MISSCI: Reconstructing Fallacies in Misrepresented Science

Max Glockner[♣], **Yufang Hou**[♢♣], **Preslav Nakov**[♠] and **Iryna Gurevych**[♣]
[♣]Ubiquitous Knowledge Processing Lab (UKP Lab),
TU Darmstadt and Hessian Center for AI (hessian.AI)
[♢]IBM Research, Ireland, [♠]MBZUAI
www.ukp.tu-darmstadt.de

Abstract

Health-related misinformation on social networks can lead to poor decision-making and real-world dangers. Such misinformation often misrepresents scientific publications and cites them as "proof" to gain perceived credibility. To effectively counter such claims automatically, a system must explain how the claim was falsely derived from the cited publication. Current methods for automated fact-checking or fallacy detection neglect to assess the (mis)used evidence in relation to misinformation claims, which is required to detect the mismatch between them. To address this gap, we introduce MISSCI, a novel argumentation theoretical model for fallacious reasoning together with a new dataset for real-world misinformation detection that misrepresents biomedical publications. Unlike previous fallacy detection datasets, MISSCI (*i*) focuses on implicit fallacies between the relevant content of the cited publication and the inaccurate claim, and (*ii*) requires models to verbalize the fallacious reasoning in addition to classifying it. We present MISSCI as a dataset to test the critical reasoning abilities of large language models (LLMs), which are required to reconstruct real-world fallacious arguments, in a zero-shot setting. We evaluate two representative LLMs and the impact of providing different levels of detail about the fallacy classes to the LLMs via prompts. Our experiments and human evaluation show promising results for GPT 4, while also demonstrating the difficulty of this task.[1]

1 Introduction

False or misleading narratives spread rapidly on social media (Vosoughi et al., 2018; Wardle, 2018), posing challenges for non-experts in discerning credible information, and exceeding the capabilities of human fact-checkers (HFC). The need to support HFC has accelerated the research in automated fact-checking (AFC) and related tasks (Guo et al., 2022; Schlichtkrull et al., 2023a). Yet, the real-world applicability of AFC systems is limited due to the lack of trustworthiness (Nakov et al., 2021), or their reliance on counter-evidence, which may not exist (Glockner et al., 2022). Often, misinformation distorts genuine information rather than creating new content (Brennen et al., 2020). For example, the claim in Figure 1 that "*hydroxychloroquine is a cure for COVID-19*" contains a kernel of truth and relies on some content (referred to as *accurate premise)* of the cited study that found "*chloroquine reduced infection of the coronavirus.*" However, further content from the study shows that it did not conduct human experiments (s_1) and focused on SARS-CoV-1 (s_2), different from the virus causing COVID-19. Only when knowing this additional information (s_1 and s_2) about the cited study, one can detect the applied fallacies (*Fallacy of Composition* and *False Equivalence*).

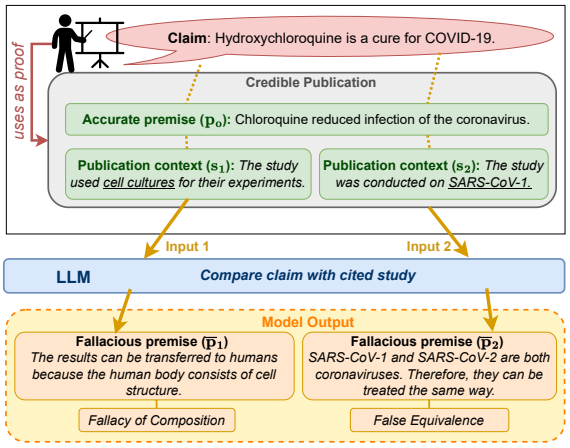

Figure 1: **Fallacious Argument Reconstruction:** The claim is falsely derived from the cited study (*green*) by relying on the content of p_0. The model generates and classifies the fallacious reasoning (*orange*) that needs to be applied when concluding the claim based on all relevant study content (including s_1 and s_2).

[1]Code and data are available at: https://github.com/UKPLab/acl2024-missci.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4372–4405
August 11-16, 2024 ©2024 Association for Computational Linguistics

In this work, we focus on inaccurate claims that misrepresent scientific publications. We assume that the misrepresented publication is presented alongside the claim as a "proof" amplifying the claim's impact through increased perceived credibility. With this assumption, we can access the cited sources, which is essential for detecting fallacious reasoning and implements human verification strategies (Silverman, 2014). Our goal is to automatically outline the fallacious reasoning and explain why the claim is incorrect, a crucial aspect of debunking misinformation (Schmid and Betsch, 2019; Lewandowsky et al., 2020).

Earlier work on fallacy detection focused on surface-level fallacies like *Ad Hominem* or *Loaded Language*. More recent work (Jin et al., 2022; Alhindi et al., 2022) also included logical fallacies like *False Cause* or *Hasty Generalization*. Yet, they did not adapt the task definition to account for the fact that these fallacies may need information beyond what is explicitly stated in the text. This hinders their applicability to real-world fallacies that rely on external information. To bridge this gap, we introduce MISSCI, a new argumentation theoretical model for fallacious reasoning, accompanied by a new fallacy detection dataset derived from real-world misinformation. Unlike prior work, we (*i*) treat inaccurate claims as the conclusion of a logical argument, encompassing all necessary information to detect the applied fallacies (*green* in Figure 1). We (*ii*) formulate a distinct fallacy inventory drawn from literature to express fallacies when misrepresenting scientific publications. Finally, inspired by Cook et al. (2018), we (*iii*) explicitly verbalize the fallacious reasoning via premises (*orange*) that only implicitly exist based on the claim to reconstruct the fallacious argument.

Our focus lies on the reasoning abilities to reconstruct fallacious arguments, and we manually paraphrase the relevant publication content (*green*) based on HFC articles, rather than using the original misrepresented document. Motivated by recent progress in zero-shot performance of large language models (LLMs) (Kojima et al., 2022), we evaluate the reasoning abilities of two state-of-the-art LLMs in reconstructing the fallacious reasoning on MISSCI and define the task in a zero-shot setup, exemplified in Figure 1. Given the claim, an accurate premise, and the publication contexts, the model must verbalize the fallacious reasoning and assign fallacy classes. Our contributions are:

- A new **argumentation theoretical model** with a new task formulation to reconstruct the fallacious arguments.
- A new **dataset** comprising complex fallacious arguments of real-world misinformation.
- **Evaluation** of two state-of-the-art LLMs and their critical reasoning abilities to reconstruct fallacious arguments.

2 Related Work

Previous work has focused on surface-level fallacies for propaganda detection (Da San Martino et al., 2019; Piskorski et al., 2023; Salman et al., 2023), for online discussions (Habernal et al., 2018a; Sahai et al., 2021), or for gamified settings (Habernal et al., 2017). The addressed fallacies typically do not require information beyond what is stated explicitly in the text. Other work targeted specific fallacies such as *Non Sequitur* in law (Nakpih and Santini, 2020) or *Ad Hominem* in social media (Habernal et al., 2018a). More similar to our work, some research (Jin et al., 2022; Musi et al., 2022; Alhindi et al., 2022) focused on logical fallacies from the real world. Yet, they neither verbalized the implicitly applied fallacies, nor considered the underlying sources beyond what is explicitly stated in the text. Moreover, our task design to generate fallacious premises differs from implicit premise detection work (Habernal et al., 2018b; Chakrabarty et al., 2021) in that the premises in MISSCI are inherently invalid and linked to applied fallacy classes (see $\overline{p}_1$ and $\overline{p}_2$ in Figure 1; *orange*). Existing work on fallacy generation focused on data generation (Huang et al., 2023; Alhindi et al., 2023) rather than on articulating applied fallacious reasoning within real-world fallacious arguments.

Scientific AFC work (Wadden et al., 2020; Sarrouti et al., 2021; Wadden et al., 2022; Lu et al., 2023; Vladika and Matthes, 2023) considered external sources, like us, but did not identify and articulate the nuanced fallacies when concluding a claim from the cited study. Detecting distortions in scientific communication is part of science communication research (Augenstein, 2021), where studies have examined exaggerations in news articles (Sumner et al., 2014; Bratton et al., 2019; Yu et al., 2020; Wright and Augenstein, 2021), analyzed reporting certainty in scientific publications (Pei and Jurgens, 2021), or quantified information mismatches between reported and actual scientific findings (Wright et al., 2022). In parallel work,

Wührl et al. (2024) quantify and compare the original paper with other reporting of their findings across fine-grained dimensions such as certainty or generalization. In addition to our distinct task setup, our problem space differs as we focus on harmful misinformation that comprises more severe distortions, which are not necessarily bound to a study's main findings.

3 Formalism of MISSCI

3.1 Misrepresented Science Arguments

Inspired by Cook et al. (2018), we reconstruct the fallacious reasoning in the form of a logical argument. For an accurate claim c, a logical argument would be structured as follows:

$$\{p_0\,,\,p_1\,,\,\ldots\,,\,p_N\} \Rightarrow c \qquad (1)$$

where $P = \{p_0, p_1, \ldots, p_N\}$ is a set of *accurate* premises that jointly entail ($\Rightarrow$) the true claim c. For inaccurate claims $\overline{c}$, the entailment relation does not hold based on accurate premises, formulated as $P \not\Rightarrow \overline{c}$, where $\not\Rightarrow$ denotes a corrupted entailment relation. To *reconstruct* a fallacious argument, a set of inaccurate (fallacious) premises $\overline{p}_i \in \overline{P}$ must be applied such that $P \cup \overline{P} \Rightarrow \overline{c}$. For example, consider the following argument (simplified from Figure 2):

- **Accurate premise p_0:** The viral COVID-19 spike protein inhibits repair of DNA damage.
- **Fallacious premise $\overline{p}_{1,2}$:** COVID-19 vaccine spike proteins are as dangerous as viral COVID-19 spike proteins.
- **Conclusion ($\overline{c}$):** Therefore, COVID-19 vaccines inhibit repair of DNA damage.

Here, the accurate premise alone lacks sufficient support for the conclusion (or claim). Establishing a support relationship between the accurate premise and the conclusion requires the fallacious premise, which employs the *False Equivalence* fallacy. To debunk a claim $\overline{c}$, the argument can be *deconstructed* by highlighting the fallacies applied in each $\overline{p}_i$. This invalidates the premises that are necessary to establish the claim as a logical conclusion and renders the argument invalid.

Misinformation that is falsely derived from a single credible publication S can be formulated as $S \not\Rightarrow \overline{c}$. Each argument in MISSCI has exactly one accurate premise, $P = \{p_0\}$, which is entailed by (parts of) S and represents the "kernel of truth" behind the inaccurate claim $\overline{c}$. Since all arguments in MISSCI constitute misinformation, each argument is illogical and contains at least one reasoning gap, which must be bridged via a fallacious premise $\overline{p}_i$ that applies a fallacy class f_i to support the claim. This reasoning gap only becomes imminent after observing relevant context (s_i) from the misrepresented publication S. For example, consider Figure 2, where a study observed harm from SARS-CoV-2 spike proteins (*accurate premise* p_0). The argument wrongly applied this finding to mRNA vaccines, assuming that viral and vaccine spike proteins behave identically (in both (alternative) fallacious premises). Without detailed content about the misrepresented publication (that it was observed on *viral* spike proteins in s_1), this fallacy is undetectable. In this work, we manually paraphrase this necessary context s_i that exhibits the reasoning gap from the publication S where each s_i is entailed by S. Because the accurate premise p_0 is always entailed by the publication, its publication context (s_0) is identical to the accurate premise. We represent each fallacious reasoning $R_i \in R$ that bridges one reasoning gap as a triplet $R_i = (s_i, \overline{p}_i, f_i)$. To bridge all reasoning gaps and fully support the claim $\overline{c}$ each fallacious reasoning $R_i \in R$ must be applied. The overall formulation of a fallacious argument $\mathcal{A}$ is shown below:

$$\left\{ \begin{array}{ccccccc} \overset{s_0}{=} & & \overset{s_1}{\downarrow} & & \overset{s_2}{\downarrow} & & \overset{s_N}{\downarrow} \\ p_0\,, & & \overline{p}_1\,, & & \overline{p}_2\,, & \ldots\,, & \overline{p}_N \\ & & \downarrow & & \downarrow & & \downarrow \\ & & f_1 & & f_2 & & f_N \end{array} \right\} \Rightarrow \overline{c} \qquad (2)$$

Each argument $\mathcal{A} = (\overline{c}, p_0, R)$ comprises an inaccurate claim $\overline{c}$, that builds on the accurate premise p_0 and applies at least one fallacious reasoning $R_i \in R$. As shown in Figure 2, in some cases, multiple fallacious premises with distinct fallacy

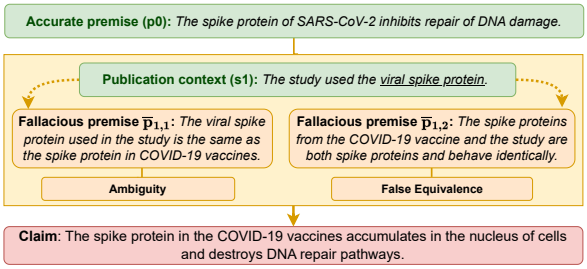

Figure 2: **Interchangeable Fallacies:** On the left, no distinction between the different "spike proteins" (from the vaccine or the virus) is made; on the right, both are assumed to behave alike. Only one of these premises is needed to bridge the reasoning gap.

classes can be used interchangeably (i.e., $\overline{p}_i = [\overline{p}_{i,1}, \overline{p}_{i,2}]$). Interchangeable fallacies always share the identical publication context (s_i), but only one of them is necessary to bridge the reasoning gap.

3.2 Non-exhaustive, Inductive Arguments

Unlike Cook et al. (2018), we do not require arguments to deduce the claim, which is unrealistic based on empirical evidence from scientific publications. Instead, we consider *strong inductive support* sufficient for ($\Rightarrow$). In inductive arguments, invalid premises, by definition, weaken the conclusion without necessarily rendering it false. To avoid labeling minor mismatches as fallacies, we ensure the relevance of each fallacious reasoning $R_i \in R$ in the strong inductive argument by deriving R exclusively from the HFC article. By relying on the HFC, the extracted fallacious reasoning lines are non-exhaustive and limited to the most pivotal ones. Importantly, fallacies within the logical arguments do not necessarily match the fallacies made by the claimant. For example, one can make the claim in Figure 1 after only skimming parts of the study without ever knowing that experiments were performed in cell cultures (s_1). In this case, the *Fallacy of Exclusion* rather than the *Fallacy of Composition* was applied. To account for these cases, we state our objective as detecting the necessary fallacies needed to conclude the claim $\overline{c}$ from all relevant content of the misrepresented publication S. This follows the *principle of total evidence*, which dictates that an inductive argument must consider all available relevant evidence (Chakraborti, 2007).

