Large Language Models are Few-Shot Training Example Generators: A Case Study in Fallacy Recognition

Anonymous ACL submission

Abstract

Recognizing fallacies is crucial for ensuring the quality and validity of arguments across various domains. However, computational fallacy recognition faces challenges due to the diverse genres, domains, and types of fallacies found in datasets. This leads to a highly multi-class, and even multi-label, setup with substantial class imbalance. In this study, we aim to enhance existing models for fallacy recognition by incorporating additional context and by leveraging large language models to generate synthetic data, thus increasing the representation of the infrequent classes. We experiment with GPT3.5 to generate synthetic examples and we examine the impact of prompt settings for this. Moreover, we explore zero-shot and few-shot scenarios to evaluate the effectiveness of using the generated examples for training smaller models within a unified fallacy recognition framework. Furthermore, we analyze the overlap between the synthetic data and existing fallacy datasets. Finally, we investigate the usefulness of providing supplementary context for detecting fallacy types that need such context, e.g., diversion fallacies. Our evaluation results demonstrate consistent improvements across fallacy types, datasets, and generators. We will release the code and synthetic dataset upon the acceptance of the paper.

1 Introduction

Fallacies are common errors in reasoning that can mislead and invalidate arguments. The capacity to discern fallacies is fundamental to sustaining the robustness and authenticity of arguments across various domains, such as public policy, legal reasoning, and scientific discourse (Bailin and Battersby, 2016). In recent years, the task of automated fallacy recognition has attracted significant interest from researchers in the fields of Natural Language Processing (NLP) and Artificial Intelligence (AI) (Amgoud and Besnard, 2013; Hamblin, 2022; Goffredo et al., 2022; Alhindi et al., 2022; Jin et al., 2022). Nevertheless, numerous challenges persist, including the multiplicity of genres, domains, and fallacy types, which contribute to a complex multi-class and multi-label task structure compounded by class imbalances in datasets.

Current state-of-the-art models struggle with the recognition of underrepresented fallacies, which may often require additional context for accurate identification, such as diversion fallacies (Walton, 1996). Moreover, the variety of fallacies coupled with the broad range of contexts in which they may occur necessitates a comprehensive and diverse dataset for training these models. One strategy to combat the challenge of sparse and imbalanced data in machine learning is data augmentation (Wang et al., 2017) by creating synthetic examples, thereby enhancing the dataset size and diversity and improving the performance of the machine learning model. Recent advancements in large language models like GPT-3,3.5,4 (Brown et al., 2020) provide a promising avenue for generating high-quality synthetic examples for fallacy recognition.

The existing work in computational models for fallacy recognition is still in its early stages, with limited datasets available. These datasets cover different types of fallacies in various contexts, such as question and answer dialog moves, name-calling in social media debates, logical fallacies from educational websites, and fallacies related to Covid-19 misinformation in social media and news. Previous work has focused on detecting fallacies in individual datasets, using techniques like fine-tuning transformers for sequence tagging (Goffredo et al., 2022), and training structure-aware classifiers (Jin et al., 2022). However, fallacy recognition is challenging due to the high number of classification labels, class imbalance in datasets, limited dataset sizes, and poor out-of-distribution generalization. Alhindi et al. (2022) proposes a multitask framework using T5, which converts fallacy types into

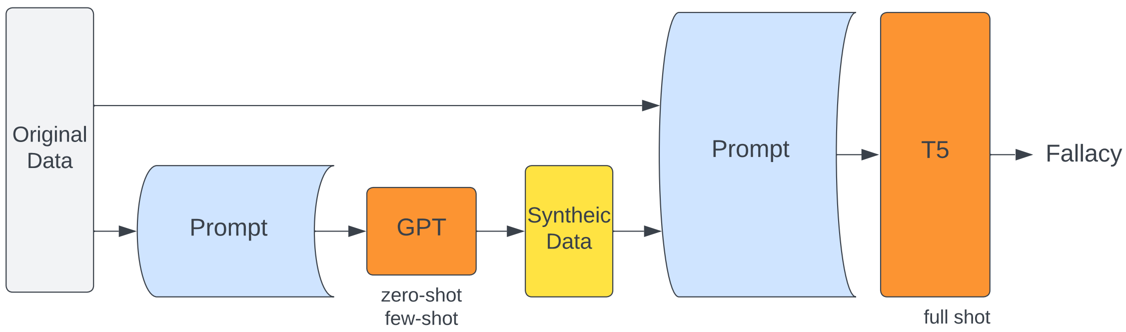

Figure 1: Our data augmentation and model training pipeline.

natural language instructions, and thus approaches the differences between fallacy datasets as different tasks, but their approach does not detect infrequent classes effectively. Goffredo et al. (2022) incorporate argumentation features to detect fallacies in political debates, while Jin et al. (2022) trains a structure-aware classifier on fallacies from educational websites; however, they both focus on a single fallacy scheme from one dataset while we include multiple fallacy schemes.

