Illuminating Logical Fallacies with the CAMPFIRE Corpus

Austin Blodgett[1], **Claire Bonial**[1], **Taylor Pellegrin**[2], **Melissa Torgbi**[3], **Harish Tayyar Madabushi**[3]
[1] U.S. Army Research Laboratory, [2] Oak Ridge Associated Universities,
[3] University of Bath
austin.blodgett.civ@army.mil, claire.n.bonial.civ@army.mil,
thudson@terpmail.umd.edu, mat66@bath.ac.uk, htm43@bath.ac.uk

Abstract

Misinformation detection remains today a challenging task for both annotators and computer systems. While there are many known markers of misinformation—e.g., logical fallacies, propaganda techniques, and improper use of sources—labeling these markers in practice has been shown to produce low agreement as it requires annotators to make several subjective judgments and rely on their own knowledge, external to the text, which may vary between annotators. In this work, we address these challenges with a collection of linguistically-inspired litmus tests. We annotate a schema of 25 logical fallacies, each of which is defined with rigorous tests applied during annotation. Our annotation methodology results in a comparatively high IAA on this task: Cohen's κ in the range .69-.86. We release a corpus of 12 documents from various domains annotated with fallacy labels. Additionally, we experiment with a large language model baseline showing that the largest, most advanced models struggle on this challenging task, achieving an F1-score with our gold standard of .08 when excluding non-fallacious examples, compared to human performance of .59-.73. However, we find that prompting methodologies requiring the model to work through our litmus tests improves performance. Our work contributes a robust fallacy annotation schema and annotated corpus, which advance capabilities in this critical research area.

1 Introduction

Identifying and addressing misinformation remains a challenging, labor-intensive task today. Particularly in situations that are fast-changing—such as natural or infrastructural disasters, disease outbreaks, military conflicts, and political crises—the spread of misinformation can easily outpace the available resources and human capital needed to address it. Automatic and human-in-the-loop strategies show some potential to reduce the cost of labor for identifying misinformation, but there remain challenges to algorithmically and robustly identifying misinformation in arbitrary text. We envision reliable tools that can facilitate the automatic markup of text with likely misinformation markers (see Figure 1).

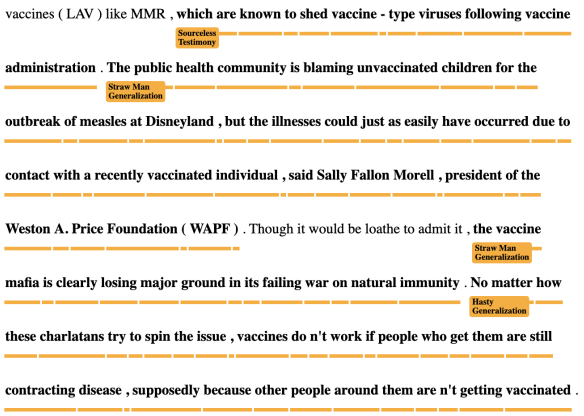

Figure 1: We show a visualization of fallacies identified in text. Although these are manual annotations shown, our corpus supports automatic markup of documents producing such a visualization for readers requiring automatic assessment of the credibility of a document, particularly in topic areas where fact-checking is not readily available.

To address these challenges, we developed the CAMPFIRE (Combined Annotations of Misinformation, Propaganda, and Fallacies Identified Robustly and Explainably) corpus—a corpus of texts on various topics (COVID-19, the Russian invasion of Ukraine, and the 2023 Ohio train derailment) annotated with markers useful for identifying misinformation. Although we divide these markers into testable and untestable beliefs, fallacies, and propaganda types, in this paper we narrow our focus to logical fallacy annotation. One advantage of focusing on logical fallacies as opposed to fact verification is that they allow us to scrutinize the soundness of a text's arguments in a content-neutral

way, even if many of the facts involved are not yet known. We address weaknesses of previous annotation schemas for fallacies by developing rigorous linguistic tests—inspired by the notion of frames and frame elements (Fillmore and Baker, 2001)—for each annotation label so that they can be applied consistently and objectively across domains. We find that our annotation methodology reduces the subjectivity of fallacy annotation, resulting in relatively high inter-annotator agreement (IAA): our agreement on a triple-annotated dataset, as measured by Cohen's κ, is in the range .69-.86 based on pairwise comparison of three annotators.

Technologies for identifying and addressing misinformation are particularly relevant today, given the popularity of generative, large language models (LLMs), the reliance of LLMs on online text, and the tendency of these systems to hallucinate. To establish baseline system performance on fallacy identification and recognition, we experiment with two of the largest, most advanced models (GPT-4o, GPT-o1) to predict CAMPFIRE fallacy labels. Performance leaves much to be desired: GPT-o1 achieves the best F1-score of .08 when excluding non-fallacious examples. Although this demonstrates the continued challenge of this task, we find that providing the litmus tests used by our annotators improves model performance.