3.3 Task Definition

We assess the ability to reconstruct the fallacious reasoning for each fallacious argument $\mathcal{A}$ on the fallacious reasoning level: For each fallacious reasoning $R_i \in R$, given the claim $\overline{c}$, the accurate premise p_0 and the publication context s_i from R_i, the model must verbalise the fallacious premise $\hat{\overline{p}}_i$ and predict the applied fallacy class $\hat{f}_i$ to bridge the reasoning gap, so that ($\hat{\overline{p}}_i$, $\hat{f}_i$) constitute valid fallacies as approximated via the annotated interchangeable fallacies ($\overline{p}_i$, f_i) of R_i.

4 Dataset

Our main annotator was a M.Sc. student in biology, covering early pilot studies and post-annotation consolidation. Additionally, two more M.Sc. students, one in biology and one in linguistics, were employed during annotation. The annotators received a pay of 12.26 EUR per hour. We used Surge AI[2] as the annotation tool. Weekly meetings involving all annotators and one of the authors were held throughout the project to provide feedback and to refine the guidelines as needed, in line with the recommendations of Klie et al. (2024). To create MISSCI, we (1) *collected* HFC articles and pre-selected links that may point to a misrepresented publication, (2) manually *selected* all links that pointed to a misrepresented publication, and (3) *reconstructed* the fallacious arguments from the HFC articles. A summary of these three steps is given in Table 1. We collected a total of 527 fact-checking articles from HealthFeedback[3] until January 2023, excluding those that address accurate claims. HealthFeedback collaborates with scientists in reviewing health and medical claims. From these HFC articles, we annotated 8,695 links from reputable sources (cf. §A.1) to determine whether a link pointed to a misrepresented scientific publication. Our annotators found 208 links pointing to misrepresented scientific publications across 150 HFC articles (cf. §A.2; Krippendorff's α was 0.728).

	Collect	Select	Reconstruct
HFC articles	527	150	147
Links	8,695	208	184
Arguments	–	–	184
Fall. Reasoning R_i	–	–	435

Table 1: **Dataset Construction:** Number of elements for all three steps during dataset construction.

4.1 Fallacious Argument Reconstruction

The annotators were instructed to generate all elements of the fallacious argument $\mathcal{A}$ that falsely concludes the claim $\overline{c}$. This included the accurate premise p_0 as well as the fallacious premise $\overline{p}_i$, fallacy class f_i, and publication context s_i for each fallacious reasoning R_i (cf. §B.1 for a list of all fallacy classes). Each element had to be justified with an extracted statement from the HFC article. Often, selecting a single definitive fallacious reasoning was ambiguous and oxymoronic, akin to identifying the "correct invalid reasoning" (cf. §3; Figure 2). This aligns with Bonial et al. (2022),

[2] https://www.surgehq.ai/
[3] https://healthfeedback.org/

who observed that due to overlapping definitions, fallacies could often be reformulated to fit the definition of a different fallacy. Hence, we allowed separate listing of interchangeable fallacies, which were merged during consolidation. As we aimed to detect the fallacious reasoning *between* a scientific publication and an inaccurate claim, our work rests on the assumption that the publication itself is trustworthy. To verify the trustworthiness, the annotators rated the credibility of the scientific document based on the HFC article, which we analyzed in §C.1. Detailed instructions and the annotation process are outlined in §B.2.

4.2 Inter-Annotator Agreement

We collected 520 annotated HITs for 208 potential arguments. After consolidation (cf. §B.3), MISSCI contained 435 distinct fallacious reasoning lines (R_i) bridging different reasoning gaps (with a total of 550 interchangeable fallacies) for 184 fallacious arguments. Each argument involved 1-5 fallacious reasoning lines (R_i), averaging at 2.4 per argument. Most of the arguments within MISSCI were related to the COVID-19 infodemic. We show the distribution of arguments over years and their relation to COVID-19 in §C.2. Calculating the inter-annotator agreement for the fallacy class annotations faced two challenges: interchangeable fallacies with different but valid labels, and annotators identifying different (non-interchangeable) fallacious reasoning lines that bridge different reasoning gaps.

To address this, we used two complementary measures: We calculated the inter-annotator agreement for the fallacy class f_i among all 253 fallacious reasoning R_i identified by at least two annotators within the consolidated arguments. When simulating a single-label classification setup, the inter-annotator agreement, measured using Krippendorff's α, was 0.520. This is comparable to Cohen's κ of 0.47 (Alhindi et al., 2022) and 0.52 (Musi et al., 2022) in similar work.[4] We additionally compared each fallacious reasoning R_i identified by each individual annotator to the consolidated argument and measured how many fallacious reasoning R_i of the consolidated argument a single annotator found on average. Here, we considered all fallacies that were merged during consolidation as identical, and did not differentiate whether the annotators selected distinct interchangeable fallacies that apply different fallacy classes. Here, we only considered 70 arguments, which were fully annotated by all three annotators, for the computation to not artificially inflate the coverage by a single annotator. On average, each annotator identified 72.5% of the fallacious reasoning lines R_i in the consolidated argument. We examined how this affected the overall recall of the detected fallacies in MISSCI in §C.3.

[4]Our agreement measure differs as we employed up to three rather than two annotators.

4.3 Fallacy Class Analysis

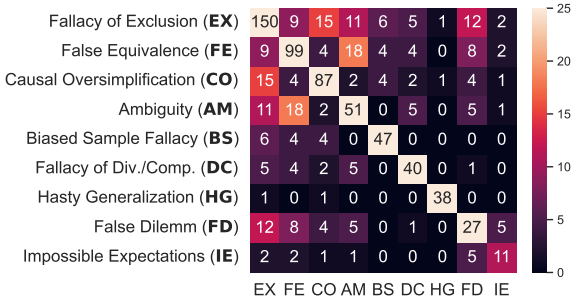

Figure 3: **Interchangeable Fallacy Classes:** Heatmap of co-occurring interchangeable fallacy classes of the consolidated arguments ordered by frequency.

To understand which fallacy classes have been annotated together as interchangeable fallacies, we show their co-occurrence matrix in Figure 3. The definitions and the examples for all fallacies are given in §B.1. We observe that most co-occurrences are between *False Equivalence* and *Ambiguity* (as discussed in Figure 2). In contrast, *Hasty Generalization* was the most clear-cut fallacy in our annotations, likely because HFC typically explicitly specify when a study lacks sufficient observations for the claim, and because it had little overlap with other fallacy definitions. The majority class of all fallacy class annotations is the *Fallacy of Exclusion*. This fallacy omits critical information when drawing a conclusion and could theoretically apply to every reasoning gap that depends on the information in the publication context s_i, because each s_i contains content that undermines the claim. To address this, during annotation the annotators were tasked to prioritize more specific fallacies before the *Fallacy of Exclusion* (cf. §B.2), yet the fallacy class remains the most common.

We found two main reasons for the prevalence of the *Fallacy of Exclusion* in MISSCI: The first and primary reason for including this fallacy in our inventory, is that parts of the misrepresented publication contradict the claim $\bar{c}$. For example, the claim

that "*spike proteins induced by RNA vaccines can damage blood vessels*" is based on a study which concludes that "*vaccination could protect against blood vessel damage*". In this case, the authors' comment must be ignored and no other fallacy in our inventory can be used to conclude a claim opposite to their conclusion. Second, the *Fallacy of Exclusion* often serves as a fallback class in MISSCI due to its broad applicability. Given the inevitability of an incomplete fallacy inventory, instances where the detected fallacies do not clearly align with the predefined fallacy classes are frequently labeled as *Fallacy of Exclusion*. This leads to co-occurrences with other fallacies in borderline cases. For example, the claim that "*Pfizer's COVID-19 vaccine effectiveness dropped from 100% to 20%*" relies on infection numbers, and ignores the reported high effectiveness against severe disease. This flawed reasoning could be interpreted as *False Equivalence*, assuming mild and severe COVID-19 cases are equivalent, or *Fallacy of Exclusion* by omitting the protection against severe diseases. A clear fallacy class can only be assigned for a specific *verbalized* fallacious premise.

5 Experiments

For each input ($\overline{c}$, p_0, s_i), comprising the incorrect claim, the accurate premise, and the publication context linked to a fallacy, the model must generate at least one fallacious premise $\hat{\overline{p}}_i$ together with the applied fallacy class $\hat{f}_i$. We only experiment in a zero-shot setting, since the dataset construction depends on high-quality HFC articles, which limits size and scalability. However, we separate 30 arguments as a validation split[5] to allow for a prompt selection without compromising the evaluation on the unseen test split, which comprises the remaining 154 arguments with 363 fallacious reasoning lines R_i bridging different reasoning gaps.

5.1 Metrics

Even though multiple interchangeable fallacies may be applicable, only one of them is required to reconstruct the fallacious argument. Hence, to evaluate the fallacy classes, we report P@1 as our primary metric, where the top-ranked predicted fallacy class $\hat{f}_{i,1}$ is considered correct if it matches any gold fallacy class $f_{i,j}$ of the interchangeable fallacy classes in R_i. Further, we model the fallacy classification as a multi-label, multi-class classification problem, in which we ask the model to identify *all* interchangeable fallacy classes within each R_i. While the single-label classification measures sufficiency, the multi-label multi-class classification relates to the comprehensiveness of the detected fallacy classes. Akin to previous fallacy detection work (Dimitrov et al., 2021; Jin et al., 2022) with high class-imbalances, we report the micro F1-score. Additionally, we assume that correctly detecting at least one fallacy is sufficient to reject the claim, and report argument-level accuracy, denoted as *Arg@1*, by considering an argument as rejected if the top-ranked fallacy class prediction of any fallacious reasoning R_i is correct.

To evaluate the generated fallacious premises, we first match the top-ranked generated premises with the gold premises as reference texts via the predicted fallacy class and cosine similarity (cf. §D.1). We then report METEOR score (Banerjee and Lavie, 2005), which was used for rationales in the real-world AFC dataset AVeriTeC (Schlichtkrull et al., 2023b), and BERTScore (Zhang et al., 2020) to account for semantic similarity. Further, we follow Honovich et al. (2022) who use a T5 (Raffel et al., 2020) model trained on NLI data and consider the predicted probability for the entailment label as measure. Rather than using the entailment probability given the reference premise $\overline{p}_i$ as *premise* and the generated premise $\hat{\overline{p}}_i$ as *hypothesis* $e(\overline{p}_i, \hat{\overline{p}}_i)$ (denoted as *NLI-A*), we additionally compute a symmetric variant (*NLI-S*) via $\max\left[e(\overline{p}_i, \hat{\overline{p}}_i); e(\hat{\overline{p}}_i, \overline{p}_i)\right]$ to not penalize a model if the generated premise is more specific than the reference premise. More details about the metrics and matching with reference text are provided in §D.1. Finally, we measure the LLM's internal consistency of the generated premise and fallacy class, by prompting the same LLM again to classify the fallacy present in the generated fallacious premise $\hat{\overline{p}}_i$ given ($\overline{c}$, p_0, $\hat{\overline{p}}_i$). We report the percentage in which the same fallacy class is predicted.

5.2 Models

We evaluated two baselines that predict a randomly selected fallacy class. For the fallacious premise, these baselines either always predict the claim $\overline{c}$ or the accurate premise p_0, both of which are topically related to the gold fallacious premise but meaningless in their verbalized reasoning.

We conducted experiments with two state-of-the-art LLMs: LLaMA 2 (70B) (Touvron et al., 2023)

[5] No misrepresented publication occurs in *test* and *dev*.

Model	*Fallacy* P@1	Arg@1	F1 (*micro*)	*Fallacious premise (@1)* METEOR	BERTScore	NLI-A	NLI-S	*Consistency* Matches@1 (%)
random + claim	0.131	0.264	0.117	0.181	0.611	0.120	0.130	–
random + p_0				0.188	0.599	0.062	0.067	–
LLaMA 2 (D)	0.223	0.416	0.233	0.222	0.617	0.123	0.148	40.5
LLaMA 2 (DE)	0.209	0.422	0.232	0.229	0.621	0.124	0.148	34.7
LLaMA 2 (DL)	0.196	0.409	0.211	0.203	0.616	0.130	0.143	41.0
LLaMA 2 (DLE)	0.209	0.416	0.233	0.207	0.616	0.129	0.145	19.3
LLaMA 2 (L)	0.193	0.377	0.208	**0.253**	**0.627**	**0.140**	**0.165**	54.5
LLaMA 2 (LE)	0.212	0.409	0.222	0.180	0.609	0.121	0.134	43.0
GPT 4 (D)	**0.317**	**0.571**	**0.297**	0.239	0.619	0.069	0.126	**61.2**
GPT 4 (L)	0.292	0.526	0.290	0.238	0.613	0.064	0.140	**61.4**

Table 2: **Argument Reconstruction Results:** Evaluation of LLaMA 2 (70B) and GPT 4 over the predicted fallacy class and the generated fallacious reasoning. We report the performance when using prompts with fallacy *(D)*efinitions, *(L)*ogical forms and/or *(E)*xamples, and provide a consistency estimate of the LLM by asking each LLM to separately classify the fallacy present in the generated premise.

as an open-source LLM which can be run on a local machine, and GPT 4 (OpenAI, 2023) as a proprietary LLM. In line with our annotation process, we prompted the LLM to generate a ranked list of multiple pairs consisting of the fallacious premise and fallacy class ($\hat{\overline{p}}_{i,j}$, $\hat{f}_{i,j}$), which may express interchangeable fallacy classes. We evaluated different prompts, varying in the amount of information provided to the LLMs about the fallacy classes. Specifically, we examined the impact of fallacy definitions (*D*), the logical form (*L*), and the examples (*E*) from our fallacy inventory (cf. §B.1), sourced from Bennett (2012) and Cook et al. (2018). The definitions offer descriptive information about the fallacies. The logical forms abstract from the content, but explicitly indicate the applied fallacious reasoning. For instance, the logical form for the *Fallacy of Composition* is *"A is part of B. A has property X. Therefore, B has property X."*. This resembles surface patterns that were found to be beneficial in logical fallacies (Jin et al., 2022). We hypothesize that different types of information have varying effects on fallacy classification and fallacious premise generation. We selected the best prompt based on the P@1 performance on the validation split (cf. §D.2). For GPT 4, we only report the results based on the respective best LLaMA 2 prompts for fallacy premise generation and fallacy class P@1 on the test set for comparison. Prompts and hyper-parameters are in §F.