We extend previous work on generic fallacy recognition by exploring the capabilities of large language models to generate synthetic data that augments manually labeled datasets. We study the effect of the data generated under zero-shot and few-shot conditions on the downstream task of fallacy recognition. We also analyze the quality of the synthetic data and its similarity to the fallacy datasets. Figure 1 shows an overview of our approach of using GPT3.5 to generate additional training examples in zero/few-shot settings, then training a T5 model (Raffel et al., 2020) for fallacy recognition on a combination of the original and the synthetic data.

The rest of the paper is organized as follows: In Section 2, we describe the fallacy datasets included in this work. Then, we present our synthetic data generation approach in Section 3 and experimental setup in Section 4. We show the results in Section 5 and analysis of the similarity between original and synthetic data in Section 6. We then present an overview of related work and conclude in Sections 7 and 8, respectively.

2 Fallacy Datasets

We experiment with the five fallacy datasets covered by Alhindi et al. (2022). They include fallacies in question-answer pairs in game settings (ARGOTARIO) (Habernal et al., 2017), 18 propaganda techniques in news articles (PROPAGANDA) (Da San Martino et al., 2019) that are recently extended to 23 (Piskorski et al., 2023) techniques, logical fallacies from educational websites (LOGIC) (Jin et al., 2022), and fallacies in misinformation around covid in social media (COVID) (Musi et al., 2022) and climate change news articles (CLIMATE) (Alhindi et al., 2022). These datasets identify different fallacy types and range from 5 to 18 fallacies. Alhindi et al. (2022) unified the fallacy types from the four schemes and introduced 28 fallacy types in one unified scheme. These dataset are different in size as they go from a few hundred examples (450-880) for CLIMATE, COVID and ARGOTARIO, to a few thousands (4,500 to 5,100) for LOGIC and PROPAGANDA. The total number of examples per fallacy type varies significantly as it ranges from less than 100 examples for some fallacies (e.g., *False Analogy, Strawman, Whataboutism*) to more than 1,000 examples (e.g., *Hasty Generalization, Name Calling or Labeling, Loaded Language*). Detailed numbers for each fallacy per dataset and split can be found in Alhindi et al. (2022).

One main challenge in these datasets is the high imbalance frequency of classes in a high multi-class, and even multi-label task. The unified model presented by Alhindi et al. (2022) improves the overall results but still performs much better on more frequent classes, thus we utilize data augmentation to address this challenge. In addition, two of the five fallacy datasets: PROPAGANDA and CLIMATE are from news articles where the fallacy is annotated at the sentence or fragment level. Therefore, we study the benefit from providing additional context to the fallacious segment (sentence or fragment) by including the preceding or succeeding sentence when available.

3 Synthetic Data Generation

To generate additional examples for infrequent fallacy classes, we leverage `gpt-3.5-turbo` (henceforth called GPT3.5), a conversational language model, as a data augmentation tool. We explore zero-shot, 1-shot, 2-shot, and 5-shot settings to generate examples that have not been seen in the original training data. These generated examples provide diversity and help address the data scarcity issue for less frequent fallacy types.

In order to understand the capabilities of pretrained large language models such as GPT3.5 in producing synthetic data, we control the information provided in the prompts as follows: i) zeroshot prompts that have no fallacy example and ask the model to generate an example in one form (e.g. sentence, tweet, question-answer pair) for a certain fallacy provided in the prompt; ii) fewshot prompts that list the fallacy type and output form in addition to providing 1 to 5 examples for the given fallacy type in the prompt. The model is asked to generate the same number of examples given in the prompt (i.e. 1-shot prompt asks the model to generate 1 example, 5-shot ask for 5 new examples and so on); iii) fewshot-context prompts that provide the examples of fallacy and their wider context if available (previous and next sentence) and asks the language model to do the same by generating both examples for a certain fallacy and their contexts. Figure 2 shows an example of the 1-shot-context prompt of the *Irrelevant Authority* fallacy from the PROPAGANDA dataset.

