argument.
- Deductive Fallacy: A deductive fallacy is a pattern of reasoning that is flawed in its logical structure, making it invalid or unsound.
- Intentional: Assuming the intent or purpose of the argument.
- Fallacy of Credibility: Dismissing an argument based on the credibility of the source rather than the strength of the argument itself.
- False Dilemma: Presenting only two options when there are actually more available.
- Fallacy of Extension: Extending an argument beyond its logical limits, often resulting in a flawed conclusion.
- Equivocation: Using a term with multiple meanings in different parts of an argument, leading to a misleading conclusion.

Understanding these fallacies can help individuals evaluate arguments more objectively and make informed decisions based on valid reasoning and evidence. However, detecting logical fallacies through linguistic methods has been challenging because it requires the detector to find faults in the reasoning process [5, 8, 22]. A successful logical fallacy detector can significantly enhance the quality of communication and decision-making in the real-world. In education, it can be employed to teach students critical thinking skills by identifying fallacious reasoning in their essays and discussions. Journalists and fact-checkers can utilize it to ensure the integrity of their reporting, quickly spotting logical inconsistencies in statements and articles. In the political arena, it can aid in debate preparation and speech writing, ensuring arguments are logically sound and persuasive. Businesses can leverage fallacy detectors to refine marketing strategies and business proposals, enhancing their persuasive power and credibility. Legal professionals can strengthen their cases by identifying weaknesses in arguments, both in their own and their opponents' presentations. On a personal level, individuals can use fallacy detectors to improve their reasoning and decision-making, fostering more rational and effective communication in everyday interactions. Lately, automated tools and natural language processing (NLP) techniques are being developed to detect logical fallacies. These tools can analyze text for patterns associated with specific fallacies. Recent progress in large language models offer a promising path for research in this direction. In this study, we develop a logical fallacy detector which can detect if a certain text contains a logical fallacy and which type of fallacy it is. We use a publicly available logical fallacy dataset [13] and adapt it to include logically sound statements. We demonstrate the use of

a fine-tuned large language model for this classification and show that zero-shot chain of thought prompting [15] improves classification accuracy by guiding the LLM to reason step by step.

1.1 Related Works

The study of logical fallacies has a long history, dating back to ancient Greek philosophers. Aristotle (384–322 BC) is considered one of the earliest thinkers to systematically analyze logical reasoning and identify fallacies. In his works “Organon” and “On Sophistical Refutations,” he described various types of fallacies, including *ad hominem*, appeal to ignorance, and begging the question [12]. During the Middle Ages, scholars like Peter Abelard (1079–1142) and John Locke (1632–1704) further developed the understanding of fallacies and their role in logical reasoning. In the 19th century, logicians such as William Hamilton (1788–1856) and John Stuart Mill (1806–1873) made significant contributions to the study of fallacies, particularly in the realm of inductive reasoning. Philosophical works such as John Stuart Mill’s “System of Logic,” employ logical fallacies as a key component in reasoning [18]. The 20th century saw a resurgence of interest in fallacies, with philosophers like Irving M. Copi (1917–2002) and Charles L. Hamblin (1922–1985) publishing influential works [11] on the subject. Copi’s book “Introduction to Logic” (1953) [4] became a widely used textbook in the study of logic and fallacies.

Contemporary scholars, such as Douglas Walton [24] and Frans van Eemeren [23], have made significant contributions to the field, exploring the pragmatic and contextual aspects of fallacies, and developing techniques for identifying and evaluating arguments in real-world situations. The identification and understanding of logical fallacies have been crucial in various fields, including computer science and argumentation theory. The evolution of argumentation logic has been instrumental in explicating reasoning processes and has found applications in artificial intelligence, underscoring the relevance of fallacies in contemporary research [2]. The application of fallacy recognition in multitask instruction-based prompting and argumentation dialogues underscores the importance of detecting fallacies for both humans and machines [1, 25].

Recently, several studies have focused on leveraging machine learning techniques for logical fallacy detection. Jin et al. [13] introduced a methodology for logical fallacy detection using neural network architectures, which laid the groundwork for applying deep learning in argument analysis. Sourati et al. [19] further expanded this by developing a robust and explainable model for identifying logical fallacies in natural language arguments, emphasizing the importance of transparency and reliability in real-world applications. Sprenkamp, Jones, and Zavolokina [20] explored the use of large language models for propaganda detection, a task closely related to logical fallacy detection, demonstrating the versatility of these models in handling various forms of argumentative discourse.