5.3 Argument Reconstruction Results

Table 2 shows the results for reconstructing the fallacious argument. LLaMA 2 achieves its best fallacy detection (P@1) and fallacious premise generation performance using (*D*) or (*L*) in the prompt, respectively, which is consequently reported for GPT 4. For both LLMs, using only the fallacy definition leads to the best fallacy classification performance. Here, GPT 4 outperforms LLaMA 2 by a large margin, correctly identifying at least one fallacy in 57% of the arguments. For fallacious premise generation, each LLM exhibits the best performance based on different prompts. In the fallacious premise generation, even GPT 4 achieves low scores, particularly compared to the random baselines. The generated premises perform primarily poorly when the predicted fallacy class does not match the gold fallacy class of the reference premise. When separately evaluating the generated fallacious premises over correctly classified fallacies with matching classes only (cf. §D.3; Table 10), GPT 4 surpasses the random baseline and outperforms LLaMA 2 in METEOR (0.264 vs. 0.243) and BERTScore (0.637 vs. 0.622), reaching comparable performance in NLI-S (0.267 vs. 0.266). Finally, both models seem to show (slightly) improved consistency (last column) when given the logical form, suggesting that it helps the fallacy generation to match the expected form. Overall, the consistency is much higher for GPT 4.

5.4 Fallacy Classification Results

We instructed LLMs to classify the applied fallacy class $f_{i,j}$ based on the *provided* gold fallacious premise $\overline{p}_{i,j}$, along with the claim $\overline{c}$, the accurate premise p_0, and the publication context linked to a fallacy s_i, and report the results in Table 3. Since each fallacious premise $p_{i,j}$ verbalizes a single fallacy class $f_{i,j}$ this becomes a single-label

LLM	Prompt	Acc.	F1
LLaMA 2	–	0.493	0.406
	Def.	0.577	0.464
	Def. + Logical	0.630	0.476
	Def. + Example	0.637	0.476
	Def. + Logical + Example	0.568	0.459
	Logical	0.601	0.472
	Logical + Example	0.645	0.499
GPT 4	Def.	0.738	0.649
	Logical	0.744	0.624
	Logical + Example	**0.771**	**0.682**

Table 3: **Fallacy Classification:** Performance when predicting the gold fallacy class $f_{i,j}$ given the claim $\overline{c}$, the fallacy context s_i and the verbalized fallacious premise $\overline{p}_{i,j}$. We report accuracy and F1-score (macro).

classification problem. These experiments help to (*i*) compare the difficulty of detecting fallacies with *explicit* fallacious reasoning provided or not (as in §5.3), and (*ii*) re-evaluate the LLMs and the prompts used to assess the consistency over the gold fallacious premises. In addition to exploring the impact of (*D*, *L*, *E*), we also evaluated the performance when only provided with the fallacy names (first row), to assess whether sufficient fallacy knowledge was acquired during pretraining. For GPT 4, we evaluated the best prompt based on LLaMA 2 performance, as well as the prompts used to measure consistency in Table 2. Both LLMs performed strong across all prompts, especially considering that this is a 9-way classification problem. In §E.3, we further provide empirical evidence that GPT 4 benefits from the premise generation task, when no gold fallacious premise is available. The accuracy over the gold premises always exceeded the consistency scores, suggesting that model-generated premises were not as clear-cut to a single fallacy class compared to gold fallacious premises, even by the LLMs' own judgement. The primary misclassification for both LLMs occured between *Ambiguity* and *False Equivalence* (cf. §D.4), two very related fallacies (cf. Figures 2 & 3). However, LLaMA 2 overpredicted *False Equivalence* in general. The best performance was reached with access to the logical form and the examples. We hypothesize that both were most influential to our annotators, and are hence helpful for detecting their generated fallacies.

Figure 4: **Performance per Fallacy:** F1-score per predicted fallacy class from a multi-label multi-class perspective considering all model predictions.

6 Analysis

6.1 Fallacy-level Performance

We assess the performances per fallacy class for the best prompts (*D*) in §5.3 for both LLMs in Figure 4. Specifically, we report the fallacy-level F1-score in a multi-label multi-class setting. GPT 4 outperforms LLaMA 2 in almost all classes. The strongest F1-score by LLaMA 2 is achieved for the *False Equivalence* class. This aligns with the (*LE*) prompted LLMs from Table 3, analyzed in §D.4, and primarily stems from a high recall for detecting this fallacy. For all other fallacy classes, GPT 4 achieves a substantially higher recall, leading to an overall higher performance in terms of F1-score, given the mostly similar precision across fallacy classes (cf. §E.2; Figure 16). We observe the biggest difference for *Ambiguity* and *Impossible Expectations*, which are frequently detected by GPT 4, but never by LLaMA 2. The same was observed when prompting LLMs to predict fallacy classes applied by the gold fallacious premises (cf. §D.4; Figures 13 & 14), suggesting that the differences were inherent to the LLMs. Interestingly, both LLMs perform best on the most frequent fallacy classes despite no fine-tuning involved.

6.2 Allowing Multiple Predictions

Generally, we assume that our annotators identified the most fitting fallacies that should be among the top model predictions. However, different applicable fallacy classes may exist and our annotations cannot guarantee full recall. To address this, in Figure 5, we evaluate all LLMs in a more lenient setting. We borrow the HasPositive@k metric from Shaar et al. (2020) and consider a detected fallacy class correct, if any of the top k predicted fallacy classes is accurate. This approach avoids penalizing models for predicting different fallacies, as long as they also predict a gold fallacy class. The

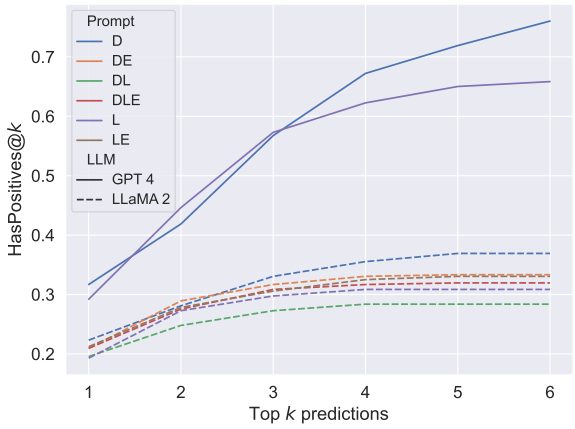

Figure 5: **Relaxed Fallacy Detection:** Performance when a fallacy is considered as correct, if the model predicts a fallacy class within the top k results that matches any of the gold interchangeable fallacy classes.

results demonstrate a consistent improvement in the performance of GPT 4 as more predictions are considered. In contrast, LLaMA 2, in most cases, fails to predict the gold fallacy, even within the top 6 predictions. This confirms that GPT 4's superiority on this task is not a result of subtle selection bias for the top-ranked fallacy class, but arises from its better ability to identify the required fallacy classes. GPT 4 especially benefits from improving the detection of *Fallacy of Exclusion* and *False Equivalence* when increasing k (cf. §E.1), which account for 45.3% of all 550 interchangeable fallacies in MISSCI, and nearly doubles the accuracy when considering the top 3 predictions. The argument-level performance (Arg@k) peaks at a maximum of 89.0% for GPT 4 and 59.7% for LLaMA 2 for $k = 6$.

6.3 Human Evaluation

We manually evaluated 240 predictions from the main experiments in Table 2 (60 per LLM with (*D*) and (*L*) prompts; 50% for correctly and incorrectly classified fallacies based on P@1). Table 4 shows the estimated overall results. Additional results and details are provided in §E.4. Our human judgment found the fallacies produced by GPT 4, in particular (*L*), the most plausible. Yet, the predicted fallacy class often did not match the premise. Overall, we observed a major quality difference of generated premises across both LLMs, with LLaMA 2 often repeating parts of the input. NLI-S showed the strongest correlation with human judgements (Pearson r=0.209; p-value=0.001; cf. §E.4). Due to the complexity of this task and its automatic evaluation,

Model	Applicable premise	Correct class
LLaMA 2 (*L*)	0.167	0.040
LLaMA 2 (*D*)	0.233	0.107
GPT 4 (*L*)	**0.867**	**0.503**
GPT 4 (D)	0.674	0.481

Table 4: **Human Evaluation:** Assessment if the generated fallacious premises are *applicable* to bridge the reasoning gap, and if the predicted fallacy *class* is applied by the generated and applicable premise.

we echo Schlichtkrull et al. (2023b) who argue that human evaluation is necessary for robustness.

7 Discussion

Following suggestions in Schlichtkrull et al. (2023a), we outline how our research contributes to combating misinformation. The analyzed *data subjects* and *data actors* are social media users. For responsible applications, we emphasize that *data owners* should ideally have domain expertise, recognizing that any system will inevitably be imperfect. We strictly did not ask the models to assign an overall rating of the claim's veracity. Instead, we kept the user in the loop for decision-making and only assisted by outlining the fallacious discrepancies between the cited publication and the claim. Clearly communicating the inaccuracies behind a claim is important for effective debunking (Schmid and Betsch, 2019; Lewandowsky et al., 2020) and can help to increase digital literacy, which is important for building resilience against misinformation (Lewandowsky and Van Der Linden, 2021; Musi et al., 2022). While previous approaches taught digital literacy using serious games (Roozenbeek and van der Linden, 2020; Musi et al., 2023) that require active participation, we envision a system that supports passive consumers of social media.

8 Conclusion and Future Work

We introduced MISSCI, a novel dataset to combat real-world misinformation that misrepresents scientific publications. We proposed a novel task formulation to automatically reconstruct the fallacious reasoning through logical arguments based on the cited publication's content. We showcased MISSCI as a testbed for evaluating the reasoning abilities of LLMs. Our experiments on two LLMs demonstrated the potential for reconstructing fallacious arguments. In future work, we plan to use MISSCI with different LLMs, domains, and languages.

Limitations

To reconstruct fallacious arguments, we solely relied on the expertise of a single fact-checking organization. MISSCI is limited to this organization's selected claims, topics, and biases. While fallacies are derived from reasoning flaws detected by the HFC, separating fallacious reasoning from valid reasoning is not always clear-cut. Generalizing or abstracting from specific observations is an essential part of reasoning (Bennett, 2012) and some argue that fallacy theory in general has limited applicability for real-world claims (Boudry et al., 2015). When selecting claims for fact-checking, the virality of claims is a major factor (Arnold, 2020). It is, therefore, likely that information about the claim, and why it is inaccurate may have been acquired by the LLMs during pretraining (Magar and Schwartz, 2022), similarly to leaked evidence effects observed in fact-checking (Glockner et al., 2022). While MISSCI addresses real-world misinformation, it is just one step toward detecting such fallacies: Our design choices exclude joint use of multiple publications to derive a claim. Assessing a claim by its cited source as done in this work is necessary but insufficient; verifying a claim in the real world requires consultation with complementary sources and domain experts (Silverman, 2014). Our approach requires knowledge of the misrepresented publication, which may not always be provided together with the claim.[6] Moreover, MISSCI does not consider the original content of the misrepresented publication, but relies on paraphrasing from HFC articles, which is not available in real-world applications. Finally, MISSCI comprises pure misinformation, and our results offer no insight into model performance over accurate claims. For practical utility, fallacy detection systems must discern whether a fallacy is present before selecting the specific type of fallacy. We note that including unbiased accurate claims is challenging, as they likely differ in topic, specificity, and may cite multiple scientific publications. Due to these limitations, neither the tested models nor any derived from MISSCI in this form should be directly applied in the real world.

Ethics Statement

The objective of this work, to combat misinformation and to increase the public resilience to it, is ethically uncritical and beneficial to society. Nevertheless, our work bears the danger of undesired side effects. Although our task definition is clearly bound to the content used when deriving a claim, our evaluation may favor models that align with the best knowledge available during COVID-19, which makes up the majority of our dataset. Yet, scientific knowledge may change over time, which will not be reflected in MISSCI. Moreover, we task the models to produce fallacious reasoning. This is important to explain the fallacious reasoning behind a claim for debunking (Lewandowsky et al., 2020), yet it may also be misused by malicious actors. Nevertheless, we argue that our work is rather a further demonstration of how generating fallacies in a controlled setup can be used for good, and aligns with previous work that generated misinformation to improve NLP-based approaches (Zellers et al., 2019; Huang et al., 2023; Alhindi et al., 2023). We did not take any steps to anonymise the collected data. All claims in MISSCI are taken from HFC articles which often focus on claims by public figures. We neither contacted the individuals making the claim, nor the HFC. Following Schlichtkrull et al. (2023b) we will remove claims from MISSCI upon request by any individual that stated the claim, is subject of the claim or created the HFC article.

Acknowledgments

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE, and by the the German Research Foundation (DFG) as part of the UKP-SQuARE project (grant GU 798/29-1). Yufang Hou is supported by the Visiting Female Professor Programme from TU Darmstadt. We gratefully acknowledge the support of Microsoft with a grant for access to OpenAI GPT models via the Azure cloud (Accelerate Foundation Model Academic Research). We are grateful to our dedicated annotators who helped to create MISSCI. Finally, we wish to thank Jan Buchmann, Nils Dycke, Aniket Pramanick, Luke Bates and Jing Yang for their valuable feedback on an early draft of this work.

[6] Although, by not providing evidence for the claim, the *Evading the Burden of Proof* fallacy applies.

References

Bea Alex, Claire Grover, Rongzhou Shen, and Mijail Kabadjov. 2010. Agile Corpus Annotation in Practice: An Overview of Manual and Automatic Annotation of CVs. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 29–37, Uppsala, Sweden. Association for Computational Linguistics.

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, Smaranda Muresan, and Preslav Nakov. 2023. Large Language Models are Few-Shot Training Example Generators: A Case Study in Fallacy Recognition. *ArXiv preprint*, abs/2311.09552.

Phoebe Arnold. 2020. The challenges of online fact checking: how technology can (and can't) help. Technical report, FullFact.

Isabelle Augenstein. 2021. Determining the Credibility of Science Communication. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 1–6, Online. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Bo Bennett. 2012. *Logically Fallacious: The Ultimate Collection of Over 300 Logical Fallacies (Academic Edition)*. eBookIt.com.

Claire Bonial, Austin Blodgett, Taylor Hudson, Stephanie M. 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, Marseille, France. European Language Resources Association.

Maarten Boudry, Fabio Paglieri, and Massimo Pigliucci. 2015. The fake, the Flimsy, and the Fallacious: Demarcating Arguments in Real Life. *Argumentation*, 29:431–456.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Luke Bratton, Rachel C Adams, Aimée Challenger, Jacky Boivin, Lewis Bott, Christopher D Chambers, and Petroc Sumner. 2019. The association between exaggeration in health-related science news and academic press releases: a replication study. *Wellcome open research*, 4(148).

J. Scott Brennen, Felix M. Simon, Philip N. Howard, and Rasmus Kleis Nielsen. 2020. Types, Sources, and Claims of COVID-19 Misinformation. Technical report, Reuters Institute for the Study of Journalism.

Tuhin Chakrabarty, Aadit Trivedi, and Smaranda Muresan. 2021. Implicit Premise Generation with Discourse-aware Commonsense Knowledge Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6247–6252, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chhanda Chakraborti. 2007. *LOGIC: Informal, Symbolic and Inductive*. PHI Learning Pvt. Ltd. Page 48.