After describing related work (Section 2), we present our theoretical framework, based upon first identifying the relevant, valid reasoning types (Section 3), followed by our annotation schema, including litmus tests ensuring diagnostic criteria for certain fallacy labels (Section 4). We then describe our corpus and annotation procedures, concluding with resulting IAA measures demonstrating the clarity and robustness of our schema (Section 5). We conduct experiments to establish baseline LLM performance in recognizing fallacies across three evaluation documents (Section 6).[1] Our discussion compares the challenges of human and system performance on this task, and we propose that our litmus tests reduce subjectivity in this task (Section 7). We conclude with suggestions for further system improvement on the critical task of fallacy and misinformation detection (Section 8).

2 Related Work

There has been an surge of research in NLP on detecting misinformation and related tasks, including fake news detection and automatic fact-checking, stance and sentiment analysis, and rumor detection, resulting in various workshops and shared tasks. Thus, there are a variety of annotation schemas and datasets focused broadly on the detection and analysis of misinformation, which may have some overlapping categories with our research. These datasets include the SemEval 2020 annotated dataset (Da San Martino et al., 2020a), and the credibility indicators outlined by Zhang et al. (2018). Here, we survey related work supporting the areas of fact-checking, propaganda techniques, and fallacy detection.

Both fact-checking generally and fake news detection more specifically require comparing claims against some ground truth, widely accepted facts. Hu et al. (2021) focus on fake news detection that compares claims against knowledge graphs. Instead of focusing on a document-level classification of fake news, Fung et al. (2021) cross-check individual elements of the document that better captures fake news where only a small portion of the document has been manipulated. One distinction between CAMPFIRE and fake news detection research is our focus on misinformation markers that do not require outside knowledge or ground truth facts to compare against. Our focus facilitates misinformation detection in subject-matter domains that are fast-changing, where the facts of a situation are not yet known or understood, such as the early weeks of the COVID-19 pandemic.

Propaganda techniques facilitate the acceptance and spread of certain claims, often in lieu of credible evidence and argumentation. Da San Martino et al. (2020b) offer a survey of relevant work on propaganda detection. Da San Martino et al. (2019) developed a corpus annotated with 18 labels describing propaganda techniques in which the annotators chose both the label and the span of the annotation, obtaining a γ inter-annotator agreement of .53. Recently, LLMs have been leveraged for propaganda detection. Sprenkamp et al. (2023) leverage GPT-3 and GPT-4 for classifying the propaganda techniques in the SemEval 2020 Task 11 dataset.[2] The best GPT-4 performance achieves an

[1] Our corpus and full experimental results and prompts can be found here: https://github.com/melissatorgbi/CAMPFIRE

[2] Many of the categories in this dataset overlap with CAMPFIRE propaganda techniques (e.g., APPEAL TO FEAR, FLAG-WAVING, REPETITION, SLOGAN), but several are classed as

F1-score of 58%, while the state-of-the-art system, which uses a fine-tuned RoBERTa model, achieves an F1-score of 63% (Abdullah et al., 2022). This demonstrates that the mere increase in scale of an LLM does not guarantee superior performance on this challenging task. Furthermore, the performance across the detection of particular techniques and fallacies varies wildly— LOADED LANGUAGE (F1-score of 72%) and NAME CALLING (F1-score of 65%) set the upper bound, while REPETITION (22% F1-score), BANDWAGON, and REDUCTIO AD HITLERUM (24% F1-score) sit on the lower bound. From this, we hypothesize that techniques with a clearer linguistic signature (as we would expect from LOADED LANGUAGE and NAME CALLING) are much easier to detect.

Like propaganda techniques, logical fallacies make a claim that may appear persuasive but is not supported by credible evidence or a logically sound argument. The *Argotario* corpus (Habernal et al., 2017, 2018) is one of the few corpora focused exclusively on logical fallacies, but their research crowd-sources annotations of just five logical fallacies. Bonial et al. (2022) attempt to replicate the *Argotario* annotation with expert annotators annotating logical fallacies in various publications, and show that the categories do not facilitate good IAA, nor can the distinctions be replicated by a system in a few-shot learning setting.

In Sahai et al. (2021), potential fallacies are collected automatically from Reddit by searching for mentions of fallacies in comments, and then these are filtered through crowdsourced judgments. Here again, IAA is somewhat low, particularly for HASTY GENERALIZATION, where agreement was measured via Cohen's κ at .38. This underscores the challenge of this annotation task. The authors explore several models for automatic prediction of fallacies, including BERT and MGN, with resulting F1-scores between 13 and 42% on the task most comparable to ours of labeling a comment with a particular fallacy. Unsurprisingly, given the correspondingly low IAA, the lowest F1-score is for HASTY GENERALIZATION.

We apply several lessons learned from related work. First, our schema supplies rigorous and detailed litmus tests facilitating objective determination of each annotation category. Second, the CAMPFIRE schema is refined until achieving satisfactory IAA, as the systems trained on data marked up with categories with relatively low IAA demonstrate correspondingly poor performance on those categories. Third, CAMPFIRE annotations focus on misinformation markers that can be identified from linguistic or structural features of a text, rather than external knowledge, as this reduces ambiguity in the annotation process and makes our schema more applicable in fast-changing domains where the facts are not yet known.