For all data augmentation settings, we generate the same number of examples per fallacy and thus study the quality of the synthetic data using different prompts for generation. The aim is to address data imbalance in these datasets so we exclude fallacies that are very frequent, especially if they only exist in one fallacy scheme and thus have a high number of examples with low diversity. Following this criteria, we exclude *Loaded Language* and *Name Calling or Labeling* that only appear in PROPAGANDA. We also do not generate examples for *Hasty Generalization* in a form similar to the one found in the LOGIC dataset, but we generate ones in Covid-19 and climate change domains since their respective datasets has this fallacy but in very low counts. For all generated fallacies we double the number of examples with respect to the number of original examples for a certain fallacy thus maintain comparable ratios of both original and synthetic data. Also, we cap the number of synthetic examples for each fallacy to 100 examples generated from each dataset thus changing the distribution of the training set to bring the very infrequent classes closer to the overall average number of examples per class.

Prompt

Your task is to perform the following actions:
1. Read the sentence that has the fallacy of Irrelevant Authority and its provided context. The sentence and context below are both delimited by <>.
2. Generate a similar example of one sentence and its context of the Irrelevant Authority fallacy. The context should be a bigger chunk of text that includes the sentence, similar to the provided example sentence and context below.

Your generated output should be in JSON format with the following keys: generated_sentence, generated_context.

Sentence: <Carlson cited Dr. Robert Epstein who has said, in Carlson's words, "Google alone could determine the outcome of almost any election just by altering its search selections and we would never know it.">

Context: <Carlson cited Dr. Robert Epstein who has said, in Carlson's words, "Google alone could determine the outcome of almost any election just by altering its search selections and we would never know it.". Dr. Robert Epstein is a social scientist and an expert on Google. >

Generated Output

{

'generated_sentence': 'As a dog groomer, I can tell you that the earth is flat.',

'generated_context': 'While most people accept that the earth is round, one dog groomer is convinced otherwise. As a dog groomer, I can tell you that the earth is flat. Dogs are great navigators, and they always sense when the ground is flat or sloping.'

}

Figure 2: Example of 1-shot context prompt from the Propaganda dataset of a sentence that has the *Irrelevant Authority* fallacy and its context.

4 Experimental Setup

Similar to Alhindi et al. (2022), we use the T5 model (Raffel et al., 2020), a versatile text-to-text transformer, as the backbone for fallacy recognition by fine-tuning instruction-based prompts on all fallacy datasets. The prompts are designed to provide explicit instructions on identifying specific fallacies, enabling targeted learning within the model. This approach is inline with a large body of research that utilizes instruction-tuning of large language models on many tasks (Wei et al., 2021; Sanh et al., 2022).

We evaluate the proposed approach on the five fallacy datasets. We train the T5 model using a combination of the original labeled data and the generated examples from GPT3.5. We compare the performance of the model under different settings, including zero-shot, 1-shot, 2-shot, and 5-shot scenarios, with and without additional context to understand the impact of prompt and data availability on fallacy recognition.

All fallacy examples, original and synthetic, are transformed into instruction-based prompts that are used to fine-tune the T5-3 Billion model (henceforth T53B) in a multitask fashion. The model and hyperparameters are fixed and we only change the training data that is fed into the model with the aim to study the ability of a smaller size model such as T53B to learn from manually annotated or crowdsourced data as well as synthetically generated data from a larger size model such as GPT3.5. We show the results of all training conditions for the PROPAGANDA and CLIMATE datasets in Table 1 and for the ARGOTARIO, LOGIC, and COVID datasets in Table 2. In both tables, we report the overall accuracy and macro F1 scores for each dataset as well as the F1 scores for each fallacy class. The results cover nine training conditions where we train on the original training set only (baseline-N (no-context)) similar to (Alhindi et al., 2022), and baseline-C with context for datasets that are from news articles. This applies to the PROPAGANDA and CLIMATE where this context is available. The remaining seven training conditions all use a different form of data augmentation depending on the number of examples provided in the prompt during synthetic data generation which includes zero, one, two, and five examples. All data augmentations experiments are done with context (C columns) and no-context (N columns), except zero-shot prompts that were not enough for GPT3.5 to provide usable examples with contexts in most cases without providing at least one example in the prompt for GPT3.5 to follow. Therefore zero-shot prompts are only reported in no-context settings. We discuss below the effect of data augmentation and adding context in more details.