John Cook, Peter Ellerton, and David Kinkead. 2018. Deconstructing climate misinformation to identify reasoning errors. *Environmental Research Letters*, 13(2).

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.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. GPT3.int8(): 8-bit Matrix Multiplication for Transformers at Scale. In *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332. Curran Associates, Inc.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting Propaganda Techniques in Memes. 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 6603–6617, Online. Association for Computational Linguistics.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5916–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking.

Transactions of the Association for Computational Linguistics, 10:178–206.

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.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018a. Before Name-Calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018b. The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating Factual Consistency Evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023. Faking Fake News for Real Fake News Detection: Propaganda-Loaded Training Data Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14571–14589, Toronto, Canada. Association for Computational Linguistics.

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.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTaiL: A Textual Entailment Dataset from Science Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing Dataset Annotation Quality Management in the Wild. *Computational Linguistics*, pages 1–48.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Stephan Lewandowsky, John Cook, Ullrich Ecker, Dolores Albarracin, Michelle Amazeen, Panayiota Kendou, Doug Lombardi, E Newman, Gordon Pennycook, Ethan Porter, et al. 2020. *The Debunking Handbook 2020*.

Stephan Lewandowsky and Sander Van Der Linden. 2021. Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology*, 32(2):348–384.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. SCITAB: A Challenging Benchmark for Compositional Reasoning and Claim Verification on Scientific Tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7787–7813, Singapore. Association for Computational Linguistics.

Inbal Magar and Roy Schwartz. 2022. Data Contamination: From Memorization to Exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. 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*, 12(3).

Elena Musi, Elinor Carmi, Chris Reed, Simeon Yates, and Kay O'Halloran. 2023. Developing Misinformation Immunity: How to Reason-Check Fallacious News in a Human–Computer Interaction Environment. *Social Media + Society*, 9(1).

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Callistus Ireneous Nakpih and Simone Santini. 2020. Automated Discovery of Logical Fallacies in Legal Argumentation. *International Journal of Artificial Intelligence and Applications (IJAIA)*, 11.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv preprint*, abs/2303.08774.

Jiaxin Pei and David Jurgens. 2021. Measuring Sentence-Level and Aspect-Level (Un)certainty in Science Communications. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9959–10011, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Jon Roozenbeek and Sander van der Linden. 2020. Breaking Harmony Square: A game that "inoculates" against political misinformation. *The Harvard Kennedy School Misinformation Review*, 1(8).

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.

Muhammad Salman, Asif Hanif, Shady Shehata, and Preslav Nakov. 2023. Detecting Propaganda Techniques in Code-Switched Social Media Text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16794–16812, Singapore. Association for Computational Linguistics.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based Fact-Checking of Health-related Claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Schlichtkrull, Nedjma Ousidhoum, and Andreas Vlachos. 2023a. The Intended Uses of Automated Fact-Checking Artefacts: Why, How and Who. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8618–8642, Singapore. Association for Computational Linguistics.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023b. AVeriTeC: A Dataset for Real-world Claim Verification with Evidence from the Web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Philipp Schmid and Cornelia Betsch. 2019. Effective strategies for rebutting science denialism in public discussions. *Nature Human Behaviour*, 3(9):931–939.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a Known Lie: Detecting Previously Fact-Checked Claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

Craig Silverman. 2014. *Verification Handbook: An Ultimate Guideline on Digital Age Sourcing for Emergency Coverage*. European Journalism Centre.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, Fred Boy, and Christopher D Chambers. 2014. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *BMJ*, 349.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay

Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *ArXiv preprint*, abs/2307.09288.

Juraj Vladika and Florian Matthes. 2023. Scientific Fact-Checking: A Survey of Resources and Approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230, Toronto, Canada. Association for Computational Linguistics.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359:1146–1151.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-Open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Claire Wardle. 2018. 5 Lessons for Reporting in an Age of Disinformation. Technical report, First Draft.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Dustin Wright and Isabelle Augenstein. 2021. Semi-Supervised Exaggeration Detection of Health Science Press Releases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10824–10836, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dustin Wright, Jiaxin Pei, David Jurgens, and Isabelle Augenstein. 2022. Modeling Information Change in Science Communication with Semantically Matched Paraphrases. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1783–1807, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Amelie Wührl, Dustin Wright, Roman Klinger, and Isabelle Augenstein. 2024. Understanding Fine-grained Distortions in Reports of Scientific Findings. *ArXiv preprint*, abs/2402.12431.

Bei Yu, Jun Wang, Lu Guo, and Yingya Li. 2020. Measuring Correlation-to-Causation Exaggeration in Press Releases. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4860–4872, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending Against Neural Fake News. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9051–9062.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

A Dataset Construction I: Selecting Misrepresented Scientific Publications From HFC Articles

A.1 URL Filtering

For reproducibility we collect all HTML webpages via the Wayback Machine[7]. We apply the following filtering on URLs that may be relevant to our instances. We initially only include URLs with the following top level domains:

- *".gov", ".org", ".int", ".edu", ".gov.uk", ".org.uk", ".gov.au", ".org.nz", ".edu.au", ".gov.in", ".org.au", ".ac.uk".*

We remove commonly occurring fact-checking organizations that are within the applied filtering: *"fullfact.org", "www.poynter.org", "factcheck.org", "npr.org"*. We finally add known publishers of scientific content that would otherwise be removed via our filtering step:

- *"nature.com", "jamanetwork.com", "thelancet.com", "researchgate.net", "academic.oup.com", "bmj.com", "onlinelibrary.wiley.com", "www.mdpi.com", "www.ijidonline.com", "link.springer.com", "sciencedirect.com", "tandfonline.com", "journals.lww.com", "cell.com", "papers.ssrn.com", "cebm.net", "thejournal.ie", "cebm.ox.ac.uk", "elsevier.com", "biomedcentral.com", "journalofinfection.com", "journals.sagepub.com", "scientificamerican.com", "pfizer.com", "www.the-scientist.com", "www.cancer.net", "www.ema.europa.eu"*

Finally, we keep archived URLs

- *"archive.is", "archive.ph", "archive.md", "archive.vn", "perma.cc", "archive.fo"*

as it is unknown from the surface form if it refers to a scientific publication or not.

A.2 Annotation Process

We selected 8,695 links for annotation, using a curated list of reputable scientific publishers and top-level domains (cf. §A.1). Few, if any, links in a fact-checking article point to a misrepresent scientific publication. HFC articles must be assessed in their entirety as critical statements may be scattered. Further, articles may cover multiple related claims, or various subclaims, each possibly misrepresenting different publications. For efficient annotation, we grouped up to 8 distinct links per fact-checking article into one HIT, highlighting each link in a different color (Figure 6). This resulted in 1,385 HITs for annotation.

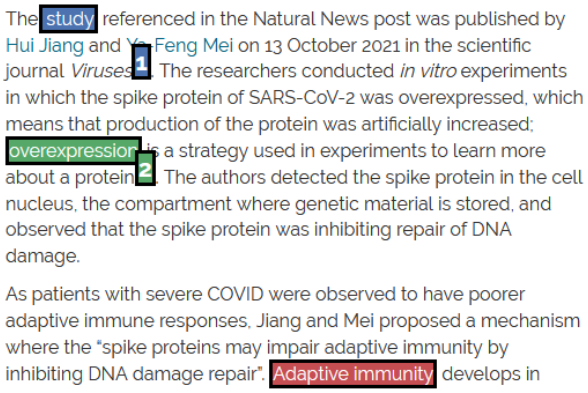

Figure 6: **Annotation (Step 1):** A preprocessed HFC article. Links for annotation are highlighted in color.

Each highlighted link was given to the annotators within the original context of the HFC article, to decide if the link points to a scientific publication that is misrepresented by a non-true claim as discussed in the article at hand. For each misrepresented publication, annotators had to provide triplets consisting of (1) a non-true claim that misrepresents (2) a scientific publication, and (3) a justification of the flawed reasoning. Each triplet requires an extracted statement from the article as justification, explaining how the document could be misused to support the claim ($S \Rightarrow \bar{c}$) and why it does not ($S \not\Rightarrow \bar{c}$). Consecutive HITs present the same HFC article to prevent context switching. Annotators label a link as "misrepresented" only if the claim explicitly relied on the linked publication's content. Indirect misrepresentation (e.g., through a press release) is also classified as "misrepresented" as the claim relied on the same content. Remaining links are grouped into three sub-categories during annotation (cf. §A.3), but these are collapsed to "not applicable" during dataset compilation and discarded from further annotation steps.

In a pilot study with all 927 selected links from 50 fact-checking articles we found that single annotation instead of double annotation (Krippendorff's α of 0.728; cf. §A.4) results in a minor loss in recall only. Hence, each instance is annotated by a single annotator. Note that redundant annotations mainly impact the recall of misrepresented scientific documents. Data quality remains unaf-

[7] https://archive.org/web/

fected, as errors in link selection are rectified in later annotation tasks. In total, we identified 208 scientific publications labeled as "misrepresented" across 150 HFC articles. 107 articles contained only one misrepresented publication, while 43 articles contained 2-4 misrepresented publications.

Definitions. When selecting the triplets consisting of (1) a non-true claim that misrepresents (2) a scientific publication, and (3) a justification of the flawed reasoning we consider the following definitions: We consider a claim as *non-true* and *misrepresenting* if it is not fully accurate (e.g., including *partly true* or *misleading* claims) and lack a valid entailment relationship ($\not\Rightarrow$) with the cited document. We exclude documents from the misrepresented category if the claim can be validly inferred from an incorrect source or if refutation requires additional external evidence. We consider a publication as *scientific* if it is published in a scientific venue, may be submitted to such a venue (preprints), or constitutes a scientific report from a credible institution, (e.g., annual CDC reports) and include non-peer-reviewed documents because they can be misrepresented before being accepted.

A.3 Link Annotation: Fine-grained labels

Annotators categorize each link into one of four categories:

- **Misrepresented:** This category is designated for scientific publications that are explicitly misrepresented by a non-true claim. Such publications may be referenced either directly or indirectly, for example, through a press release.
- **Misrepresentable:** This category is assigned to publications that were not misrepresented but have the potential to be misrepresented. This occurs when the HFC discuss related scientific documents for a comprehensive overview. While these documents could be susceptible to misrepresentation by similar claims, they haven't been misrepresented by the claimant.
- **Maybe Misrepresentable:** Annotators can choose this category when uncertain. Uncertainty may stem from ambiguity regarding a document's scientific status or doubts about misrepresentation.
- **Not Applicable:** This category applies to all other links not covered by the previous categories.

Annotators must provide an explanation for labels other than "not applicable". In MISSCI, we focus exclusively on real-world misinformation involving genuinely misrepresented scientific documents. Instances not labeled as "misrepresented" are collapsed into the "not applicable" class and excluded. While we exclude links labeled "misrepresentable" or "maybe misrepresentable" from MISSCI, we provide all annotations using the fine-grained taxonomy.

A.4 Link Annotation: Pilot Study and Final Results

Agreement. The annotators, alongside one author, annotated all 221 links from 16 randomly chosen fact-checking articles. The inter-annotator agreement, assessed with Krippendorff's α, was 0.360. Disagreement primarily arose from cases initially marked as "misrepresentable" or "maybe misrepresentable" later grouped into the "not applicable" category (as per §A.3). Using the grouped labels, we calculated binary inter-annotator agreement between "misrepresented" labels and "not applicable" labels. This resulted in an inter-annotator agreement of 0.751. The annotators then double-annotated all 706 links from additional 34 randomly selected fact-checking articles, achieving comparable inter-annotator agreement of 0.728.

Single annotations are sufficient. We assess the value of having two annotations versus one using the double-annotated data. Specifically, if at least one annotator labels an instance as "misrepresented" we classify it as such. A single annotator identifies 78.3% of the same instances as "misrepresented". When we also consider instances labeled as "misrepresentable" by the single annotator, 95.5% of the presumed "misprepresented" double-annotated instances are detected. These additional cases were labeled as "misrepresentable" only by the second annotator, indicating more uncertainty. To reduce the workload while maintaining sufficient coverage, all remaining instances were annotated by a single annotator.

Results. In total, we found 208 (2.4%) scientific publications labelled as "misrepresented", 425 (4.9%) labelled as "misrepresentable", and 596 (6.9%) labelled as "maybe misrepresentable". The remaining 7,466 (85.9%) links were unrelated to our problem.

B Dataset Construction II: Fallacious Argument Reconstruction

B.1 Fallacious Reasoning for Misrepresented Science

Fallacy Inventory Selection. To select a suitable fallacy inventory, we begin by examining the fallacies employed by Cook et al. (2018) as they pertain to misinformation within the scientific domain. A distinction lies in the relation to science, as they focus on climate-change denial whereas our focus lies on the misrepresentation of scientific documents that seemingly *support* the claims. Consequently, we exclude fallacies like *Red Herring* which divert attention from opposing arguments, or *Fake experts*, which contradicts our requirement for credible evidence. We select the remaining fallacies by examining instances of misrepresentation of scientific publications, guided by the collection of logical fallacies from Bennett (2012). An overview of all selected fallacies can be found in Table 5, along with examples in Table 6.

Merged Fallacies. After annotation, we merge several fallacy classes due to difficulties in differentiation based solely on the information from the fact-checking article, or because they share similar traits and one is very infrequent:

- **Fallacy of Division/Composition**: Combines *Fallacy of Division* and *Fallacy of Composition* as both involve generalizations through the *part-of* relationship.
- **Causal Oversimplification**: Merges *Single Cause* and *False Cause*as they are often indistinguishable based solely on the fact-checking article.
- **Ambiguity**: Combines *Ambiguity* with its subtype *Equivocation*, which relies on the same vocabulary in the claim and premises, a detail not accessible during annotation.

Fallacies annotated as *Other* were resolved into one of the existing fallacy classes. This was always possible, as *Fallacy of Exclusion* can be applied in almost all cases (as it ignores compromising content (or *context*) of S.

Fallacy Inventory Discussion. Several fallacies, such as *False Causality*, *Hasty Generalization*, *False Dilemma* or *Ambiguity* are present in most existing NLP fallacy inventories in some form. A distinction of our selected fallacies lies in our exclusive focus on corrupted support ($\not\Rightarrow$) relationships. For this reason, we exclude several fallacies that are commonly used in fallacy detection datasets and important for propaganda techniques (Da San Martino et al., 2019; Piskorski et al., 2023) or misinformation in general (Musi et al., 2022; Alhindi et al., 2022):

1. Fallacies that **attack** (e.g. *Ad Hominem*).
2. Fallacies that **divert** (e.g. *Strawman Fallacy*).
3. Fallacies that use **manipulation techniques** like slogans or emotional language (e.g. *Appeal to Emotion*).
4. Fallacies that utilize **non-credible evidence** (e.g. *False Authority*), or **omit evidence** altogether (e.g. *Evading the Burden of Proof*).