CAMPFIRE fallacies (e.g., BAND WAGON and REDUCTIO AD HITLERUM).

3 Theoretical Framework

A fallacy is an error in reasoning, argument, or methodology that leads to an unsound inference. A fallacy may be intentional or unintentional. Because fallacies are erroneous forms of inference, it is useful to categorize fallacies based on the type of inference they attempt to make. CAMPFIRE's fallacy taxonomy groups fallacies based on five inference types:

- **Deductive** inference draws a conclusion as a logical consequence of a premise. This includes inference using logical connectives *and*, *not*, *if*... *then*, etc., propositions that are true by definition (e.g., *cats are mammals*), as well as mathematical proof. A deductive fallacy can involve use of contradictions, skipping steps in an inference, or presenting an intuition, association, or bias as a universal principle. Deductive fallacies in CAMPFIRE include: FALSE DILEMMA, APPEAL TO NATURE, APPEAL TO NOVELTY, APPEAL TO TRADITION, THOUGHT-TERMINATING CLICHE.

- **Inductive** inference draws a conclusion that *likely* follows from a premise. For example, inductive inference might use observations about a population to infer a general claim that is supported by the observations. An inductive fallacy can involve relying on insufficient observations or relying on a biased sample of observations that are not representative of the population the general principle is meant to describe. Inductive fallacies in CAMPFIRE include: HASTY GENERALIZATION, CORRELATION-CAUSATION FALLACY, SLIPPERY SLOPE.

- **Abductive** inference draws a hypothesis that is meant to explain a set of observations, but is not observed directly. Note that in abductive reasoning, unlike inductive reasoning, the

hypothesis is only *consistent with* the observations and functions as a guess of how to explain them. Thus abductive inferences still need to be tested inductively before being considered credible. An abductive fallacy involves concluding that a hypothesis is true because it is consistent with observations without providing evidence for it. Abductive fallacies in CAMPFIRE include: APPEAL TO IGNORANCE, CONSPIRACY THEORY, SCAPEGOAT.

- **Testimony** is the process of obtaining information from a source. As an inference type, testimony can be thought of having the premises *source A says X* and *source A is credible and qualified* and the conclusion *X is true*. A testimony fallacy can involve relying on an uncredible or unqualified source, relying on testimony without identifying the source, or using the commonality of a belief as evidence that it is true. Testimonial fallacies in CAMPFIRE include: BANDWAGON, IRRELEVANT AUTHORITY, SOURCELESS TESTIMONY, AMBIGUOUS SOURCE, APPEAL TO CONFIDENCE/DISBELIEF, PLAIN FOLKS.

- **Rebuttal** is the process of critique of an argument in order to invalidate it. Rebuttal might involve identifying contradictions or inconsistencies in an argument (rebuttal of *deduction*), presenting counter-evidence or scrutinizing the reliability of evidence (rebuttal of *induction*), posing a more plausible hypothesis (rebuttal of *abduction*), or scrutinizing the credentials and credibility of sources of testimony (rebuttal of *testimony*). Rebuttal fallacies often involve rejecting evidence, arguments, or testimony for irrelevant or frivolous reasons. Rebuttal fallacies in CAMPFIRE include: APPEAL TO ACCIDENT, APPEAL TO FABRICATION, APPEAL TO COVER-UP, REJECTION BY AD HOMINEM, GUILT BY ASSOCIATION, GUILT BY ANALOGY, STRAW MAN GENERALIZATION, TWO WRONGS MAKE A RIGHT.

Fallacies are grouped into the five categories above based on inference type—deductive, inductive, abductive, testimony, or rebuttal. Each fallacy is assumed to be an unsound attempt to draw some inference, and different types of fallacies are organized by the type of inference they attempt to draw. Organizing the taxonomy this way also allows us to explain why techniques in each category are fallacious, because we can compare them to credible forms of inference and identify the differences.

4 Annotation Schema

We recognize three major challenging sources of ambiguity in the annotation of fallacies:

- In what circumstances should a given fallacy apply—how similar must the text be to the fallacy schema?
- What span of text should a fallacy be 'anchored' to—what span should receive the fallacy label?
- How much external knowledge should annotators rely on when annotating?

These challenges inform the design of our annotation schema. We address them using a collection of strategies meant to reduce the annotators' burden to make subjective judgments.

Annotating clauses. The annotation anchor of each CAMPFIRE fallacy label is always a *clause*. Each clause is a span of tokens within a sentence. We use a preprocessing script to first identify clauses in a text before annoatating. This script parses text into universal dependency trees (de Marneffe et al., 2021). Dependencies that correspond to a clause (*root, csubj, csubj:pass, ccomp, advcl, advcl:relcl, acl, acl:relcl, xcomp, parataxis*) are used to select the token span under that subtree. We also include coordinated clauses (under *conj*) and—for the sake of identifying testimonial fallacies—prepositional phrases evoking a reporting events (e.g., 'according to . . . ') are also treated as "clauses" for purposes of annotation. This procedure produces a (possibly nested) list of text spans each with the potential to be an annotation anchor. This allows for more fine-grained annotation than annotation by sentence, but involves less subjectivity than asking annotators to choose an arbitrary span by hand.[3] Because some fallacies can conceivably span over many clauses or sentences, each fallacy guideline also includes rules for identifying its conventional annotation anchor in order to further reduce this source of ambiguity.