5 Results

5.1 Data Augmentation

Overall Results. Adding synthetic data to the original dataset improves the results over the baselines where only the original training data is used regardless of the data augmentation method. This is true for both the overall accuracy and macro-F1 scores in all five datasets as shown in Tables 1 and 2 whether the context is provided or not. Interestingly, 1-shot prompts seem to yield the best results when compared to both zero-shot and other few-shot settings. This results is counter to what we initially expected. We hypothesized that 5-shot prompts that have five examples of a fallacy and ask GPT3.5 to generate five similar examples to yield synthetic data that is more generic to the fallacy (the one factor that is common among the five examples in the prompt), and therefore would help train a model for fallacy recognition to be more resilient. The first part of our hypothesis seem to be correct i.e. the synthetic examples are less similar than the ones provided in the prompt as we explore in detail in Section 6. However, having less similar examples to the training data does not help the model perform better on these benchmark test sets for fallacy recognition.

Per-Class Results. Some fallacies show massive gains after data augmentation compared to others. This is true in the LOGIC dataset where the *Equivocation* and *Fallacy of Extension* are among fallacies with the biggest gains over baselines. These two fallacies are also the least frequent in the LOGIC dataset and thus the impact of data augmentation is bigger. The diversion fallacies in PROPAGANDA e.g., *Red Herring, Strawman, Whataboutism* are particularly challenging in baseline settings due to their low counts and complexity since they typically require external information to the fallacious segment to be properly recognized, which is especially the case for *Strawman* where all models fail to make any correct prediction with or without data augmentation. However, for the other two (*Red Herring* and *Whataboutism*), significant gains are observed with data augmentation particularly for *Whataboutism* in 1-shot settings where the f1-scores jumps to 0.63 compared to 0 in the baseline models.

Some fallacy types present a different level of challenge across datasets due to their format in a particular dataset, frequency, and the fallacies they

Dataset	Fallacy	baseline		Data Augmentation: zero-shot	1-shot		2-shot		5-shot	
		N	C	N	N	C	N	C	N	C
Propaganda	Black and White Fallacy	.14	.34	**.39**	**.39**	.29	.35	.33	.36	.34
	Causal Oversimplification	.34	.27	.41	**.48**	.27	.39	.23	.44	.29
	Doubt	.61	.66	.67	.66	.69	.66	**.71**	.69	.68
	Exaggerate/Minimization	.34	.32	.44	**.58**	.55	**.58**	.47	**.58**	.49
	Fear or Prejudice	.49	.44	.49	.54	.49	**.67**	.46	.51	.50
	Flag-Waving	.64	.67	.67	.68	.67	.67	.67	**.69**	**.69**
	Irrelevant Authority	.26	.30	.44	.46	.44	**.47**	.40	.41	.38
	Loaded Language	.79	.76	.81	**.83**	.81	**.83**	.79	**.83**	.80
	Name Calling, Labeling	.83	.79	.83	.85	.82	.85	.81	**.86**	.81
	Red Herring	0	0	0	**.29**	.22	0	0	0	0
	Reductio Ad Hitlerum	.17	.18	.29	.40	.27	**.44**	.25	.40	.22
	Slogans	.49	.45	.59	.56	.52	.55	.51	**.67**	.48
	Strawman	0	0	0	0	0	0	0	0	0
	Thought-Termin. Cliches	.29	.34	.29	**.50**	.40	.36	.41	.38	.39
	Whataboutism	0	0	.29	**.63**	.62	.53	.53	.48	.47
	Accuracy	.68	.67	.71	**.74**	.71	.73	.70	**.74**	.70
	Macro	.36	.37	.44	**.52**	.47	.48	.44	.49	.44
Climate	Causal Oversimplification	.35	.33	.40	.53	.32	.42	.30	**.60**	.37
	Cherry Picking	.44	.41	.43	**.48**	.44	.43	.41	.46	.45
	Evading Burden of Proof	0	0	0	**.17**	.12	0	.10	0	0
	False Analogy	0	0	.36	**.62**	.18	.35	.17	.43	.17
	Hasty Generalization	0	0	0	0	0	0	0	0	0
	Irrelevant Authority	.22	.25	.31	.31	.31	.31	.31	**.43**	.33
	Red Herring	0	0	.12	.11	0	**.18**	0	**.18**	0
	Strawman	.22	0	.40	.40	.40	.36	.50	**.55**	.40
	Vagueness	.37	.39	.34	**.40**	.29	.36	.36	.36	.24
	Accuracy	.30	.28	.34	**.40**	.29	.34	.29	.39	.30
	Macro	.18	.15	.26	**.33**	.23	.27	.24	**.33**	.22

Table 1: F1 scores on the Propaganda and Climate datasets using multitask training of T53B model. **N**: no context to the fallacious segment added. **C**: context of previous and next sentence to the fallacious segment provided.