Fallacy Inventory Comparison. In contrast to other works (Musi et al., 2022; Alhindi et al., 2022), a strong focus lies on a detailed analysis of generalization fallacies. We employ various specific generalization fallacies, such as *Hasty Generalization*, *Biased Sample Fallacy* and *False Equivalence*. The fallacy of *Cherry Picking* has been addressed in prior research. However, our interpretation of this fallacy aligns more closely with *Slothful Induction*. This distinction arises because our focus is not on presenting selectively chosen information but rather on the omission of crucial aspects of the study that weaken the claim's validity. We find *False Analogy* as employed by e.g. Alhindi et al. (2022), to be comparable to *False Equivalence*. Both fallacies can be applied to the same problems. Without access to the claimant's specific reasoning, we find it impossible to prioritize one over the other when generating fallacious premises. For similar reasons, we adopt a broad definition of the *False Dilemma* fallacy, which assumes that only two options (or outcomes) exist when, in reality, more options are available. We consolidate it with the fallacy of *Affirming the Disjunct*, which assumes an "either/or" possibility among different options, even when these options are not mutually exclusive. Despite their differences (in *False Dilemma* the options are indeed mutually exclusive but not exhaustive), they share similar characteristics in exhibiting black-and-white thinking and may only differ in their specificity of the explicit reasoning that is unavailable to us.

B.2 Argument Reconstruction Guidelines

Annotators should base their reconstruction of fallacious arguments on content in the fact-checking article. The final argument should be:

Definition	Logical Form
AMBIGUITY When an unclear phrase with multiple definitions is used within the argument; therefore, does not support the conclusion.	*Claim X is made. Y is concluded based on an ambiguous understanding of X.*
EQUIVOCATION (merged with **AMBIGUITY**) When the same word (here used also for phrase) is used with two different meanings. Equivocation is a subset of the ambiguity fallacy.	*Term X is used to mean Y in the premise. Term X is used to mean Z in the conclusion.*
IMPOSSIBLE EXPECTATIONS / NIRVANA FALLACY Comparing a realistic solution with an idealized one, and discounting or even dismissing the realistic solution as a result of comparing to a "perfect world" or impossible standard, ignoring the fact that improvements are often good enough reason.	*X is what we have. Y is the perfect situation. Therefore, X is not good enough.*
FALSE EQUIVALENCE Assumes that two subjects that share a single trait are equivalent.	*X and Y both share characteristic A. Therefore, X and Y are [behave] equal.*
FALSE DILEMMA Presents only two alternatives, while there may be another alternative, another way of framing the situation, or both options may be simultaneously viable.	*Either X or Y is true.*
BIASED SAMPLE FALLACY Drawing a conclusion about a population based on a sample that is biased, or chosen in order to make it appear the population on average is different than it actually is.	*Sample S, which is biased, is taken from population P. Conclusion C is drawn about population P based on S.*
HASTY GENERALIZATION Drawing a conclusion based on a small sample size, rather than looking at statistics that are much more in line with the typical or average situation.	*Sample S is taken from population P. Sample S is a very small part of population P. Conclusion C is drawn from sample S and applied to population P.*
FALSE CAUSE FALLACY (use as **CAUSAL SIMPLIFICATION**) Post hoc ergo propter hoc — after this therefore because of this. Automatically attributes causality to a sequence or conjunction of events.	*A is regularly associated with B; therefore, A causes B.*
SINGLE CAUSE FALLACY (use as **CAUSAL SIMPLIFICATION**) Assumes there is a single, simple cause of an outcome.	*X is a contributing factor to Y. X and Y are present. Therefore, to remove Y, remove X.*
FALLACY OF COMPOSITION Inferring that something is true of the whole from the fact that it is true of some part of the whole.	*A is part of B. A has property X. Therefore, B has property X.*
FALLACY OF DIVISION (merged with **FALLACY OF COMPOSITION**) Inferring that something is true of one or more of the parts from the fact that it is true of the whole.	*A is part of B. B has property X. Therefore, A has property X.*
FALLACY OF EXCLUSION / CHERRY PICKING / SLOTHFUL INDUCTION When only select evidence is presented in order to persuade the audience to accept a position, and evidence that would go against the position is withheld (Cherry Picking). Ignores relevant and significant evidence when inferring to a conclusion (Slothful Induction – focus on neglect).	*Evidence A and evidence B is available. Evidence A supports the claim of person 1. Evidence B supports the counterclaim of person 2. Therefore, person 1 presents only evidence A.*

Table 5: Fallacy Overview. Definition and logical form taken from Bennett (2012) and Cook et al. (2018).

AMBIGUITY *It is said that we have a good understanding of our universe. Therefore, we know exactly how it began and exactly when.*
EQUIVOCATION *A feather is light. What is light cannot be dark. Therefore, a feather cannot be dark.*
IMPOSSIBLE EXPECTATIONS / NIRVANA FALLACY *Seat belts are a bad idea. People are still going to die in car crashes.*
FALSE EQUIVALENCE *They are both Felidae, mammals in the order Carnivora, therefore there's little difference between having a pet cat and a pet jaguar.*
FALSE DILEMMA *I thought you were a good person, but you weren't at church today.*
BIASED SAMPLE FALLACY *Based on a survey of 1000 American homeowners, 99% of those surveyed have two or more automobiles worth on average $100,000 each. Therefore, Americans are very wealthy.*
HASTY GENERALIZATION *My father smoked four packs of cigarettes a day since age fourteen and lived until age sixty-nine. Therefore, smoking really can't be that bad for you.*
FALSE CAUSE FALLACY *Every time I go to sleep, the sun goes down. Therefore, my going to sleep causes the sun to set.*
SINGLE CAUSE FALLACY *Smoking has been empirically proven to cause lung cancer. Therefore, if we eradicate smoking, we will eradicate lung cancer.*
FALLACY OF COMPOSITION *Hydrogen is not wet. Oxygen is not wet. Therefore, water (H2O) is not wet.*
FALLACY OF DIVISION *His house is about half the size of most houses in the neighborhood. Therefore, his doors must all be about 3 1/2 feet high.*
FALLACY OF EXCLUSION / CHERRY PICKING / SLOTHFUL INDUCTION *Employer: "It says here on your resume that you are a hard worker, you pay attention to detail, and you don't mind working long hours."* *Andy: "Yes sir."* *Employer: "I spoke to your previous employer. He says that you constantly change things that should not be changed, you could care less about other people's privacy, and you had the lowest score in customer relations."* *Andy: "Yes, that is all true, as well."*

Table 6: Fallacy Examples (taken from Bennett (2012)).

- **Comprehensive:** All fallacies identified by the fact-checker should be incorporated.
- **Self-contained:** Subsequent steps should not necessitate the use of the fact-checking articles.

For all text generation tasks our annotators utilize Grammarly[8], integrated into the Surge AI tool, to ensure high-quality text. Each of the previously detected 208 misrepresented links was provided to the annotators together with the the preprocessed fact-checking article with highlighted links, and the justification for labeling the link as "misrepresented" (from §A.2) and annotators were tasked to reconstruct all parts of the fallacious argument. Annotations were conducted in batches of 30-40 arguments per annotator per week with weekly meetings, adhering to agile annotation principles (Alex et al., 2010). Each argument was independently assessed by at least two annotators. At the end of annotation, annotators reviewed their own annotations, excluding the last batch, to rectify errors resulting from initial guideline misunderstandings and enhance consistency.

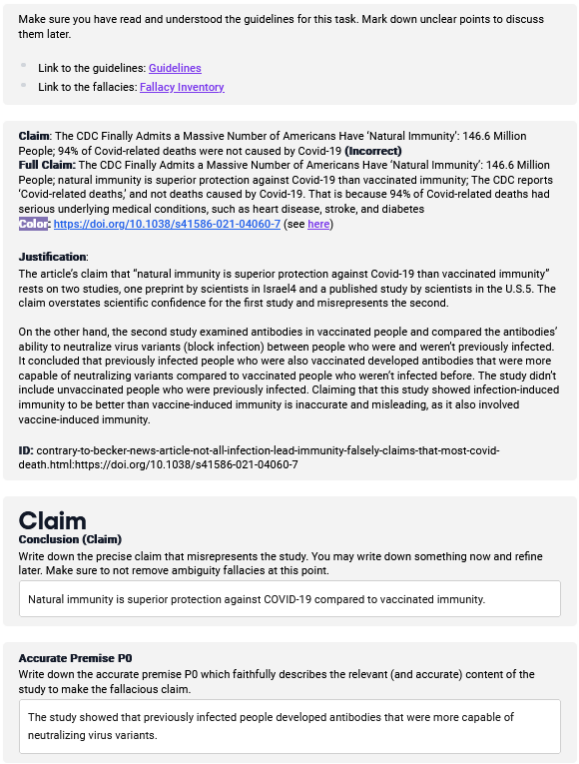

Figure 7: Claim annotation interface from Surge AI.

Claim Rewriting. Fact-checking articles might discuss multiple related claims or complex arguments with subclaims. Annotators should first understand the main claim of the fact-checking article and the misrepresented scientific document. The misrepresenting claim $\overline{c}$ should be formulated as such that $S \cup \overline{P} \Rightarrow \overline{c}$. Annotators should use the main claim of the fact-checking article if possible, and make minimal changes if necessary. While the formulation of $\overline{c}$ is a prerequisite for detecting fallacious reasoning lines R, the validity of $S \cup \overline{P} \Rightarrow \overline{c}$ can only be checked after constructing the argument. Therefore, annotators should re-evaluate $\overline{c}$ after identifying all fallacies. The annotation interface, including a link to the pre-processed fact-checking article and relevant information is shown in Figure 7.

Accurate Premise Writing. The accurate premise p_0 provides a correct description of the misrepresented scientific document S. Its purpose is to offer logical support for the claim ($p_0 \Rightarrow \overline{c}$) but it falls short due to the presence of fallacious reasoning. Fact-checkers always include an accurate description of the misrepresented scientific document. Annotators must locate all relevant information and formulate p_0 s.t.

1. The wording is as precise as possible and uses the HFC vocabulary.
2. All accurate content that strengthens $p_0 \Rightarrow \overline{c}$ is included.
3. Any accurate content that weakens $p_0 \Rightarrow \overline{c}$ is excluded.

We guide annotators to include information in the accurate premise p_0 only if it *increases* the plausibility of $\overline{c}$. Any information (from S) that *decreases* plausibility is paraphrased as s_i and is part of R_i. The publication context s_i is optional and required only if additional information beyond the given p_0 is necessary.

Hidden Premise Writing. Fact-checkers explain how the claim $\overline{c}$ relates to the scientific publication S. This may include additional knowledge not found in $\overline{c}$ or S.[9] For example, to understand why the claim "Cucumber kills lung cancer cells." was made based on the scientific finding that "cucubitacin B promoted lung tumor cell death." one must know that cucumbers contain cucubitacin B. This information is likely not provided by the misrepresented publication itself. Annotators can provide any number of hidden premises which are

[8] https://www.grammarly.com

[9] This relies on subjective judgment since annotators aren't required to read the scientific document.

concise, accurate statements that complement S and are essential in understanding the connection between $\overline{c}$ and S. Each hidden premise should be a single sentence derived from the fact-checking article.

Fallacy Class Selection Preference Annotators are directed to prioritize fallacies that engage with the content (all fallacies except *Fallacy of Exclusion*) rather than ignoring crucial aspects. We allow interchangeable fallacies with distinct classes, but we instruct annotators to prioritize more specific fallacies over broader ones. When multiple fallacies share the same flawed reasoning, annotators should select the most specific fallacy class. For example, when a conclusion is drawn from a biased sample, it can be labeled as the *Biased Sample Fallacy*. Alternatively, it might be seen as the *Single Cause Fallacy* assuming that the properties for which the sample is biased do not impact the conclusion. In this example, both fallacy classes do not differ in their applied reasoning, but only in their level of specificity. Therefore, the more specific *Biased Sample Fallacy* should be preferred.[10]. We provided a taxonomy to the annotators to specify how to choose the more specific fallacy class if multiple apply.

Fallacious Premise Writing. Annotators should thoroughly review the fact-checking article to identify all sections discussing the claim $\overline{c}$ misrepresenting the scientific publication S. These discussions may be distributed throughout the article. Annotators must focus solely on fallacies discernible from the content of S. This excludes other scientific documents used by the HFC to refute $\overline{c}$ or assessing the credibility of S (e.g., if S is retracted or inaccurate). Each fallacy should be separately formulated by the annotator and must be accompanied by a justification extracted from the fact-checking article. The context for each fallacy must be constructed akin to p_0, emphasizing the weakening of the claim rather than its strengthening. The fallacious premise must align with the selected fallacy class and make the fallacious reasoning explicit. Given that selecting the fallacious reasoning is subjective since different fallacious premises p_i can re-instantiate $S \cup p_i \Rightarrow \overline{c}$, annotators are permitted to formulate multiple alternative variants that they consider plausible. The annotation interface for a fallacy is visualized in Figure 8. After constructing all fallacies, annotators are tasked with reviewing their own constructed argument to ensure the coherence of the fallacies, claim, and accurate premise, in their HIT.

[10]This is different to the example in Figure 2, in which different fallacies apply different reasoning.

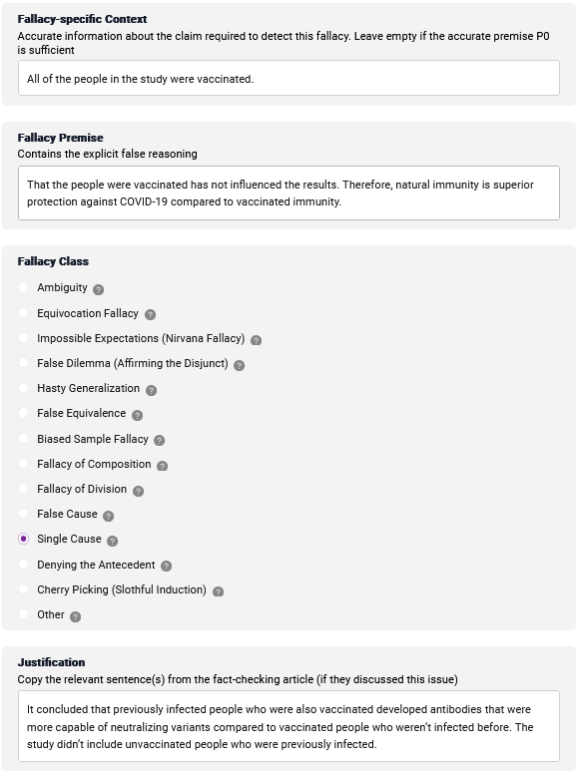

Figure 8: Fallacy annotation interface from Surge AI.