Fallacy Guidelines. In practice, identifying fallacies can be a very challenging task because ar-

[3] See Furman et al. (2023) for discussion of span disagreement that motivated our decision to simplify the annotation span by using the clause as an anchor, and thereby reduce this source of disagreement.

guments in the real world that invoke a fallacy do not all take the same structural form or rely on the same lexical items or linguistic markers. Additionally, a real-world argument might have degrees of similarity to a known fallacy, in which case annotators might disagree about how similar it must be in order to deserve a fallacy label. To address this challenge, we develop rigorous annotation guidelines for each fallacy in our schema to drastically reduce this source of ambiguity. We start by observing that each fallacy has a logical form with premises and a conclusion. Each fallacy also has 'frame elements,' concepts evoked by the fallacy that must be in a particular relationship with each other for the fallacy label to apply.

Figure 2, for example, shows the guidelines for the SLIPPERY SLOPE fallacy. Text that is labeled as SLIPPERY SLOPE must evoke frame elements: Person/group **A** who initiates the events and Events **E** and **E'** which are the starting and resulting events of the slippery slope. The advantage of relying on frame elements and other litmus tests is that annotators are asked whether they can identify concepts in the text corresponding to the correct frame elements and whether these elements meet particular criteria, greatly reducing the subjectivity of the task.

During annotation, annotators consider a fallacy's logical form, frame elements, and tests to decide if that fallacy label can be applied. During adjudication, annotators again consult the guidelines to resolve disputes. Although frame elements are not annotated explicitly, they provide a rigorous litmus test to identify fallacies as objectively as possible.

Limiting External Knowledge. Another major challenge in the design of this schema was the issue of reliance on external knowledge. Early group annotations of fallacies revealed that often correctly identifying a fallacy in some text depended greatly on annotators' knowledge about the particular subject being discussed. Annotators with different levels of expertise or different preconceptions tended to make different judgments, resulting in lower agreement. We decided early on to reduce this source of ambiguity by focusing on fallacies that could be identified without relying on external knowledge or relying on it as little as possible. For example, an early version of our schema included the label STRAW MAN which is a fallacy of relevance where an opponent's position is mischaracterized in order to make it seem weaker than

logical form

A allows/causes event **E** therefore A will allow/cause event **E**′.

[*If we allow pet cats*]$_{premise}$, [*it's just a matter of time until someone has a pet alligator.*]$_{conclusion}$

frame elements

- Person/Group **A**: Initiator of the events **E** and **E**′
- Event **E** : Starting event
- Event **E**′ : Resulting event
  - Test 1: **E** and **E**′ are intentional or presented as intentional.
  - Test 2: **E**′ is presented as a more extreme version of **E**.
  - Test 3: **E** is presented as an indirect cause of **E**′, i.e. if **E** does not occur, **E**′ is assumed not to occur.

Anchor: **E**′

Figure 2: For each fallacy, our guidelines present the logical form and an example illustrating it. Additionally, required frame elements and litmus tests for determining if those frame elements are present in a sentence are provided.

it is and therefore easier to critique. But identifying STRAW MAN fallacies places a burden on the annotator to know what the opponent's true position is. Since that level of external knowledge is not practical and may vary between annotators, we narrowed this fallacy to STRAW MAN GENERALIZATION which can be identified with little external knowledge. See the Table 4 in the Appendix for the full list of fallacies, definitions, and examples.

5 Corpus

In this section, we present the corpus of our research into the detection of misinformation across a diverse range of documents. The corpus in total comprises fourteen documents sourced from a variety of publications, including scholarly works, tabloids, and major news organizations. Our corpus distribution across topics is summarized in Table 1. These documents were selected to represent the multiple avenues for the dissemination of misinformation across the population as well as to cover opposing positions on a number of topics. The corpus we present here is a subset of what is planned for the CAMPFIRE corpus which we continue to develop. Additionally, we note again that while

Annotation Task	Topic
Triple Annotations	Covid (1) Ukrainian Conflict (1) Ohio Train Derailment (1)
Double Annotations	Covid (4) Ukrainian Conflict (2) Ohio Train Derailment (0)
Single Annotations	Covid (4) Ukrainian Conflict (1) Ohio Train Derailment (0)

Table 1: A summary of our corpus of fourteen documents focusing on three topics. Double and triple annotations are annotated by multiple annotators independently and then adjudicated together.