are listed with in the prompt at inference time. For example, *Red Herring* is easier to detect in ARGOTARIO and LOGIC in the baseline model to begin with due to a lower number of fallacies in ARGOTARIO, and lack of other diversion fallacies in those two scheme thus making them more distinct than the other fallacies and easier to distinguish. However, for PROPAGANDA and CLIMATE, the baselines get 0 f1-scores for *Red Herring* and data augmentation helps in improving the results to 0.29 in 1-shot for PROPAGANDA and 0.18 in 2-shot and 5-shot settings in CLIMATE. Some fallacy types remain challenging to detect with any kind of data augmentation, such as *Strawman* in PROPAGANDA, and *Hasty Generalization* in CLIMATE given their low counts in the test set (e.g., 2-5 examples) and therefore the test sets might have one particular form of this fallacy rather than represent the fallacy type in general. Having a train-test split that can truly evaluate the performance of machine learning models for this task is not trivial due to the high number of classes and the severe data imbalance.

5.2 Effect of Additional Context

The use of context during training is different for PROPAGANDA and CLIMATE in Table 1 compared to the other three datasets shown in Table 2. The difference between each N and C columns in Table 1 is rather than only providing a fallacious segment, we provide a wider context window of the previous and next sentence when available for the two datasets listed in the table. However, there is no difference

Dataset	Fallacy	baseline		Data Augmentation						
				zero-shot	1-shot		2-shot		5-shot	
		N	C	N	N	C	N	C	N	C
Argotario	Ad Hominem	.59	.63	.63	**.71**	.64	.62	.64	.65	.63
	Emotional Language	.64	.68	.70	**.71**	.70	.67	.69	.60	.65
	Hasty Generalization	.46	.44	.51	.47	.54	.47	.52	**.55**	.49
	Irrelevant Authority	.71	.72	**.80**	.75	.78	.74	.77	.75	.78
	Red Herring	.32	.42	.47	.50	.46	.44	.51	**.53**	.52
	Accuracy	.56	.57	.61	.61	.60	.59	**.63**	.61	.62
	Macro	.54	.58	.62	.61	.61	.58	**.63**	.61	.61
Logic	Ad Hominem	.77	.81	.87	**.88**	**.88**	**.88**	.85	.86	**.88**
	Ad Populum	.81	.80	.82	**.89**	.87	.86	.86	**.89**	.85
	Black and White Fallacy	.84	.84	.91	.91	.89	**.92**	.89	.91	**.92**
	Causal Oversimplification	.65	.70	.81	.79	**.82**	.81	.80	.78	.79
	Circular Reasoning	.57	.56	.68	**.84**	**.84**	.80	.77	.76	.83
	Deductive Fallacy	.32	.29	.48	**.69**	.57	.57	.54	.56	.57
	Emotional Language	.55	.53	.65	.76	**.77**	.72	.74	.71	.68
	Equivocation	.22	0	.27	**.57**	.43	.55	.39	.43	.42
	Fallacy of Extension	.08	.04	.48	**.68**	**.68**	.64	.60	.58	.64
	Hasty Generalization	.64	.63	.72	**.80**	.75	.77	.75	.77	.79
	Intentional Fallacy	.09	.15	.16	**.55**	.48	.46	.33	.35	.33
	Irrelevant Authority	.56	.54	.61	**.74**	.68	.68	.72	.68	.66
	Red Herring	.24	.30	.58	**.78**	.67	.67	.61	.65	.62
	Accuracy	.58	.58	.68	**.79**	.76	.75	.73	.73	.74
	Macro	.45	.48	.62	**.76**	.72	.72	.68	.69	.69
Covid	Causal Oversimplification	.45	.53	.40	.56	**.59**	.53	.53	.50	.50
	Cherry Picking	.35	.37	.37	.31	.36	.28	.34	**.38**	**.38**
	Evading Burden of Proof	0	0	.31	.45	.53	.46	**.57**	.49	.40
	False Analogy	.33	**.50**	.25	.29	.29	.25	.29	.29	.25
	Hasty Generalization	.17	0	.11	.17	.16	.11	**.25**	.10	.11
	Irrelevant Authority	0	0	0	0	0	0	0	0	0
	Red Herring	0	0	0	0	0	0	0	0	0
	Strawman	0	0	0	**.20**	0	0	.17	.17	.15
	Vagueness	.09	0	.09	.27	**.33**	.22	.30	.15	.19
	Accuracy	.23	.25	.26	.30	.34	.27	**.36**	.31	.30
	Macro	.16	.15	.17	.25	.25	.20	**.27**	.23	.22