In this study, we build on these works by fine-tuning GPT-3.5 turbo for logical fallacy detection, a novel approach that combines the strengths of previous methodologies with the enhanced capabilities of recent large language models.

Unlike prior studies, our work specifically focuses on the application of zero-shot chain of thought (CoT) prompting to improve classification accuracy.

1.2 Large Language Models in Language Comprehension

Large language models (LLMs) such as BERT, GPT, and LLAMA have been successfully applied to various tasks, including question answering, language inference, and code evaluation and repair [6]. LLMs are based on transformer architectures, which have proven to be more effective than traditional recurrent or convolutional neural networks. LLMs have shown the potential to acquire human-like language abilities through statistical learning from vast linguistic experience [14]. Furthermore, they have enabled few-shot learning, significantly reducing the need for task-specific training examples [3]. LLMs have also been utilized in the translation of unstructured natural language to formal specifications [10], demonstrating their versatility and potential in understanding and processing complex linguistic structures [17]. The impact of LLMs has also been observed in conversation comprehension, where advancements in neural language modeling have led to significant progress [9]. The fine-tuning of pre-trained LLMs remains the standard approach for downstream tasks, showcasing their versatility and adaptability [16].

1.3 Fine Tuning LLMs

The fine-tuning process typically involves adjusting the pre-trained model with an additional output layer to suit the specific task at hand. The improved performance obtained by fine-tuning large language models has been demonstrated in various NLP tasks, including text classification, short answer grading, offensive language detection, and dialogue systems [21, 26].

2 Methodology

2.1 Dataset

Data Sources and Preprocessing

This dataset was initially sourced from the dataset by Jin et al. [13] had several issues:

- Quiz questions and definitions instead of fallacy examples.
- Spanish text and duplicate examples.
- Only questions without examples.
- Incorrect or mislabeled fallacies.
- Poor source referencing.

This initial dataset underwent extensive cleaning and annotation to improve its quality for machine learning training. The cleaning process included identifying and removing incorrect examples, translating non-English text, ensuring

each entry contained a clear fallacy example, correcting mislabeled entries, and improving source referencing. This resulted in the more robust and tidy “Fixed-Logic” dataset.

Additionally, logically sound samples were extracted from the Stanford Natural Language Inference (SNLI) corpus (version 1.0), which contains 570,000 human-written English sentence pairs. Each pair was manually labeled for balanced classification across three categories: entailment, contradiction, and neutral. To generate logically sound samples, we specifically selected sentence pairs labeled as ‘entailment.’ These sentences were then concatenated, ensuring that the second sentence logically followed from the first, thereby forming a cohesive and logically sound statement. These statements are labeled Non-fallacy in our dataset. Given below are example statements for each class in the dataset,

ad populum: “This makes you think you need to believe or buy something because everyone else is. Which technique is it?”

faulty generalization: “A few students are misbehaving...therefore the whole class is bad.”

ad hominem: “My teacher always brings up stuff about Women’s Rights. She only cares about it because she is a woman. The stats she shows us are probably biased.”

Non-Fallacy: Seeing that group of men sitting at a table turn to have their picture taken., it can be deduced that the group of men are posing for a picture.

false dilemma: “Judge Danforth in the play, ‘The Crucible,’ offers Proctor two choices: you are either with this court or against it.”

intentional: “Don: If you drink alcohol, it will kill any virus you might have. Tony: What evidence do you have to support that? Don: I don’t need evidence. It is common sense.”

circular reasoning: “Plagiarism is deceitful because it is dishonest.”

fallacy of credibility: “The medicine man rolled into town on his bandwagon offering various natural remedies, such as very special plain water. He said that it was only natural that people should be wary of ‘artificial’ medicines such as antibiotics.”

false causality: ‘Bush was “determined to knock down Saddam Hussein” because of his “nuclear bomb potential.”’

appeal to emotion: “If we don’t teach teens to work harder, the human race is doomed.”

fallacy of extension: “Al Gore feels that all companies are irresponsible and should be punished for allowing emissions which causes global warming.”

fallacy of relevance: “My opponent says I am weak on crime, but I have been one of the most reliable participants in city council meetings.”

equivocation: “All stars are exploding balls of gas. Miley Cyrus is a star. Therefore, Miley Cyrus is an exploding ball of gas.”

miscellaneous: “I know this relationship isn’t working anymore and that we’re both miserable. No marriage. No kids. No steady job. But I’ve been with him for seven years, so I’d better stay with him.”