	Annotator Pair		
Cohen's κ	**A1-A2**	**A2-A3**	**A1-A3**
Overall	.78	.86	.69
- **Fallacy Y/N**	.77	.89	.72
- **Fallacy Label**	.61	.72	.47

Table 2: We break our IAA evaluation into three metrics: 1) The overall Cohen's κ which accounts for the judgment of whether a fallacy is present or not and the correct fallacy label. 2) Fallacy Y/N measures Cohen's κ IAA on whether a fallacy is present. 3) Fallacy Label evaluates Cohen's κ IAA for only examples where either the gold or predicted label is a fallacy. We show IAA scores for each pair of annotators.

our full corpus annotation includes the annotation layers of beliefs types and propaganda techniques, in the present paper we focus only on the Fallacy annotations.

The process of document selection began with the selection of a range of medical documents on the topic COVID-19 at the start of the pandemic. The topics of these papers spanned the safety in wearing masks, the effectiveness of herd immunity, vaccination safety, and long-term illnesses. As we've developed our misinformation guidelines, we've broadened our annotation work to include the international conflict of the Russo-Ukrainian War, and an ecological disaster, known as the Ohio train derailment.

5.1 Annotation Procedure

The annotation process itself was a multi-stage endeavor that involved a team of three native English-speaking annotators with undergraduate or graduate-level training in linguistics. The annotators were trained over the course of two weeks to identify and annotate misinformation markers. Each annotator worked independently to annotate the documents according to the provided guidelines. This initial round of solo annotation allowed them to individually develop their expertise in recognizing and marking instances of misinformation across the four layers. After the initial annotations were completed, the annotators convened to discuss their findings and collaboratively establish a Gold standard for a subset of documents that were double and triple annotated. IAA scores were also collected to establish which fallacy labels were fairly clear, and which required updates either to the guidelines or to the categorization itself.

5.2 Agreement Metrics

All three annotators independently annotated three documents (containing a total of 194 annotation targets) and then convened to develop agreed-upon, gold standard annotations. We leverage these to establish IAA and to use as our evaluation set in Section 6. Table 2 shows our agreement results. We measured agreement in several ways. First, we measured the overall Cohen's κ IAA for each pair of our three annotators with results ranging from .69-.86. Because most clauses do not contain a fallacy and annotators usually agree on whether a fallacy is present, this overall IAA score is skewed by the vast number of NONE labels. To account for this in our evaluation, we also measure IAA on the judgement of whether a fallacy is present or not (Fallacy Y/N in Table 2) with results ranging from .72-.89. Lastly, we evaluate IAA on fallacy labels excluding cases where both annotators agree that a fallacy is not present (Fallacy Label in Table 2) with results ranging from .47-.72. This was the most challenging of the three metrics.

Overall, our level of agreement exceeds reported scores for other comparable annotations schemas and demonstrates the clarity and reliability of our schema, despite having 25 annotation category labels in a challenging task.

Additionally, Figure 3 shows confusion matrices for human and GPT-o1 performance respectively against our gold labels. What can readily be seen from this figure is that, for humans, the largest source of confusion of labels is the decision of whether the text should be labeled with a fallacy or should be labeled NONE, whereas for our experiments with GPT-o1, both the decision of whether a fallacy is present and the decision of which fallacy to apply are large sources of confusion.

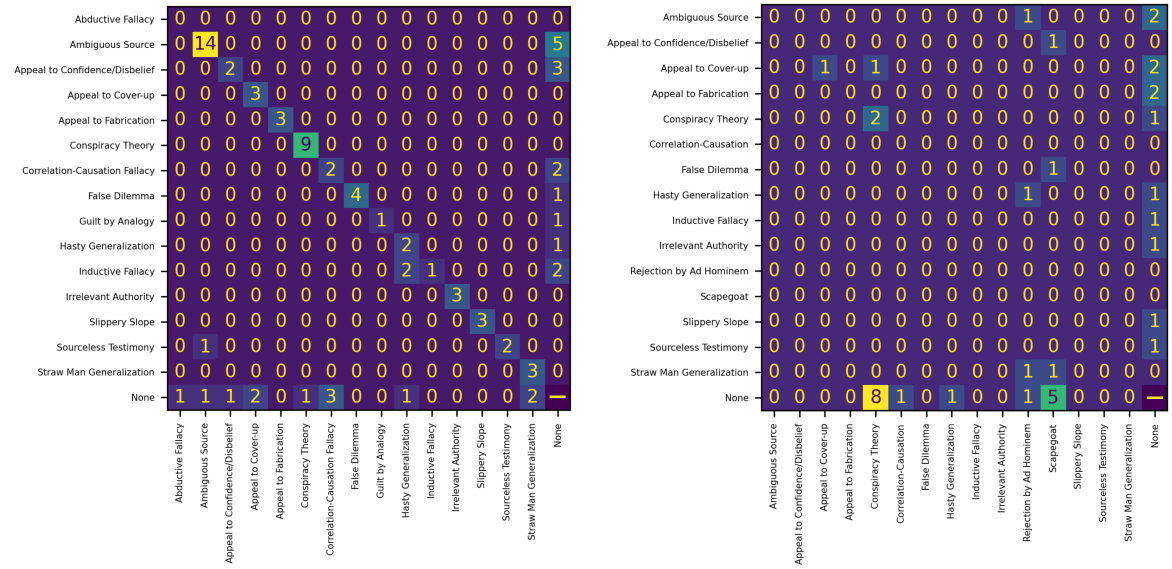

Figure 3: Confusion matrices for human performance (Left) and GPT-o1 performance (Right) respectively. The left matrix shows human annotations (columns) compared to gold adjudicated labels (rows) based on triple-annotated and double annotated documents. For comparison, the right matrix shows GPT-o1 predicted labels (columns) compared to gold (rows) based on triple-annotated documents. The dash in the lower right corner of each matrix stands in for the vast majority of NONE examples (1,104 examples for humans, 222 for GPT-o1) where both the gold and predicted labels agree that a fallacy is not present to prevent skewing the results.