B.3 Argument Consolidation

Our primary annotator, with the most project experience, handled the consolidation process. The consolidator aligned all annotated fallacious reasoning lines, and selected the best verbalized candidate for each $\overline{c}$, $\overline{p}_i$ and s_i, or paraphrased multiple such candidates into an improved version. Interchangeable fallacies with different classes were preserved. Only clearly unjustified fallacy annotations were discarded or corrected if possible. Each consolidated argument underwent double-checking by an author. The consolidator and one of the authors collaborated on a final round of fallacious premise curation. Due to different URLs used to link to the same scientific document, some arguments with duplicate annotations were merged using normalized URLs (cf. §B.4), resulting in 193 distinct fallacious arguments $\mathcal{A}$. Individual annotators might discard arguments during annotation if they cannot reconstruct a fallacious argument, for example because of insufficient information provided by the HFC or

violations of the credibility assumption of the cited publication (we analyze the impact of such cases where fewer annotations are available in §C.3). In seven cases no annotator could reconstruct $\mathcal{A}$, and two more fallacious arguments were discarded during consolidation.

B.4 Scientific Document Annotation.

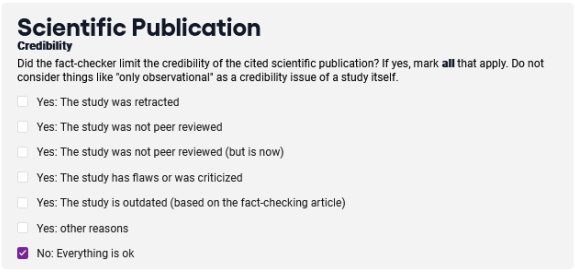

Figure 9: Credibility annotation via Surge AI.

Fact-checkers may use varying links of the same publication within their fact-checking articles. To detect duplicates these links must be normalized. To assess the content of publicly available studies, we require them to be available via PMC. Annotators are tasked to find a URL pointing to the misrepresented scientific publication S and select the first applicable URL of the following list:

1. PMC (available as *fulltext*)
2. sciencedirect (available as *fulltext*)
3. Original Publisher (available as *fulltext*)
4. PubMed (available as *abstract only*)
5. PMC (available as *abstract only*)
6. sciencedirect (available as *abstract only*)
7. Original Publisher (available as *abstract only*)
8. as-is

We also asked whether the HFC highlighted flaws in the study itself (see Figure 9).

C Dataset Analysis

C.1 Publication Credibility

Augenstein (2021) identifies two key challenges in science communication: (1) assessing the credibility of scientific publications, and (2) preventing the misrepresentation of a study's findings. This study addresses the second challenge, while assuming the credibility. To examine to which degree our assumption holds, we report the credibility annotations of the used publications in Table 7. Note that a publication may exhibit more than one credibility issue (e.g., a preprint, that was criticised)

Annotation	Documents	Krippendorff's α
Flawed/Criticised	39	0.504
Preprint	28	0.582
Retracted	2	0.665
Outdated	1	0.0
Any from Above	**61**	**0.577**
Credible	**123**	

Table 7: **Publication Credibility:** Results of the credibility annotations of the misrepresented publications per fallacious argument alongside with the inter-annotator agreement per label.

We consider a publication S as "credible" only if no annotator detected any aspect compromising its credibility in the HFC article, providing a conservative estimate. For 123 arguments, no credibility violations were identified. Even identified violations do not necessarily mean that the publication is not credible or not misrepresented by $\overline{c}$. For instance, preprints lack formal community approval but may still be credible. At the time of annotation, 16 out of 28 publications marked as preprints by the HFC were accepted. Further, criticism of a study does not make it scientifically unsound. In fact, some critics highlight omitted limitations that could lead to misunderstandings similar to the misrepresentations studied in this work. The overall agreement in distinguishing credible documents from those with any credibility issue is 0.577 (Krippendorff's α). Figure 10 displays the co-occurring credibility

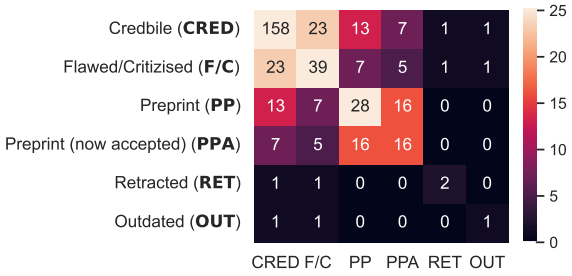

Figure 10: Heatmap of co-occurring credibility annotations per misrepresented document a fallacious argument.

annotations assigned to fallacious arguments. We count an occurrence of a credibility label if at least one annotator assigned it to the misrepresented document, and, hence, ignore duplicate annotations of the same label.

C.2 Claims about the COVID-19 infodemic

MISSCI comprises fallacious arguments from 2014 to 2022. We manually evaluate each argument

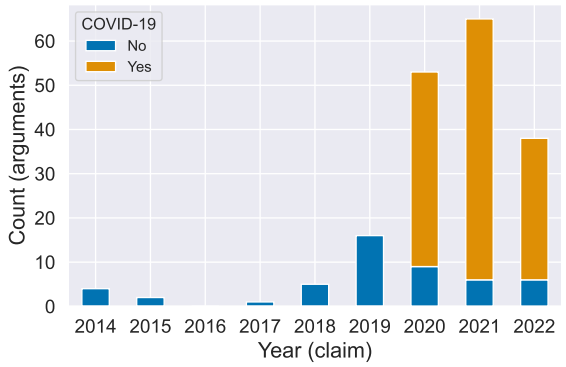

Figure 11: **COVID-19 Arguments:** Collected arguments per year and whether they constitute COVID-19 related misinformation.

Annotators		*Result*
Assigned	**Successful**	**Num. Arguments**
2	1	9
3	1	6
3	2	16
Incl. failed reconstructions		**31**
2	2	83
3	3	70
No failed reconstruction		**153**

Table 8: **Annotation Assignments:** Number of successfully reconstructed argument versions by the number of initially assigned annotators.

for its association with COVID-19 misinformation, considering context provided by the HFC, rather than explicit COVID-19 mentions. For instance, we consider the claim "Ivermectin sterilizes most men who take it" as COVID-19-related because it only exists due to the misinformation about Ivermectin as a COVID-19 cure. Figure 11 illustrates argument distribution over time and their COVID-19 relevance.

C.3 On the Coverage of Fallacies

In some cases, annotators may decide that an argument cannot be reconstructed and discard the HIT. This can for example happen, if the annotator considers the information provided by the HFC insufficient to reconstruct the argument, or if the cited publication itself is non-credible. Table 8 compares the annotation assignments and successful argument reconstructions. A total of 153 arguments were reconstructed by all assigned annotators. In 25 cases, one annotator, and in 6 cases, two annotators could not reconstruct the argument. Given the limited number of annotations and ambiguity when assigning fallacy classes, we aim to shed light into the recall of our detected fallacies. To this end, we compare the number of successfully reconstructed fallacious arguments with the number of detected fallacies (excluding interchangeable fallacies) in Figure 12. More annotators generally lead to more identified fallacies. The majority of arguments with successful reconstructions of two and three annotators comprise two or three distinct fallacies respectively.

The analysis indicates that MISSCI may lack comprehensiveness of fallacies, in arguments with fewer annotations. However, this does not affect our main objective since annotators are likely to identify the most severe fallacies, highlighted by the HFC, that violate *strong inductive support* (cf. §3.2). A single study, under specific conditions, rarely gives unconditional support for any claim c without potential generalizations. While not necessarily or severely fallacious, these generalizations may be less emphasized by HFC, which makes them harder to detect as fallacious reasoning for the annotators. In real-world texts, identifying different fallacies among annotators is not unusual (Da San Martino et al., 2019), and distinguishing between fallacious and accurate reasoning can be subtle (Boudry et al., 2015).

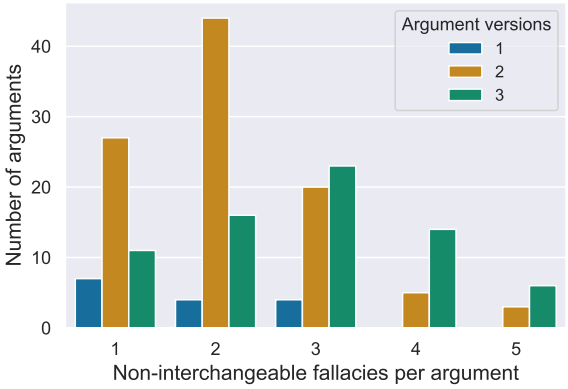

Figure 12: **Argument Annotations:** Number of successfully constructed arguments (*y-axis*) compared to the number of assigned annotators who manually reconstruct each argument (*color*) per distinct fallacious reasoning lines R_i in the argument post consolidation (*x-axis*). The number of arguments shows if one (*blue*), two (*orange*) or three (*green*) distinct arguments versions of the same argument were constructed by the annotators.

D Experiments

D.1 Fallacious Premise Evaluation

We use BERTScore (Zhang et al., 2020) F1 with version 0.3.13, based on DeBERTa (He et al., 2021), fine-tuned on MNLI (Williams et al., 2018), as recommended at the time of this writing.[11] For the NLI-based metric, we use the predicted probability of the entailment label, following (Honovich et al., 2022). We use their provided T5-XXL model (Raffel et al., 2020), fine-tuned on SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), FEVER (Thorne et al., 2018), Scitail (Khot et al., 2018), PAWS (Zhang et al., 2019), and VitaminC (Schuster et al., 2021). The predicted probability for the token "`1`" (representing *entailment*) is used for evaluation.

Multiple interchangeable reference fallacious premises $\overline{p}_{i,j}$ may exist. Therefore, before evaluating the generated fallacious premise $\hat{\overline{p}}_{i,1}$, it needs to be matched with a reference text. For the first-ranked predicted fallacy by the model, which consists of a fallacious premise and fallacy class ($\hat{\overline{p}}_{i,1}$, $\hat{f}_{i,1}$), we use the gold fallacious premise $\overline{p}_{i,j}$ as the reference text if the accompanied gold fallacy class $f_{i,j}$ matches the predicted fallacy class $\hat{f}_{i,1}$. In the absence of matches based on the fallacy class, we select the $\overline{p}_{i,j}$ with the highest cosine similarity to $\hat{\overline{p}}_{i,1}$ from all interchangeable fallacious premises $\overline{p}_{i,j}$. Cosine similarity is measured using the sentence embeddings produced by SBERT (Reimers and Gurevych, 2019) (`all-mpnet-base-v2`).

D.2 Prompt Engineering for Fallacy Generation and Classification

The only existing prompts related to fallacy detection stem from Alhindi et al. (2022). Yet, they were (a) not used in zero-shot experiments, instead relying on fine-tuning, and (b) were used to only classify the fallacy within a single text. Since our task differs and demands a zero-shot setup, we manually select novel prompts for the task. We initially experiment on few instances to derive promising prompt templates. We assess the performance of each prompt with different combinations of (*D*, *L*, and *E*) using LLaMA 2 on the 30 arguments of the validation split in Table 9. All prompts and hyper-parameters are listed in §F.

Template	P@1	Arg-1
p1-basic (D)	0.208	0.467
p2-support (D)	0.222	0.467
p3-undermine (D)	0.208	0.467
p4-connect (D)	**0.278**	**0.533**
p5-auto (D)	0.194	0.433
p5-auto-connect (D)	0.264	0.467
p1-basic (DE)	0.208	0.467
p2-support (DE)	0.194	0.433
p3-undermine (DE)	0.208	0.467
p4-connect (DE)	**0.292**	**0.600**
p5-auto (DE)	0.194	0.433
p5-auto-connect (DE)	0.278	0.500
p1-basic (DL)	0.208	0.467
p2-support (DL)	0.222	0.500
p3-undermine (DL)	0.236	0.533
p4-connect (DL)	**0.361**	**0.700**
p5-auto (DL)	0.250	0.567
p5-auto-connect (DL)	0.306	0.633
p1-basic (DLE)	0.236	0.500
p2-support (DLE)	0.236	0.533
p3-undermine (DLE)	0.250	0.533
p4-connect (DLE)	**0.306**	**0.600**
p5-auto (DLE)	0.222	0.500
p5-auto-connect (DLE)	0.236	0.500
p1-basic (L)	0.181	0.367
p2-support (L)	0.208	0.433
p3-undermine (L)	0.236	0.500
p4-connect (L)	**0.278**	**0.567**
p5-auto (L)	0.194	0.400
p5-auto-connect (L)	0.236	0.433
p1-basic (LE)	0.208	0.467
p2-support (LE)	0.250	0.567
p3-undermine (LE)	0.222	0.500
p4-connect (LE)	0.236	0.500
p5-auto (LE)	0.222	0.500
p5-auto-connect (LE)	**0.264**	**0.600**

Table 9: **Argument Reconstruction:** Prompt tuning using LLaMA 2 (70B)

[11] https://github.com/Tiiiger/bert_score

D.3 Premise Evaluation on Correct/Incorrect Fallacy Classes

We closely examine the predictions in §5.3 and analyze the generated fallacious premises under two conditions: (a) when accompanied by a correct fallacy class, and (b) when not accompanied by an accurate fallacy class. The results are presented in Table 10. Specifically, for the metrics METEOR, BERTScore, and NLI-S, the fallacious premises produced by GPT 4 outperform those generated by LLaMA 2 when the fallacy class was correctly predicted. This differs strongly from the results when not separating correct from incorrect fallacy predictions. LLaMA 2 (*D*), the previously superior model in fallacious premise generation (from the main results in Table 2), maintains a good performance in this evaluation. In almost all cases, the performance of both LLMs is higher for correctly classified fallacies compared to incorrectly classified premises. This is likely attributed to the fact that we only have matching fallacious premises as reference text when the fallacy class is accurately predicted, but also indicates a certain consistency between the fallacy class and fallacious premise generation. While we believe it is important to report the performance as done here, we chose to report the overall performance in the main results in §5.3. This decision is made because (a) models may (and do, see §E.4) produce correct premises but assign the inaccurate fallacy class, and (b) to ensure comparability among models, as all scores are based on identical instances and not influenced by the model's specific fallacy classification performance.

D.4 Class-wise analysis of the fallacy classification

We examine the class-wise predictions of the best-performing models concerning the fallacies applied in the gold fallacious premises $\overline{p}_i$. The confusion matrices of LLaMA 2 (*LE*) and GPT 4 (*LE*) are presented in Figures 13 & 14. Both LLMs exhibit strong performances across the majority of fallacy classes. For GPT 4, the most confusion arises between *Ambiguity* and *False Equivalence*, as well as between *Biased Sample Fallacy* and *Hasty Generalization*. The commonalities between the former two fallacies align with the frequently co-occurring fallacies identified by our annotators in §4.3. Conversely, *Hasty Generalization*, although clear-cut in our annotations, is often used interchangeably with its superclass *Faulty Generalization* in the wild. We hypothesize that GPT 4 may have acquired knowledge during pretraining from discussions where claims generalizing from biased subsets are labeled as *Hasty Generalization*. Similarly, for LLaMA 2, most confusion primarily arises between *False Equivalence* and *Ambiguity*. However, we observe that LLaMA 2 tends to overpredict *False Equivalence* in general.