For model fine-tuning, we compiled a master dataset comprising 1960 records. Table 1 shows the distribution of classes in the dataset expressed as percentages:

Table 1. Distribution of Logical Fallacies in the dataset

Logical Fallacy Type	Percentage
Faulty Generalization	16.48%
Non Fallacy	15.05%
Ad Hominem	10.87%
Ad Populum	8.47%
False Causality	7.76%
Circular Reasoning	6.48%
Appeal to Emotion	6.43%
Fallacy of Relevance	5.97%
Intentional	5.41%
Fallacy of Credibility	5.20%
False Dilemma	5.15%
Fallacy of Extension	4.69%
Equivocation	1.94%
Miscellaneous	0.10%

2.2 Data Preprocessing

The dataset preparation involved several key steps to ensure the data was suitable for fine-tuning the GPT-3.5 model in a chat-completion format. The goal was to create a model that could accurately classify logical fallacies from textual content. Here's how the dataset was prepared: The original dataset was split into training and validation sets using an 80/20 ratio. Stratified sampling based on the updated label column ensured that each class of logical fallacies was proportionately represented in both sets. To avoid having rows of similar classes clustered together, both the training and validation datasets were shuffled. This randomization helps prevent the model from learning any potential order biases in the data. The training data was formatted into a specific JSONL structure to meet the input requirements of the GPT-3.5 Turbo model. Each entry was structured as a chat conversation where: The system presents a list of possible logical fallacies and asks the user to classify a given statement. The user provides a sentence from the source article. The assistant (model) responds with the logical fallacy classification based on the ground truth.

2.3 Training and Fine-Tuning LLMs

In this study, we applied fine-tuning techniques to OpenAI's GPT-3.5-turbo-0125 model to tailor it for the task of logical fallacy detection. Fine-tuning was conducted as follows:

Base Model: GPT-3.5 **Training Data:** The training dataset consisted of logically annotated sentence pairs, as previously described. **Training Tokens:** 853,347 tokens were used during the training process. **Epochs:** The model was trained for 3 epochs to balance between underfitting and overfitting. **Batch Size:** A batch size of 4 was employed, optimizing the balance between training speed and memory utilization. **Learning Rate Multiplier:** A learning rate multiplier of 2 was used to adjust the learning rate during the training process. **Random Seed:** The seed used for initializing the training process was 448608511, ensuring reproducibility of the training results. This fine-tuning process was designed to enhance the model's ability to accurately identify and classify different types of logical fallacies based on the nuances captured in the training data. The custom training parameters were selected to maximize model performance while maintaining computational efficiency.

For the validation phase of our study, a specific prompt was crafted to test the fine-tuned GPT-3.5 model's ability to classify sentences based on logical fallacies. The prompt began with an instruction to the model, "Remember to think step by step," aimed at guiding the model to process the input methodically. Following this, a list of possible classifications was provided. See Fig. 1 for the structure of the prompt.

The model was then presented with a sentence (referred to as the source article) and required to respond with only one of the listed logical fallacies that best described the logical flaw (if any) within the sentence. This prompt format was meticulously designed to assess the model's precision in identifying the correct type of logical fallacy from a set of predefined categories, thus measuring its classification accuracy effectively.

```
[faulty generalization,
false causality,
circular reasoning,
ad populum,
ad hominem,
Fallacy of extension
fallacy of logic,
appeal to emotion,
false dilemma,
equivocation,
fallacy of extension,
fallacy of relevance,
fallacy of credibility,
intentional,
miscellaneous,
Non Fallacy]
Please only respond with one of the above logical fallacies for the
following sentence, Remember to think step by step:
```

Fig. 1. Zero shot chain of thought prompt (highlighted in yellow) (Color figure online)

2.4 Model Evaluation Metrics

Precision measures the ability of the classifier to correctly identify instances of a specific class among all instances it classified as that class. It is defined as the ratio of true positives to the sum of true positives and false positives for a given class:

$$\text{Precision}_i = \frac{\text{True Positives}_i}{\text{True Positives}_i + \text{False Positives}_i}$$

where True Positives_i are the number of correctly predicted instances of class i , and False Positives_i are the number of instances incorrectly predicted as class i .