6 Experiments: LLM Baseline

To establish baseline system performance on the task of recognizing and labeling fallacies, we use OpenAI's gpt-4o-2024-08-06 (GPT-4o) and o1-2024-12-17 (GPT-o1). These models were selected as representative of current LLM capabilities due to their large size. GPT-o1 was chosen alongside GPT-4o for its reported ability to handle complex reasoning which may be beneficial for this task. The temperature for GPT-4o and GPT-o1 were 0 and 1 respectively, which were the lowest options for each model to make the outputs more deterministic. Three documents that had been triple annotated and adjudicated were selected for evaluation, thereby giving us a clear picture of how LLM performance compares to manual annotation. A total of 22 tests were run, including experiments to investigate what information from the guidelines to include in the prompt.

6.1 Prompt Variations

Initial experiments were conducted to determine the amount and type of information to include in the prompt. These experiments were primarily tested on a single pilot document that contained the most fallacies of the three evaluation documents, and later extended to include the other two documents for final evaluation.[4] The prompt experiments involved varying combinations of the following elements, all drawn from the annotation guidelines:

- Fallacy Names
- 1-2 Sentence Fallacy Definitions
- Frame Element Listing
- Fallacy Examples

In one variation, we also instructed the model to output frame elements as instantiated by the annotation target sentence.

In the prompt, the model was given the whole document in text, and then a list of the clauses to label. We experimented with giving the model the full list of clauses in a single prompt, as well as iterating over each clause with a full list of fallacies and iterating over each clause and each fallacy, then asked the model to produce a label for a single clause and a single fallacy each time. The model was instructed to label each clause with a fallacy name or NONE which was then compared to a

[4]We acknowledge that leveraging items from our test set in our prompt experimentation could have led to over-optimization and better performance on those specific items. Ideally, we would conduct prompt experimentation on a separate set; however, our corpus size limited this possibility. Additionally, we note that the relatively poor performance overall indicates that optimizing on the test items did not dramatically skew performance.

gold label. The prompt variation that produced the highest F1-score on the pilot document was selected for further experiments.

Overall, our prompt experiments demonstrated that, in comparison to just providing the fallacy names, providing the fallacy definition improved performance, as does adding the frame element description and asking the model to output the frame elements in its response. Somewhat surprisingly, we found that adding examples of the fallacies did not improve performance. We tested two variants of this: first leveraging the simple, invented examples from the guidelines (see Table 4 in the Appendix for examples), and then adding corpus examples of the fallacies. Neither variation improved performance, and in fact the additional corpus examples decreased performance further. We posit that adding examples hurts performance because it cues the model into lexical similarities with examples, whereas the fallacies are based to a greater extent on semantic properties of the reasoning chain across clauses.

We found that providing a list of fallacies produced better results than iterating over individual fallacies. We also found that providing a listing of all clauses and asking the model to label all of them individually in one output response greatly improved performance over presenting the entire document and then asking the model to annotate a single clause at a time, iterating over clauses. We attribute this to the importance of the overall document context in understanding fallacies.

Thus, the best-performing prompt variation selected provided a task description, followed by a listing of all fallacies, each supplemented with its definition and a description of the required frame elements. The entire document was given in text, followed by the same text split into a listing of clauses. The model was then asked to output the fallacy label or "none" for each clause, and provide the instantiated frame elements for each detected fallacy.[5]

6.2 Results: Baseline Performance

Table 3 reports evaluation metrics for the two models tested using the best prompt variation. Similar to our IAA evaluation in section 5.2, we measure F1-scores in several ways. First, we measured the overall F1-score comparing annotators and models against our gold data. Because most clauses do not contain a fallacy and annotators usually agree on whether a fallacy is present, this overall F1-score is skewed by the vast number of NONE labels. To account for this in our evaluation, we also measure F1 on the judgement of whether a fallacy is present or not (Fallacy Y/N in Table 3). Lastly, we measure F1 on predicting fallacy labels excluding cases where both gold and predicted labels agree that a fallacy is not present (Fallacy Label in Table 2). This last metric presents the most challenging problem for both humans and LLMs.

We measure F1-scores among three annotators of .96-.98, but this score is greatly skewed by the presence of NONE labels. When drilling deeper, we find scores of .98-.99 on the judgement of whether a fallacy is present and .59-.73 on the more challenging task of predicting the correct label, excluding cases where both the annotated and gold labels agree that a fallacy is not present.