E Experiment Analysis

E.1 GPT 4 (D) Fallacy Detection Performance over k

We report the number of correctly detected fallacies per distinct fallacy class in Figure 15 for the GPT-4 (*D*) model with the best overall performance from §5.3. By allowing the model to provide k additional fallacy class predictions, the number of correctly identified fallacies increases, especially for the *Fallacy of Exclusion* and *False Equivalence*.

E.2 Fallacy-wise classification performance

We visualize the precision (top) and recall (bottom) from a multi-class multi-label perspective for each distinct fallacy class of LLaMA-2 and GPT-4 (both (*D*)) in Figure 16.

E.3 The impact of the fallacious premise

To better understand the impact of the generated fallacious premise $\overline{p}_i$, we additionally compare and evaluate the LLMs to see how well they can predict a correct fallacy class given only the inaccurate claim $\overline{c}$, the accurate premise p_0 and the context s_i, without generating the fallacious premise $\overline{p}_i$. We compare the results with the accuracy achieved in the fallacy reconstruction (Table 2), and when the fallacious premise is provided (Table 3). The results are depicted in Table 11. We report P@1 when the gold $\overline{p}_i$ is not provided to the model (i.e. for *Reconstruct* and *n/a*), since each of the interchangeable fallacies is equally valid. When $\overline{p}_i$ is provided (*Given*), we report the stricter accuracy metric in which only the one applied fallacy is correct. Interestingly, LLaMA 2 consistently performs better when it is not required to generate the fallacious reasoning for the fallacy classification task. We hypothesize that this is due to the poor utility of the fallacious premises generated by LLaMA 2 (cf. manual analysis in §E.4). For GPT 4, which produced substantially better fallacious premises according to our human evaluation,

Model	*Correct Fallacy Class* METEOR	BERTScore	NLI-A	NLI-S	*Incorrect Fallacy Class* METEOR	BERTScore	NLI-A	NLI-S
LLaMA 2 (*D*)	0.214	0.616	0.181	0.231	0.224	0.617	0.106	0.124
LLaMA 2 (*DE*)	0.238	0.615	0.144	0.174	0.227	0.622	0.119	0.142
LLaMA 2 (*DL*)	0.218	0.608	0.105	0.120	0.199	0.618	0.136	**0.148**
LLaMA 2 (*DLE*)	0.236	0.624	0.130	0.154	0.199	0.614	**0.128**	0.143
LLaMA 2 (*L*)	0.243	0.622	**0.239**	**0.266**	**0.256**	**0.628**	0.116	0.141
LLaMA 2 (*LE*)	0.194	0.605	0.149	0.168	0.177	0.610	0.114	0.125
GPT 4 (*D*)	0.259	**0.636**	0.094	0.200	0.229	0.611	0.057	0.092
GPT 4 (*L*)	**0.264**	**0.637**	0.109	**0.267**	0.227	0.604	0.046	0.088

Table 10: **Fallacious Premise Evaluation based on Fallacy Class Correctness:** Automatic metrics separately evaluated over fallacious premises when the predicted fallacy class was correct or not.

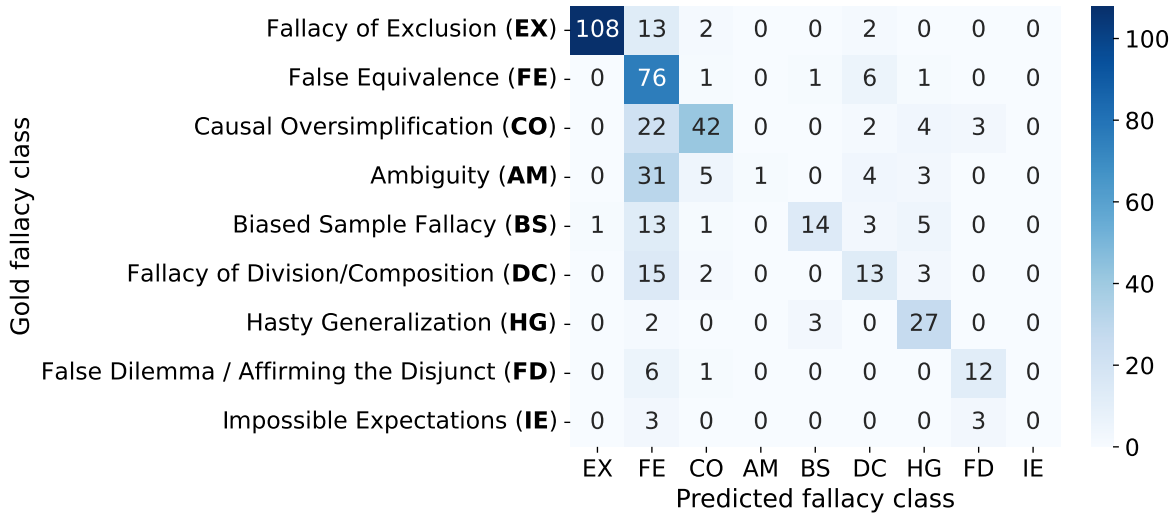

Figure 13: **Fallacy Classification for LLaMA 2** Confusion matrix based on LLaMA 2 (*Logical* + *Example*) of predicted and gold fallacy classes when provided with p_0, s_i, $\overline{p_i}$ and $\overline{c}$.

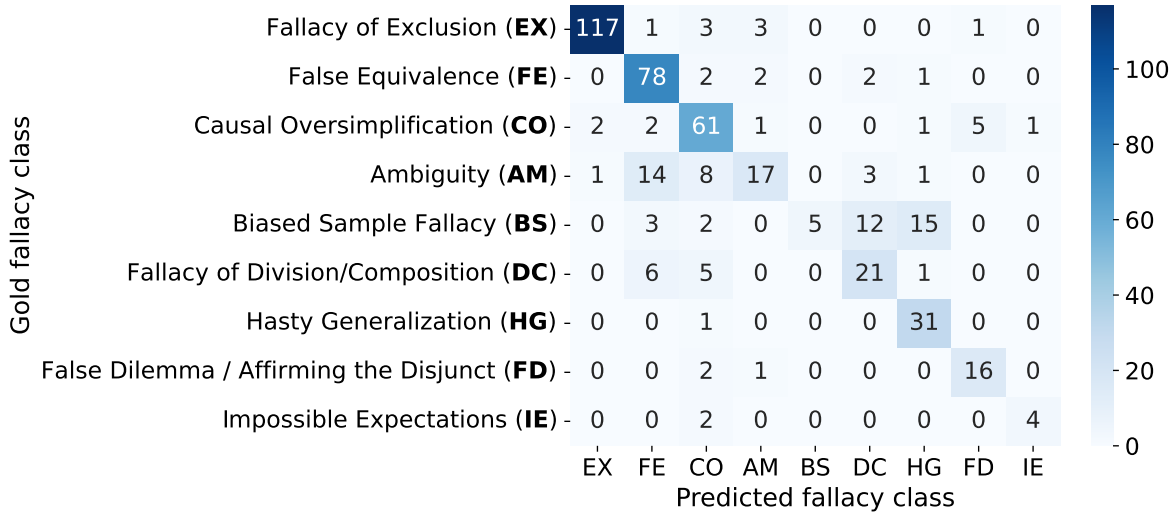

Figure 14: **Fallacy Classification for GPT 4** Confusion matrix based on GPT 4 (*Logical* + *Example*) of predicted and gold fallacy classes when provided with p_0, s_i, $\overline{p_i}$ and $\overline{c}$.

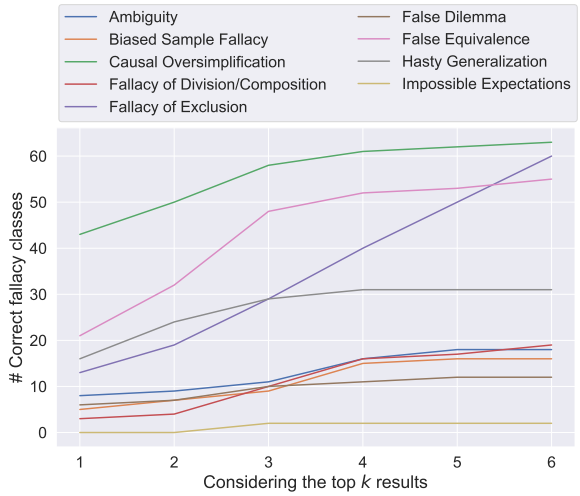

Figure 15: Number of correctly predicted fallacy classes of the GPT 4 (*D*) model when considering all correctly predicted fallacies among the top k predictions.

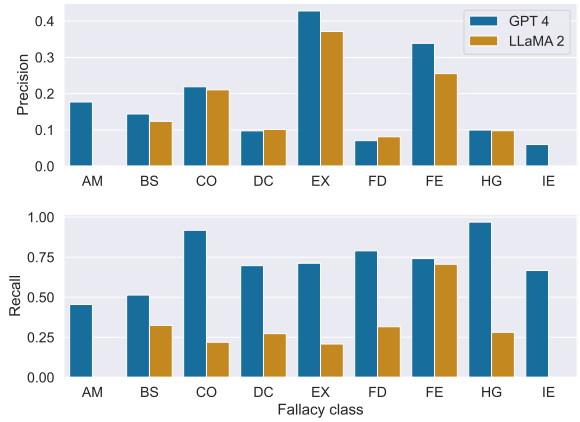

Figure 16: **Per-Fallacy Performance:** F1 score per predicted fallacy class from a multi-label multi-class classification perspective considering all predictions by the respective model.

Model	*Role of* $\overline{p}_i$ Reconstruct	Given	n/a
LLaMA 2 (*D*)	0.223	0.577	0.248
LLaMA 2 (*DE*)	0.209	0.637	0.264
LLaMA 2 (*DL*)	0.196	0.630	0.237
LLaMA 2 (*DLE*)	0.209	0.568	0.259
LLaMA 2 (*LE*)	0.212	0.645	**0.267**
LLaMA 2 (*L*)	0.193	0.601	0.262
GPT 4 (*D*)	**0.317**	0.738	**0.267**
GPT 4 (*L*)	0.292	**0.744**	0.245

Table 11: **Fallacy classification without fallacious premises:** Performance when predicting the applied fallacy class only, when required to *reconstruct* the fallacious premise, when the gold fallacious premise is *given* to the model, or when the fallacious premise is *n/a* and does not need to be generated.

Model	Setup	*Matching* $\hat{\overline{p}}_i$ Yes	No
GPT 4 (*D*)	classify f_i and gen. $\overline{p}_i$	**0.880**	0.229
	classify f_i w/o $\overline{p}_i$	0.640	0.114
	classify f_i given $\overline{p}_i$	0.788	**0.689**
GPT 4 (*L*)	classify f_i and gen. $\overline{p}_i$	**0.867**	0.133
	classify f_i w/o $\overline{p}_i$	0.533	0.133
	classify f_i given $\overline{p}_i$	0.732	**0.722**

Table 12: **Comparison to fallacy generation difficulty**: We separate the 60 instances from our manual analysis per GPT 4 by whether the generated fallacious premises apply the same fallacious reasoning as the reference gold fallacious premises (*Yes*) or not (*No*) and report the performance in three setups for both evaluated prompts.

the model exhibits improved performance, similar to chain-of-thought prompting (Wei et al., 2022), when tasked with generating the fallacious premise for both evaluated prompts.

To further shed light on this, In Table 12 we compare the performance of GPT 4 when (a) it was able to generate a fallacious premise that matched the reference premise according to human judgment, and (b) when it did not. We report the P@1 (or accuracy) for the GPT 4 models[12] when tasked to additionally generate the fallacious premise, only classify the fallacy, or classify the fallacy given the gold $\overline{p}_i$ in Table 11. The results show a substantial performance difference among both evaluated splits of instances, regardless of whether the fallacious premises needed to be generated or not, suggesting that these instances are challenging for both objectives. When provided with the gold reference $\overline{p}_i$, however, the difference in performance decreases, showing that the verbalized fallacious reasoning overcomes these challenges. For both prompts, the best fallacy classification performance was achieved when the model was tasked with additionally generating the fallacious premise and did so correctly. Note that we report P@1 if the fallacious premise $\overline{p}_i$ is not provided, and accuracy otherwise.

E.4 Manual Evaluation of Generated Fallacies

For the first ranked generated fallacious premise $\hat{\overline{p}}_i$ and predicted fallacy class $\hat{f}_i$ of 240 model predictions we assess if

1. the LLM outputs a fallacious premise (**Q1**),

[12]We only report GPT 4 results here because we found too few instances in which LLaMA 2 generated a matching fallacious premise in §E.4.

2. the generated premise $\hat{\overline{p}}_i$ represents an applicable premise within the argument bridges the reasoning gap based on the context provided via s_i (if exists) (**Q2**),
3. Q2 and the generated premise $\hat{\overline{p}}_i$ applies the predicted fallacy class $\hat{f}_i$ (**Q3**),
4. the generated premise $\hat{\overline{p}}_i$ expresses the same fallacious reasoning as the reference $\overline{p}_i$ (**Q4**).

We answer each question with yes/no, without awareness of the model or prompt for each prediction. Only in a single instance, the LLM failed to generate any fallacious premise (Q1). We provide the results of our manual analysis (Q2 and Q3) in Table 14. Most differences can be seen among both LLMs. Premises generated when a gold fallacy was assigned were considered valid slightly more often compared to when no gold fallacy was predicted. When the gold fallacy was predicted, the predicted fallacy class almost always matched the fallacy of the generated premise. However, when no gold fallacy class was predicted, often the predicted fallacy class did also not match the generated premise. Since we used stratified sampling when selecting the predictions, we approximate the overall performance s:

$$s = \text{P@1} \times h^{correct} + (1 - \text{P@1}) \times h^{incorrect}$$

where $h^{correct}$ (or $h^{incorrect}$) represent the manual evaluation among all instances considered correct (or incorrect) by the automatic evaluation via P@1. Examples are provided in Table 15. The first example contains relatively similar premises generated by the LLM and the annotators. In the second example, the LLM provides the same fallacious reasoning that applies results from mice to humans, yet the premise is much more specific and tailored to the provided content of the misrepresented publication. However, the human premise does not entail the LLM-generated premises. The third example exhibits semantically similar premises. However, the LLM-generated premise makes a causal assumption, differing from the *Biased Sample Fallacy* employed in the human-written premise. In the last example, the LLM-generated premise looks similar but does not logically support the claim, which reasons that in order to stop COVID-19, we need higher temperatures (and hence climate change) since higher temperatures stop the spread of COVID-19.