Recall measures the ability of the classifier to correctly identify instances of a specific class among all instances that truly belong to that class. It is defined as the ratio of true positives to the sum of true positives and false negatives for a given class:

$$\text{Recall}_i = \frac{\text{True Positives}_i}{\text{True Positives}_i + \text{False Negatives}_i}$$

where False Negatives_i are the number of instances of class i that were incorrectly classified as another class. The F1 score is the harmonic mean of precision and recall for a given class. It is defined as:

$$F1_i = \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

The overall F1 score can be computed by averaging the F1 scores for all classes, usually using either micro-averaging, macro-averaging, or weighted averaging, depending on the specific evaluation scenario.

3 Results

The overall accuracy of the model was 80.82%, with a weighted average precision of 82.14%, recall of 80.82%, and F1 score of 0.8123. From the confusion matrix in Fig. 2 it is clear that the gpt3.5 model is successful at distinguishing between the logical fallacy classes. Table 2 further shows that the model classified non-fallacy examples with the highest level of accuracy. While equivocations and miscellaneous classes had the lowest level of accuracy. This is understandable due to the small number of training examples in that class.

The ROC curve in Fig. 3 confirms the above results, with the non-fallacy curve having the highest area under the curve and equivocation and miscellaneous having the lowest.

3.1 Ablation Study

To understand the effects of the zero shot CoT prompt of the classification performance, we conducted an ablation study. Without the CoT prompt, the

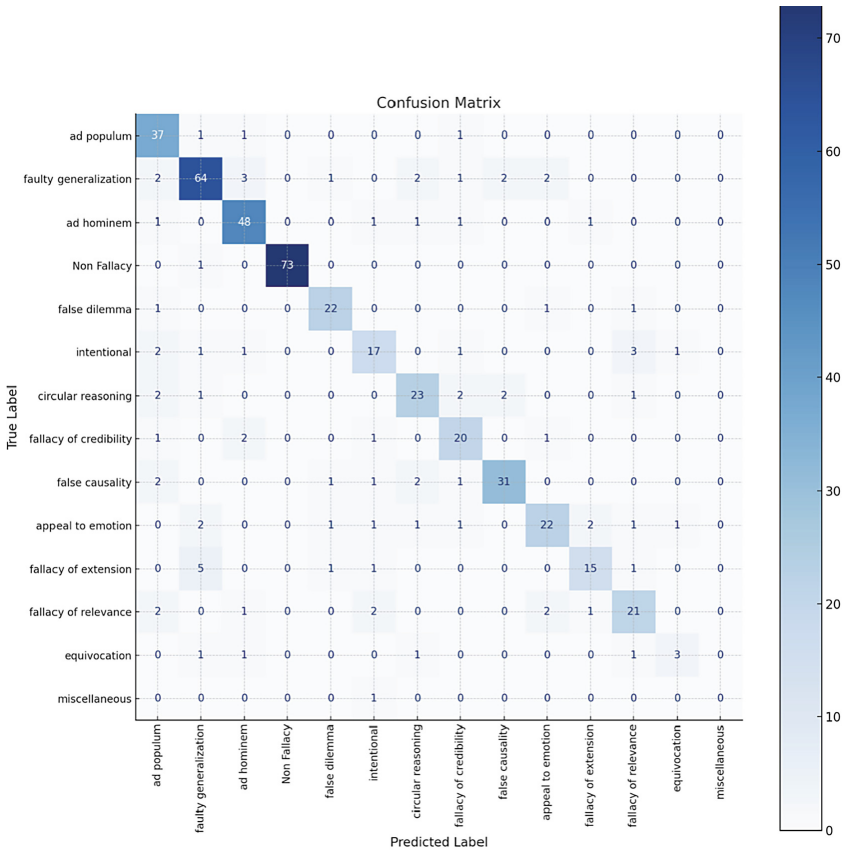


Fig. 2. Confusion matrix for the fine-tuned GPT-3.5 model

Table 2. Performance metrics of the fine tuned model for each logical fallacy

Class	Precision	Recall	F1 Score
Non Fallacy	1.000	0.986	0.993
Ad Hominem	0.842	0.906	0.873
Ad Populum	0.740	0.902	0.813
Appeal to Emotion	0.786	0.688	0.733
Circular Reasoning	0.767	0.719	0.742
Equivocation	0.600	0.333	0.429
Fallacy of Credibility	0.714	0.800	0.755
Fallacy of Extension	0.789	0.652	0.714
Fallacy of Relevance	0.724	0.700	0.712
False Causality	0.886	0.816	0.849
False Dilemma	0.846	0.880	0.863
Faulty Generalization	0.842	0.790	0.815
Intentional	0.680	0.654	0.667
Miscellaneous	0.000	0.000	0.000

model F1 score fell to 0.7947. Accuracy was 78.57%. Precision was 81.17% and Recall was 78.57%. These results show that the CoT prompting does improve the classification performance for this problem.