Table 2: F1 scores on the Argotario, Logic and Covid datasets. **N**: no additional context provided. **C**: context of previous and next sentence provided where available (Propaganda and Climate only and added with no-context training sets of the three datasets shown).

in the training data between the N and C columns for the three datasets listed in Table 2 i.e. ARGOTARIO, LOGIC, and COVID. The only difference is that they are combined in the multitask training model with two other datasets (PROPAGANDA and CLIMATE) where the context is provided. Since the training is done on all datasets combined with some overlap between fallacy types across datasets, we report the results on the ARGOTARIO, LOGIC, and COVID datasets for context-based experiments on all five datasets.

With minor exceptions (e.g. *Doubt* in PROPAGANDA, *Vagueness* and overall scores in COVID), adding context does not improve the results for fallacy recognition. This could be related to the fact that some fallacy types require different context than others. For example, *Cherry Picking* requires understanding of the trend and *Strawman* requires the retrieval of the original argument, while *Evading the Burden of Proof* needs information regard-

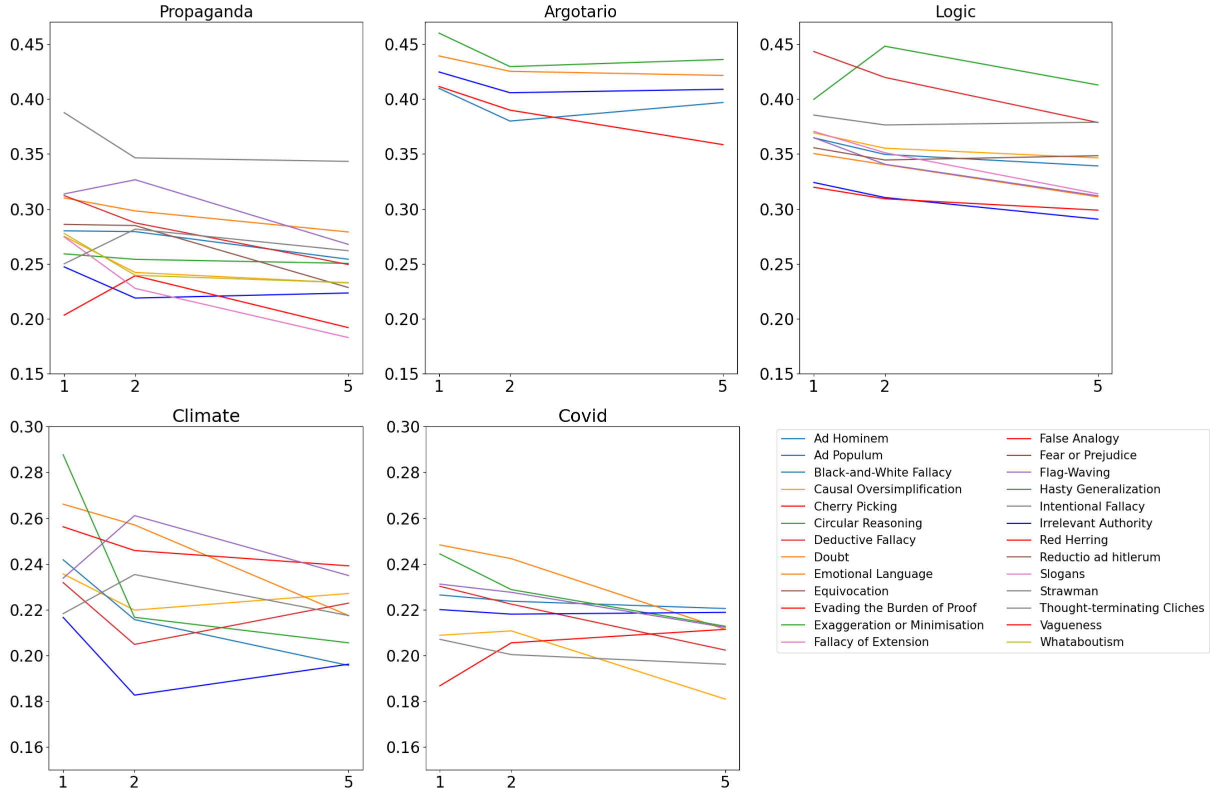

Figure 3: Average BLEURT score (y-axis) between original and synthetic data for each fallacy type in few-shot prompts (x-axis: 1-shot, 2shot, 5-shot).

ing the structure of the argument to assess its validity (Goffredo et al., 2022; Alhindi et al., 2021). Therefore, a unified form of context across 28 fallacy types does not have consistent improvement over experiments conducted under similar conditions.