In comparison, when we calculate F1-scores for GPT-4o and -o1 against the gold standard, the models achieve .90 and .89 overall F1 respectively. Again, this is greatly skewed by the vast majority of NONE labels from non-fallacious sentences. When we inspect further, we find that models each achieve .95 scores when judging whether a fallacy is present. But on the more challenging metric of choosing the correct label excluding cases where both the predicted and gold labels agree that a fallacy is not present, GPT-4o and GPT-o1 score only .05 and .08 respectively, demonstrating that this task is far from solved.

When we drill down to examine how often the model can correctly predict that a fallacy is present and what the fallacy label is, we find that GPT-4o only correctly labels 1 of 14 gold fallacy labels from our evaluation set, while GPT-o1 correctly labels just 3. Qualitative analysis is provided in the Discussion.

7 Discussion

Our results show that our annotation schema and methodology— moving from a decision tree supporting recognition of a fallacy, to inference type, and finally to litmus tests involving frame elements to decide upon the specific fallacy—support relatively high overall annotator IAA on this challenging and generally subjective task. Additionally, our prompt variation experiments support the notion that having litmus tests for particular fallacies, in the form of required frame elements, also supports

[5]Full prompts can be viewed on our github: https://github.com/melissatorgbi/CAMPFIRE.

F1-score	GPT-4o	GPT-o1	Human
Overall	.90	.89	.96-.98
- Fallacy Y/N	.95	.95	.98-.99
- Fallacy Label	.05	.08	.59-.73

Table 3: Evaluation of two models against 3 linguist annotators. We break our evaluation into three metrics: 1) The overall F1-score which accounts for the judgment of whether a fallacy is present or not and the correct fallacy label. 2) Fallacy Y/N measures F1-score on whether a fallacy is present. 3) Fallacy Label evaluates F1-score for only examples where either the gold or predicted label was a fallacy.

model performance. When our annotation team disagreed upon the appropriate fallacy label, adjudication involved presenting the frame elements found in that sentence in support of a particular fallacy. Similarly, requiring the model to output the frame elements boosts performance. Thus, we posit that breaking the annotation task down in multiple steps and criteria for decision making decreases subjectivity in fallacy classification.

We readily acknowledge, however, that our analysis regarding model performance must be tempered by the fact that GPT-o1, the best-performing model, is only able to accurately label 3 of 14 gold-standard fallacies. Of the three fallacies that -o1 correctly identified, two are CONSPIRACY THEORY, an Abductive fallacy, and one is APPEAL TO COVER-UP, a Rebuttal fallacy. The three correctly identified cases are given below:

1. *The media...doesn't want you talking about East Palestine and Nordstream* - APPEAL TO COVER-UP
2. *A pandemic is their last attempt for total control* - CONSPIRACY THEORY
3. *A coordinated censorship attack is being waged against the entire independent media by Google, YouTube and Facebook* - CONSPIRACY THEORY

Example (3) above was the only fallacy correctly labeled by GPT-4o as well. We note that all three annotators agreed on these labels for each of these three cases.

When we explore several cases where the model posited that a fallacy existed where there was none, we find that GPT-o1 most often labeled clauses as CONSPIRACY THEORY fallacies: 8 of 17 predicted fallacies were assigned this label. Indeed, the model seems to have the best handle on the notion of a CONSPIRACY THEORY, as there was no clear set of lexical triggers associated with this set, and conceptually the false positives did involve the powerful, conspiratorial entity frame element, but no clear conspiratorial event required for annotation. Next most frequently, GPT-o1 assigned SCAPEGOAT fallacies where the word "blame" was mentioned in 7 of 17 predicted fallacies. Finally, AD HOMINEM was assigned in 4 cases where there were insulting names such as "charlatan." Thus, in many of these cases, while one frame element was found in the clause (often cued by a key lexical item), all required elements were not present.

8 Conclusion & Future Work

When we consider our manual and model annotation results overall, we posit that model performance could be brought closer to human performance with prompting strategies as well as structured output that required frame elements and litmus tests to be passed. Only if the model can provide all frame elements can the annotation of a particular fallacy be assigned. This process of requiring the model to "show its work" when it comes to the fallacy assigned is quite similar to how annotators argued for and settled disputes over fallacy labels.

In addition to exploring more sophisticated prompting strategies, we are currently working to further expand our corpus to levels adequate to experiment with finetuning a model. We are eager to see if a fine-tuned model can excel at this task, or if larger models with more advanced "reasoning" capabilities can outpace even fine-tuned models given the right prompting strategies.

With improved model performance over a larger corpus, we will also begin to explore if there is any difference in performance in detecting fallacies that are missteps in different reasoning types. It has been posited that LLMs are inductive, bottom-up reasoners moving from specific observations to generalizations (Olsson et al., 2022); thus, we may expect performance on inductive fallacies to be superior to deductive and abductive fallacies. However, we also note an opportunity to leverage fallacy recognition evaluation in order to further explore whether or not these models are "reasoning" at all (cf. Lu et al. (2024)).