4 Discussion

The model achieved good classification performance across various types of fallacies, with an F1 score of 0.99 for “Non Fallacy,” indicating robust performance in distinguishing non-fallacious statements from those containing logical fallacies. This high F1 score suggests that the model is adept at accurately identifying non-fallacious statements, minimizing both false positives and false negatives in this category. However, some fallacies, such as “Equivocation” and “Miscellaneous,” showed lower performance. It remains to be seen if more data could improve the classification performance for these classes.

The use of zero-shot chain of thought prompting improved the model’s classification accuracy, suggesting that prompting strategies can help improve model performance for this task. This is understandable due to the logical nature of the problem. This finding aligns with recent advancements in LLMs, where prompting techniques have been shown to guide models towards more accurate and context-aware responses.

Though the models perform better than previous studies, they lack interpretability due to the proprietary nature of gpt3.5. The ability to detect fallacies in real-time within online discourse, such as social media platforms, can help curb the spread of misinformation and improve the quality of public discus-

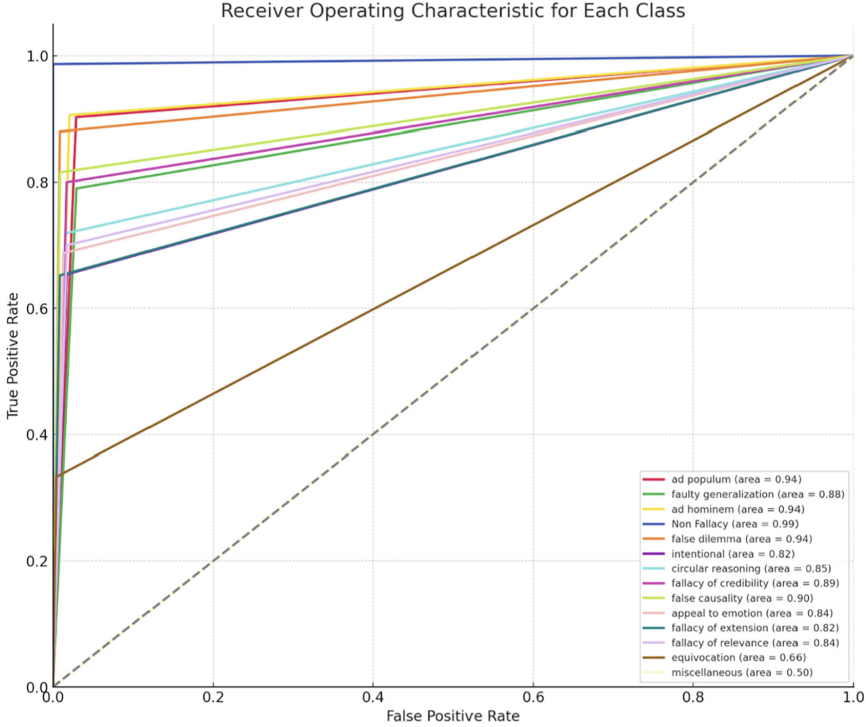


Fig. 3. Receiver Operating Characteristics curve for the fine-tuned GPT-3.5 model

sions. Implementing these systems effectively requires addressing interpretability challenges.

5 Conclusion

This study highlights the potential of fine-tuned LLMs like GPT-3.5 to serve as effective tools for detecting logical fallacies. Beyond achieving high accuracy in distinguishing fallacious and non-fallacious statements, this work underscores the broader importance of automated reasoning systems in promoting critical discourse across various domains. The real-world applications of this approach extend from education and journalism to social media moderation and legal analysis, where enhancing argument quality is vital. By addressing existing challenges, such as dataset imbalances and model interpretability, the field can move closer to creating robust, universally applicable solutions. Future research should explore integrating logical fallacy detection into multi-task systems that combine sentiment analysis, misinformation detection, and cross-linguistic argument evaluation. Such advancements hold promise for fostering more rational, informed discussions, not just in academic and professional circles, but in everyday interactions. This work serves as a foundation, inviting further innovation and collab-