Overall Observations If we examine all per-class results, we notice some inconsistency of the results under similar training settings. However, there are two general observations that are consistent across all results. First, data augmentation through large language models helps train smaller models on more data that is beneficial to fallacy recognition. Second, simple context of previous or next sentence does not provide valuable insight for this task and thus customization of the type of context based on the fallacy type is needed. In the next section, we take a closer look at the generated examples and how similar they are to the original ones.

6 Original and Synthetic Data Similarity

In order to understand the reason for 1-shot prompts to generate synthetic data that is more beneficial to the task, we analyze the similarity between the generated data and the original training examples shown at the prompts. We use BLEURT as our metric to calculate the similarity as it has the most consistent results with human evaluation (Sellam et al., 2020).

6.1 Similarity with the Training Sets

We calculate the BLEURT score for each original-synthetic example pair where the original example is the one included in the prompt to generate the synthetic example. Thus, we only run this analysis on the 1-shot, 2-shot, and 5-shot prompts. For the 2-shot and 5-shot, we report the average score for a generated example with respect to all original examples included in the prompt that generated it. Figure 3 shows average BLEURT scores for each fallacy type in all five datasets. We notice high similarity scores in ARGOTARTIO and LOGIC that range between 0.45 and 0.30 and much lower scores for CLIMATE and COVID that range between 0.18 and 0.30. This shows that it is harder in general to produce examples that are similar to those that naturally appear in misinformation found in news and social media in comparison to ones from dialogue in game settings or educational fallacy websites. However, what is shared across all datasets is that 1-shot prompts tend to be produce more similar examples to the ones included in the prompts.

On the other hand using 2-shot prompts generates examples that are both less similar than 1-shots prompts as well as 5-shot prompts. The regain in similarity in 5-shot prompts could be due to that including more examples improves the average similarity score for each generated example compared with all original examples in the prompt. In other words, each synthetic example generated by a 2-shot prompt is compared with the two examples in the prompt. Thus is could be very similar to one and very different from the other, which will keep the average similarity score at a lower value. However, each synthetic example from 5-shot prompts is generated after seeing five examples from the original data for the given fallacy and therefore might not be too similar to a single one of them but rather comparably similar to all five, which will result in a higher average similarity.

The drop in similarity scores is more significant in some fallacies than others such as *Hasty Generalization* in CLIMATE, and *Causal Oversimplification* in both CLIMATE and COVID. Some fallacy types break the general pattern of having the highest score in 1-shot prompts and the lowest in 2-shot prompts such as *Circular Reasoning* in LOGIC, which tends to have more homogeneous examples in the original training set and thus average similarity score increase with the number of shots in the prompt.

7 Related Work

With the significant focus on the development of generative large language models (LLM)s in recent years (Brown et al., 2020; Chowdhery et al., 2022; Zhang et al., 2022; Touvron et al., 2023), there has been an increase in the utilization of these models to annotate data (Feng et al., 2021; Chen et al., 2023; He et al., 2023; Bansal and Sharma, 2023; Zhang et al., 2023), or generate additional data instances that can be added to existing training sets for various tasks (Kumar et al., 2020; Schick and Schütze, 2021; Wang et al., 2023, 2021; Ye et al., 2022; Gao et al., 2022; Sahu et al., 2023).

We build on the line of work that uses language models for generating additional training data. Our work differs form previous work in the following aspects: i) we particularly focus on the ability of using synthetic data generated by language models to address data imbalance challenges, ii) we use zero-shot and few-shot settings to generate synthetic data but use full-shot training on a mix of original and synthetic data for the downstream task, and iii) we tackle a challenging task of fallacy recognition to understand the gains from using large language models for data augmentation.

8 Conclusion and Future Work

Fallacy recognition remains a challenging problem due to the high number of classes, severe data imbalance and the need in some cases for external information to the fallacious segment. To mitigate the effect of data imbalance, we studied the capabilities of large language models to generate synthetic data that can be used to train smaller models on a combination of original and synthetic data for fallacy recognition across multiple tasks. The main observation is that data augmentation through large language models is beneficial for this task. However, the conditions under which the data is generated impacts the quality of the synthetic data significantly. Providing one example in the prompt (1-shot) for a certain fallacy from the original data and asking GPT3.5 to generate a similar example results in synthetic data that is more similar to the original data in comparison to no examples (zero-shot) or more examples (e.g., 2-shot or 5-shot), and benefits downstream models in detecting fallacy types in test sets from the same distribution. The value in having synthetic data that is less similar to the original training data and possibly more generic to the task needs to be tested on data from unseen fallacy schemes or domains, which presents a potential avenue for future work. Overall, large language models show great potential to generate additional training data for the task of fallacy recognition, which can be used to train smaller size open-source models for this task.