Limitations

Although we annotated a schema of 25 fallacy types and demonstrated improvement of inter-annotator agreement over previous work, there is still much room for improvement in the types of fallacies to identify, the agreement and objectivity of annotators, and the reliability of automated systems in performing this task. So far, our annotations have focused on single-author texts. We hope to add annotations of multi-author debate and discourse in future work.

Ethical Considerations

All annotators who participated in this research were paid adequately for their work and were included as authors. Annotators met regularly to discuss ways to improve the annotation process and make it easier, and their expert input was relied on throughout the development of our schema. Misinformation detection is a complex issue with important societal implications, and we recognize the possibility for bias to influence our data creation. We take steps to reduce the possibility for bias wherever possible. We believe our approach of focusing on logical structures of arguments has allowed us to annotate in a content-neutral way and thus reduce potential sources of bias.

References

Malak Abdullah, Ola Altiti, and Rasha Obiedat. 2022. Detecting propaganda techniques in english news articles using pre-trained transformers. In *2022 13th International Conference on Information and Communication Systems (ICICS)*, pages 301–308. IEEE.

Claire Bonial, Austin Blodgett, Taylor Hudson, Stephanie Lukin, Jeffrey Micher, Douglas Summers-Stay, Peter Sutor, and Clare Voss. 2022. The search for agreement on logical fallacy annotation of an infodemic. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4430–4438.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020a. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020b. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4826–4832. International Joint Conferences on Artificial Intelligence Organization. Survey track.

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.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Charles J. Fillmore and Collin F. Baker. 2001. Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop*, Pittsburgh. NAACL, NAACL.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil. 2021. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. 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 1683–1698, Online. Association for Computational Linguistics.

Damián Ariel Furman, Pablo Torres, José A. Rodríguez, Laura Alonso Alemany, Diego Letzen, and Vanina Martínez. 2023. Which argumentative aspects of hate speech in social media can be reliably identified? In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 136–153, Nancy, France. Association for Computational Linguistics.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *EMNLP (System Demonstrations)*.

Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018. Adapting serious game for fallacious argumentation to german: pitfalls, insights, and best practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to the knowledge: Graph neural fake news detection with external knowledge. 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 754–763, Online. Association for Computational Linguistics.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2024. Are emergent abilities in large language models just in-context learning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5139.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

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

Kilian Sprenkamp, Daniel Gordon Jones, and Liudmila Zavolokina. 2023. Large language models for propaganda detection. *arXiv preprint arXiv:2310.06422*.

Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, and 1 others. 2018. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018*, pages 603–612.

A Fallacy Definitions and Examples

We provide a listing of all our fallacy labels, organized by fallacy type, as well as guidelines examples of each fallacy in Table 4.

Inference Type	Fallacy Label	Guidelines Example
Deductive	FALSE DILEMMA	If we don't get a cat then we have to get a dog.
	APPEAL TO NATURE / NOVELTY / TRADITION	Raw meat is more natural for cats / We have to get that new cat food / Old-fashioned cat food is the best.
	THOUGHT-TERMINATING CLICHE	It just is the way it is.
Inductive	HASTY GENERALIZATION	My cat is black, so all cats are black.
	CORRELATION-CAUSATION	Many cat owners have asthma.
	SLIPPERY SLOPE	If we allow pet cats, it's just a matter of time until someone has a pet alligator.
Abductive	APPEAL TO IGNORANCE	No one has proven that cats can't understand humans.
	CONSPIRACY THEORY	There is an evil, secret organization of people who want to kidnap our pet cats.
	SCAPEGOAT	The shortage of cat food is all because of immigrants.
Testimony	BANDWAGON	90% of people prefer cats.
	IRRELEVANT AUTHORITY	I heard from a friend that cats can sense radio waves.
	SOURCELESS TESTIMONY	It is known that cats can sense radio waves.
	AMBIGUOUS SOURCE	Scientists say that cats can sense radio waves.
	APPEAL TO CONFIDENCE-DISBELIEF	Cats couldn't possibly be a good pet.
	PLAIN FOLKS	You can trust me, I'm just an ordinary pet owner like you.
Rebuttal	APPEAL TO ACCIDENT / FABRICATION / COVER-UP	Some people say cats are mean, but those are just the bad cats / People who like cats are brainwashed by the pro-cat shadow government / The news never tells you about all the people who were murdered by their cats.
	REJECTION BY AD-HOMINEM	I don't trust the opinion of a cat person.
	GUILT BY ASSOCIATION / ANALOGY	John's brother stole a dog, so John can't be trusted! / Cat owners are like fascists, always creating rules for their pets.
	STRAW MAN GENERALIZATION	Dog lovers think that cats are evil!
	TWO WRONGS MAKE A RIGHT	People say cats can be mean, but what about dogs?!

Table 4: Listing of the fallacy labels used in our schema; these are categorized by the inference type involved, where each fallacy represents a fallacious step in that type of reasoning. We also provide a simple, invented example of the fallacy listed in our guidelines.