We use Q4 to assess the quality of each applied metric in Table 13. We separately report each metric among premise pairs when the generated premises involves the same fallacious reasoning as the reference premise (score should be high), and when not (score should be low). As commonly done (Banerjee and Lavie, 2005; Zhang et al., 2020), we report Pearson correlation to human judgment. When computing the Pearson correlation, we convert "yes" and "no" into numerical values of 1 and 0. Unlike semantic similarity, we observed that deciding whether the same reasoning flaw is verbalized is more often a binary decision than a decision on a continuous scale.

Metric	*Matching premises* Yes	No	*Pearson r* r	P-value
METEOR	0.280	0.227	0.153	0.017*
BERTScore	0.638	0.622	0.092	0.158
NLI-A	0.155	0.140	0.022	0.736
NLI-S	0.317	0.150	0.209	0.001*

Table 13: **Human evaluation (Q4):** Comparison of the used metrics reported over 68 generated premises that match the reference premise and 172 premises that do not match the reference premise. We measure the correlation with the human judgment via Person *r* and mark significant results with an asterisk.

We found NLI-S to be most correlated with human judgment, while the asymmetric counterpart NLI-A exhibited the least correlation. This confirms our initial hypothesis that LLMs may produce more specific premises than our annotators (see examples in §E.4, Table 15), and vice versa, which should be accounted for by the metric.

F Reproducibility

F.1 Hyperparameters

For reproducibility, experiments on LLaMA 2 use a batch size of 1, zero temperature, and a beam size of 1, with a fixed random seed set to 1. LLaMA 2 experiments were implemented using the Huggingface transformers library (Wolf et al., 2020). To address computational constraints, we employ 4-bit quantization (Dettmers et al., 2022) for LLaMA 2 (70B), which we found comparable to 8-bit quantization in preliminary experiments. All prompting experiments wit LLaMA 2 models are executed on A100 GPUs. For GPT 4 we use the GPT version=`gpt-4` and the API version=`2023-10-01-preview` with a maximum new token length of 1000 and content filtering

Model	*Gold Fallacy Prediction* Plausible	Matching	*No Gold Fallacy Prediction* Plausible	Matching	*Overall* (approximated) Plausible	Matching
LLaMA 2 (*L*)	5 (17%)	5 (17%)	2 (7%)	1 (3%)	*0.167*	*0.040*
LLaMA 2 (*D*)	7 (23%)	7 (23%)	4 (13%)	3 (10%)	*0.233*	*0.107*
GPT 4 (*L*)	26 (87%)	26 (87%)	25 (83%)	11 (37%)	*0.867*	*0.503*
GPT 4 (*D*)	25 (83%)	24 (80%)	18 (60%)	10 (33%)	*0.674*	*0.481*

Table 14: Results of our manual analysis based on 240 ($30 \times 4 \times 2$) manually assessed model predictions. We report the number of generated premises deemed **plausible**, and the number of plausible fallacious premises in which the applied fallacy class **matches** the predicted fallacy class.

Matches	LLM	Annotator
yes	Since THC is a compound in marijuana and HU210 is more potent than THC, it follows that marijuana has the same brain cell growth properties as HU210.	HU210 and THC have similar properties. Since HU210 increases the number of brain cells, THC also causes brain cells to grow.
yes	Because HU210, a synthetic compound found in cannabis, increases the number of cells in the hippocampus of mice, the use of any cannabis-related substance, like marijuana, will yield the same results in humans.	Mice and humans are both mammals. Therefore, marijuana grows brain cells in humans.
no	The fluoride in the pineal gland comes from the fluoride in toothpaste.	Elderly people are a subset of all people. Therefore, fluoride in toothpaste harms the pineal gland in all humans.
no	Efforts to combat climate change will result in warmer average temperatures, therefore increasing the prevalence of the SARS-CoV-2 virus.	Lower SARS-CoV-2 transmission was associated with lower temperature. Therefore, lower temperature facilitates SARS-CoV-2 transmission.

Table 15: **Matching of Fallacious Premises:** Examples of the manual evaluation with fallacious premises produced by an LLM and by the annotators, together with our rating whether they express the same reasoning flaw or not.

turned off, using the OpenAI API[13].

F.2 Prompts

F.2.1 Prompts for Fallacious Argument Reconstruction

We include *definitions*, *logical forms* or *examples* from literature (Bennett, 2012; Cook et al., 2018) (cf. §B.1). Notably, examples do not constitute real-world fallacies, but constitute educational examples that clearly outline the fallacious reasoning behind them, such as "A feather is light. What is light cannot be dark. Therefore, a feather cannot be dark" (cf. Table 6). We assess different combinations thereof in various prompt templates outlined in Figure 17. Using this template, we evaluate various task descriptions (Figures 18-23; key differences are highlighted in **bold**) for the different combinations of definitions, logical forms and examples. We do not argue that these are the respective best prompts to solve the introduced task. Rather, they serve as useful baselines that for future research that and are identical for both evaluated LLMs to allow for direct comparison as well as the analysis of the impact of *definitions*, *logical forms* or *examples* within these prompts.

F.2.2 Prompt for Consistency

When evaluation the internal consistency of LLMs we use the prompt in Figure 24. As Premise 3 we insert the LLMs generated premise (Table 2) or the annotated gold fallacious premise (Table 3). We measuring the consistency of the LLM we always provide the same level of detail about the fallacies (*definition*, *logical form*, *example*) that were also available in the prompt when generating the fallacious premise.

F.2.3 Prompt for Fallacy Classification without Premise

We adapt the prompts from §F.2.1 to only instruct LLMs to classify the fallacy, but not generate the missing fallacious premise by adapting the task instructions as depicted in Figure 25.

F.3 Paper Writing

We used ChatGPT as assistant when condensing our paper content. We thoroughly reviewed and

[13] https://platform.openai.com/docs/api-reference

Fallacy Inventory:

Ambiguity:
Definition 1: *When an unclear phrase with multiple definitions is used within the argument; therefore, does not support the conclusion.*
Logical Form 1: *Claim X is made. Y is concluded based on an ambiguous understanding of X.*
Example 1: *It is said that we have a good understanding of our universe. Therefore, we know exactly how it began and exactly when.*
Definition 2: *When the same word (here used also for phrase) is used with two different meanings.*
Logical Form 2: *Term X is used to mean Y in the premise. Term X is used to mean Z in the conclusion.*
Example 2: *A feather is light. What is light cannot be dark. Therefore, a feather cannot be dark.*

Impossible Expectations:
Definition 1: *Comparing a realistic solution with an idealized one, and discounting or even dismissing the realistic solution as a result of comparing to a "perfect world" or impossible standard, ignoring the fact that improvements are often good enough reason.*
Logical Form 1: *X is what we have. Y is the perfect situation. Therefore, X is not good enough.*
Example 1: *Seat belts are a bad idea. People are still going to die in car crashes.*

`<more fallacies>`

Task:
`<task-instruction with instance>`

Figure 17: **Argument Reconstruction Template:** Overall template used to reconstruct fallacious arguments in which *definition*, *logical form* and *example* are used. We evaluate different combinations thereof, as well as different task descriptions.

Examine the following fallacious argument:
Premise 1: "<p_0>"
Premise 2: "<s_i>"
Premise 3: ""
Therefore: "`<claim>`"

Premises 1 and 2 are sourced from the same credible scientific document. The claim is based on the information in Premise 1. However, Premise 2 suggests that the claim is an invalid conclusion from the scientific document.

Your task is to identify and verbalize the fallacious reasoning in Premise 3 (the fallacious premise) that is necessary to support the claim, despite the conflicting information in Premise 2. Only consider fallacies from the provided fallacy inventory.

Present each fallacious premise along with the applied fallacy class in this format:

Fallacious Premise: <fallacious premise>; Applied Fallacy Class: <applied fallacy class>.

If there are multiple applicable fallacies, list them in order of relevance.

Figure 18: **P1 Basic:** Our most basic task instruction to reconstruct the fallacious argument.

adjusted the paraphrased text for accuracy.

Examine the following fallacious argument:
Premise 1: "<p_0>"
Premise 2: "<s_i>"
Premise 3: ""
Therefore: "<claim>"

Premises 1 and 2 are sourced from the same credible scientific document. The claim is based on the information in Premise 1. However, Premise 2 suggests that the claim is an invalid conclusion from the scientific document.

Your task is to identify and verbalize the fallacious reasoning in Premise 3 (the fallacious premise) that is necessary to support the claim, despite the conflicting information in Premise 2. **This reasoning should be strong enough to support the claim and counter any uncertainties raised by Premise 2.** Only consider fallacies from the provided fallacy inventory.

Present each fallacious premise along with the applied fallacy class in this format:

Fallacious Premise: <fallacious premise>; Applied Fallacy Class: <applied fallacy class>.

If there are multiple applicable fallacies, list them in order of relevance.

Figure 19: **P2 Support:** The model is tasked to produce a fallacious premise that increases the support behind the claim.

Examine the following fallacious argument:
Premise 1: "<p_0>"
Premise 2: "<s_i>"
Premise 3: ""
Therefore: "<claim>"

Premises 1 and 2 are sourced from the same credible scientific document. The claim is based on the information in Premise 1. However, Premise 2 suggests that the claim is an invalid conclusion from the scientific document.

Your task is to identify and verbalize the fallacious reasoning in Premise 3 (the fallacious premise) that is necessary to support the claim, despite the conflicting information in Premise 2. **This reasoning should effectively support the claim, ensuring that Premise 2 does not undermine the claim as a valid conclusion.** Only consider fallacies from the provided fallacy inventory.

Present each fallacious premise along with the applied fallacy class in this format:

Fallacious Premise: <fallacious premise>; Applied Fallacy Class: <applied fallacy class>.

If there are multiple applicable fallacies, list them in order of relevance.

Figure 20: **P3 Undermine:** The task is rather phrased negatively. The model must generate the fallacious premise to avoid that s_i undermines the claim.

Examine the following fallacious argument:
Premise 1: "<p_0>"
Premise 2: "<s_i>"
Premise 3: ""
Therefore: "<claim>"

Premises 1 and 2 are sourced from the same credible scientific document. The claim is based on the information in Premise 1. However, Premise 2 suggests that the claim is an invalid conclusion from the scientific document.

Your task is to identify and verbalize the fallacious reasoning in Premise 3 (the fallacious premise) that is necessary to support the claim, despite the conflicting information in Premise 2. Do not repeat the claim itself, Premise 1, or Premise 2 when generating the fallacious Premise 3. **Make sure the generated Premise 3 connects Premise 1 and Premise 2 to robustly support the claim, and ensure that Premise 2 does not undermine the claim as a valid conclusion.** Only consider fallacies from the provided fallacy inventory.

Present each fallacious premise along with the applied fallacy class in this format:

Fallacious Premise: <fallacious premise>; Applied Fallacy Class: <applied fallacy class>.

If there are multiple applicable fallacies, list them in order of relevance.

Figure 21: **P4 Connect:** The task definition explicitly requires the model to connect Premises 1 and 2 with the conclusion via the fallacious premise.

Carefully analyze the following fallacious argument:
Premise 1: "<p_0>"
Premise 2: "<s_i>"
Premise 3: ""
Therefore: "<claim>"

Both Premise 1 and Premise 2 originate from a reputable scientific document. The claim is deduced from information presented in Premise 1. However, Premise 2 introduces doubt, suggesting that the claim is an invalid conclusion based on the scientific document.

Your objective is to precisely identify and articulate the fallacious reasoning in Premise 3 (the fallacious premise). **This reasoning must robustly support the claim, ensuring that Premise 2 does not undermine the claim as a valid conclusion.** Consider only fallacies from the provided fallacy inventory. Present each fallacious premise alongside the applied fallacy class in this format:

Fallacious Premise: <fallacious premise>; Applied Fallacy Class: <applied fallacy class>.

If multiple fallacies are applicable, list them in order of relevance.

Figure 22: **P5 Auto:** We automatically optimized the *P2 Support* template by asking ChatGPT to improve the prompt for clarity and conciseness.

Carefully analyze the following fallacious argument:
Premise 1: "<p_0>"
Premise 2: "<s_i>"
Premise 3: ""
Therefore: "<claim>"

Both Premise 1 and Premise 2 originate from a reputable scientific document. The claim is deduced from information presented in Premise 1. However, Premise 2 introduces doubt, suggesting that the claim is an invalid conclusion based on the scientific document.

Your objective is to precisely identify and articulate the fallacious reasoning in Premise 3 (the fallacious premise). Do not repeat the claim itself, Premise 1, or Premise 2 when generating the fallacious Premise 3. **Make sure the generated Premise 3 connects Premise 1 and Premise 2 to robustly support the claim, and ensure that Premise 2 does not undermine the claim as a valid conclusion.** Present each fallacious premise alongside the applied fallacy class in this format:

Fallacious Premise: <fallacious premise>; Applied Fallacy Class: <applied fallacy class>.

If multiple fallacies are applicable, list them in order of relevance.

Figure 23: **P6 Auto-Connect:** An extension of *P5 Auto* that explicitly requires the model to connect Premises 1 and 2 with the conclusion via the generated fallacious premise (as in *P4 Connect*).

Given the following argument and <definitions with their logical forms and examples>, determine which of the fallacies defined below occurs in Premise 3 of the provided argument. The argument may contain multiple fallacies. Only detect the most fitting fallacy within Premise 3. Explain your decision and conclude with the applied fallacy in a separate line at the end as "Fallacy: <fallacy class>".

Fallacies:
<Fallacies with definition and/or logical form and/or example>

Argument:
Premise 1: "<p_0>"
Premise 2: "<s_i>"
Premise 3: "<$\overline{p}_i$>"
Therefore: "<claim>"

Figure 24: **Prompt Template for Consistency:** Template used to assess the consistency of the LLMs.

Task:
Examine the following fallacious argument:

Premise 1: "<p_0>"
Premise 2: "<s_i>"
Premise 3: ""
Therefore: "<claim>"

Premises 1 and 2 are sourced from the same credible scientific document. The claim is based on the information in Premise 1. However, Premise 2 suggests that the claim is an invalid conclusion from the scientific document.

A fallacy must be applied when connecting Premise 1 and Premise 2 to robustly support the claim. Your task is to identify the applied fallacy class. Only consider fallacies from the provided fallacy inventory.

Present each applied fallacy class in this format:

Fallacy: <applied fallacy class>.

If there are multiple applicable fallacies, list them in order of relevance.

Figure 25: **Fallacy classification only:** Task instructions when the LLM is only tasked to classify the applied fallacy (without generating the fallacious premise).