In future work, we want to test the resilience of data augmentation on out-of-domain test sets such as fallacy in political debates. Also, we want study the ability of LLMs to generate examples that could be labeled by multiple fallacies and train machine learning for this tasks with multi-labeling. Finally, we want to experiment with ability of LLMs to provide more useful context for fallacy recognition.

Limitations

This work addresses challenges related to datasets with imbalance class ratios in high multi-class classifications using data augmentation generated by large language models. However, this work does not address other challenges in fallacy recognition. These include incorporating external knowledge to the fallacious segment which is essential is detecting some diversion fallacies such as *Cherry Picking* and *Strawman*. In addition, this work assumes a single fallacy label for each segment of text. However, in reality fallacies can overlap and thus handling the multi-label aspect of this task is not covered in this work. Finally, labeling fallacy by humans is inherently subjective and thus concurrent work suggests incorporating subjectivity in fallacy labels (Helwe et al., 2023), and thus treating human annotations as certain gold labels might provide a limited prospective for fallacy recognition models.

Ethics and Broader Impact

Using large language models to generate fallacy examples comes at a risk of having improper or hateful language. We have inspected a sample of the synthetic data and modified the prompts to minimize these aspects in the generated data. However, it is hard to guarantee the nonexistence of harsh language in data from large language models at scale. Some fallacy techniques in the datasets used in this paper have harsh or impolite language by definition e.g., *Name Calling*. Studying fallacy and training machine learning models for fallacy recognition could potentially lead to the promotion of the topic and the misuse of these models. While we acknowledge the risks, we believe this study contributes to increasing the awareness of fallacious techniques for both readers and writers and better equip them with proper tools to increase their immunity against potential harms.

References

Tariq Alhindi, Tuhin Chakrabarty, Elena Musi, and Smaranda Muresan. 2022. Multitask instruction-based prompting for fallacy recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8172–8187, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tariq Alhindi, Brennan McManus, and Smaranda Muresan. 2021. What to fact-check: Guiding check-worthy information detection in news articles through argumentative discourse structure. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 380–391, Singapore and Online. Association for Computational Linguistics.

Leila Amgoud and Philippe Besnard. 2013. Logical limits of abstract argumentation frameworks. *Journal of Applied Non-Classical Logics*, 23(3):229–267.

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

Parikshit Bansal and Amit Sharma. 2023. Large language models as annotators: Enhancing generalization of nlp models at minimal cost. *arXiv preprint arXiv:2306.15766*.

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.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. PLACES: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language model as an annotator: Exploring DialoGPT for dialogue summarization. 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 1479–1491, Online. Association for Computational Linguistics.

Jiahui Gao, Renjie Pi, LIN Yong, Hang Xu, Jiacheng Ye, Zhiyong Wu, WEIZHONG ZHANG, Xiaodan

Liang, Zhenguo Li, and Lingpeng Kong. 2022. Self-guided noise-free data generation for efficient zero-shot learning. In *The Eleventh International Conference on Learning Representations*.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4143–4149.

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

Charles Leonard Hamblin. 2022. *Fallacies*. Advanced Reasoning Forum.

Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.

Chadi Helwe, Tom Calamai, Pierre-Henri Paris, Chloé Clavel, and Fabian Suchanek. 2023. Mafalda: A benchmark and comprehensive study of fallacy detection and classification. *arXiv preprint arXiv:2311.09761*.

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

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

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*.

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022, Toronto, Canada. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Gaurav Sahu, Olga Vechtomova, Dzmitry Bahdanau, and Issam H Laradji. 2023. Promptmix: A class boundary augmentation method for large language model distillation. *arXiv preprint arXiv:2310.14192*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations*.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Douglas N Walton. 1996. *Argument structure: A pragmatic theory*. University of Toronto Press Toronto.

Jason Wang, Luis Perez, et al. 2017. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, 11(2017):1–8.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards zero-label language learning. *arXiv preprint arXiv:2109.09193*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. Llmaaa: Making large language models as active annotators. *arXiv preprint arXiv:2310.19596*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.