oration toward a future where flawed reasoning is minimized and critical thinking is maximized.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alhindi, T., Chakrabarty, T., Musi, E., Muresan, S.: Multitask instruction-based prompting for fallacy recognition (2023). N/A
2. Almpiani, S., Lisanyuk, E., Schumann, A.: Trends in argumentation logic. *Studia Humana* **11**, 1–5 (2022)
3. Chowdhery, A., et al.: Palm: scaling language modeling with pathways (2022). N/A
4. Copi, I.M., Cohen, C., McMahon, K.: *Introduction to Logic*. Routledge (2016)
5. Damer, A.F.R.: *A Practical Guide to Fallacy-Free Reasoning*. Wadsworth Cengage Learning (2009)
6. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding (2018). N/A
7. Gena, C., Grillo, P., Lieto, A., Mattutino, C., Vernerio, F.: When personalization is not an option: an in-the-wild study on persuasive news recommendation. *Information* **10**, 300 (2019)
8. Govier, T., et al.: *A Practical Study of Argument*. Wadsworth Belmont, CA (2010)
9. Gupta, S., Rawat, B.P.S., Yu, H.: Conversational machine comprehension: a literature review. In: *Proceedings of the 28th International Conference on Computational Linguistics* (2020)
10. Hahn, C., Schmitt, F., Tillman, J., Niklas, M., Siber, J., Finkbeiner, B.: Formal specifications from natural language (2022). N/A
11. Hamblin, C.L.: Fallacies. In: *Advanced Reasoning Forum* (2022)
12. Hansen, H.V.: Aristotle, whately, and the taxonomy of fallacies. In: *Practical Reasoning*, pp. 318–330 (1996)
13. Jin, Z., et al.: Logical fallacy detection. arXiv preprint [arXiv:2202.13758](https://arxiv.org/abs/2202.13758) (2022)
14. Kallens, P.C., Kristensen-McLachlan, R.D., Christiansen, M.H.: Large language models demonstrate the potential of statistical learning in language. *Cognit. Sci.* **47**, e13256 (2023)
15. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. *Adv. Neural. Inf. Process. Syst.* **35**, 22199–22213 (2022)
16. Li, X.L., Liang, P.: Prefix-tuning: optimizing continuous prompts for generation. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Confer* (2021)
17. Matthias, C., Hahn, C., Mendoza, D., Schmitt, F., Trippel, C.: Nl2spec: inter-actively translating unstructured natural language to temporal logics with large language models. In: Enea, C., Lal, A. (eds.) *Computer Aided Verification. CAV 2023. LNCS*, vol. 13965. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-37703-7_18
18. Rosen, F.: The philosophy of error and liberty of thought: J.S. mill on logical fallacies. *Informal Logic* **26**, 121 (2008)
19. Sourati, Z., Venkatesh, V., et al.: Robust and explainable identification of logical fallacies in natural language arguments. *Knowl. Based Syst.* **266**, 110418 (2023)

20. Sprenkamp, K., Jones, D.G., Zavolokina, L.: Large language models for propaganda detection. arXiv preprint [arXiv:2310.06422](https://arxiv.org/abs/2310.06422) (2023)
21. Sung, C., Dhamecha, T.I., Saha, S., Ma, T., Reddy, V.L., Arora, R.: Pre-training BERT on domain resources for short answer grading. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conferen (2019)
22. Van Eemeren, F.H., Grootendorst, R., Johnson, R.H., Plantin, C., Willard, C.A.: Fundamentals of Argumentation Theory: A Handbook of Historical Backgrounds and Contemporary Developments. Routledge (2013)
23. Van Eemeren, F.H., Grootendorst, R., Grootendorst, R.: A Systematic Theory of Argumentation: The Pragma-Dialectical Approach. Cambridge University Press (2004)
24. Walton, D.N.: A Pragmatic Theory of Fallacy Studies in. N/A (1995)
25. Wardeh, M., Bench-Capon, T., Coenen, F.: Padua: a protocol for argumentation dialogue using association rules. *Artif. Intell. Law* **17**, 183–215 (2009)
26. Wiedemann, G., Yimam, S.M., Biemann, C.: UHH-LT at Semeval-2020 task 12: fine-tuning of pre-trained transformer networks for offensive language detection. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation (2020)