Grounding Fallacies Misrepresenting Scientific Publications in Evidence

Anonymous ACL submission

Abstract

Health-related misinformation claims often falsely cite a credible biomedical publication as evidence, which superficially appears to support the false claim. The publication does not really support the claim, but a reader could believe it thanks to logical fallacies. Here, we aim to detect and to highlight such fallacies, which requires assessing the exact content of the misrepresented publications. To achieve this, we introduce MISSCIPLUS, an extension of the fallacy detection dataset MISSCI. MISSCIPLUS extends MISSCI by grounding the applied fallacies in real-world passages from misrepresented studies. This creates a realistic test-bed for detecting and verbalizing fallacies under real-world input conditions, and enables new and realistic passage-retrieval tasks. MISSCIPLUS is the first logical fallacy dataset which pairs the real-world misrepresented evidence with incorrect claims, identical to the input to evidence-based fact-checking models. With MISSCIPLUS, we *i*) benchmark retrieval models in identifying passages that support claims only with fallacious reasoning, *ii)* evaluate how well LLMs verbalize fallacious reasoning based on misrepresented scientific passages, and *iii)* assess the effectiveness of fact-checking models in refuting claims that misrepresent biomedical research. Our findings show that current fact-checking models struggle to use misrepresented scientific passages to refute misinformation. Moreover, these passages can mislead LLMs into accepting false claims as true.[1]

1 Introduction

Health-related misinformation has caused significant harm in our society (Zarocostas, 2020). Human fact-checking (HFC), which is time-consuming, must prioritize the most impactful claims and struggles with the rapid spread of misinformation (Arnold, 2020; Vosoughi et al., 2018). Two main approaches can automatically combat health-related misinformation: Scientific automated fact-checking (AFC) retrieves scientific evidence documents to support or to refute a claim, but faces challenges with mismatched specificity between claim and evidence (Wadden et al., 2022), and the reliance on the availability of refuting evidence (Glockner et al., 2022). Reason-checking (Musi et al., 2023) rejects claims that use fallacious reasoning, which is particularly relevant since facts are often misrepresented or skewed (Brennen et al., 2020). Unlike most logical fallacy detection datasets, which assume that fallacious reasoning is explicitly stated (Jin et al., 2022; Alhindi et al., 2022), real-world scenarios often involve fallacies that are not explicitly articulated.

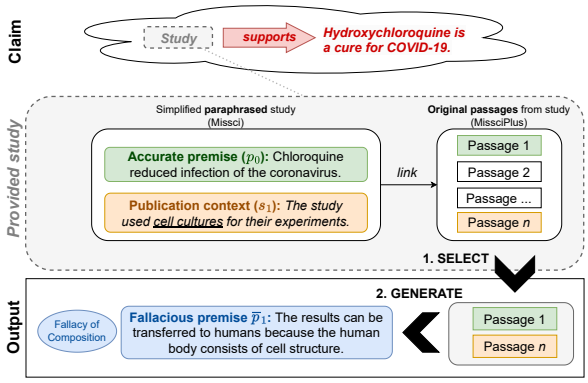

Figure 1: We link the paraphrased context from MISSCI to real-world passages. The LLM must (*i*) find relevant passages from the original study and (*ii*) generate a fallacious premise to (falsely) support the claim.

The MISSCI dataset (Glockner et al., 2024a) addresses implicit fallacious reasoning by reconstructing the *fallacious logical argument*, explicitly verbalizing the fallacies that led to the incorrect claim. However, MISSCI provides pre-selected and simplified short phrases instead of the actual misrepresented study as evidence. In a real-world scenario, models must first *(i)* identify the relevant

[1]Code and data are provided with this submission.

passages within the entire misrepresented study and *(ii)* verbalize the fallacious reasoning based on the original scientific text. To address this, we present MISSCIPLUS (§4), an extension of MISSCI with 2,257 human-annotated links between the simplified phrases from MISSCI and the real-world passages from the misrepresented publications.

Figure 1 shows (parts of) a fallacious logical argument from MISSCI for the claim that "*hydroxychloroquine is a cure for COVID-19*". Each claim in MISSCI has a kernel of truth which is anchored in the study's content ("*Chloroquine reduced infection of the coronavirus*"), denoted as the *accurate premise* in green. However, additional information from the study undermines the claim's validity and reveals a reasoning gap ("*The study used cell cultures for their experiments*"), denoted as *publication context* in orange. To bridge the reasoning gap between the study content and the claim, the model must verbalize a second premise ("*The results can be transferred to humans because the human body consists of cell structure.*"), denoted as *fallacious premise* in blue, and classify the fallacy associated with this fallacious premise ("*fallacy of composition*"). Unlike MISSCI, where the accurate premise and publication context are provided directly, MISSCIPLUS requires LLMs to retrieve and reason over real-world passages from the misrepresented study. MISSCIPLUS is the first fallacy dataset to pair real-world misinformation claims with publication passages as evidence. This setup is identical to the input used by evidence-based scientific AFC models and enables their evaluation over fallacious claims.

We leverage MISSCIPLUS to answer the following research questions:

1. How well can existing ranking approaches select the correct evidence passages (§5)?
2. How well can LLMs reconstruct fallacious arguments over original studies compared to paraphrased information in MISSCI (§6)?
3. Can evidence-based scientific AFC models use the content of misrepresented publications to detect the claims as misinformation (§7)?

Our findings suggest that lexical/semantic similarity based ranking models perform the best at identifying the required evidence passages. However, all models cannot leverage these passages to refute misinformation. In summary, we contribute MISSCIPLUS, an extension of MISSCI to ground fallacies in real-world evidence, which bridges the gap to AFC. We propose novel task definitions, along with extensive experiments on retrieving the required passages from the misrepresented study needed to detect misrepresented publications in the wild, as well as experiments using AFC models and LLMs in detecting science distortions.

2 Related Work

Fallacy detection research has primarily focused on surface-level fallacies (Habernal et al., 2017, 2018; Da San Martino et al., 2019; Sahai et al., 2021; Piskorski et al., 2023; Salman et al., 2023). Others extended these inventories to include logical fallacies that may require additional context for detection. However, all of this research relied on educational examples, fake news websites (Jin et al., 2022), or fact-checking articles (Musi et al., 2022; Alhindi et al., 2022) and assumed that the explicitly stated text was sufficient to detect the fallacies. MISSCI (Glockner et al., 2024a) developed models to verbalize the implicit fallacious reasoning. MISSCIPLUS differs from existing fallacy datasets by grounding implicit fallacies in real-world evidence documents.

Automated scientific fact-checking (Wadden et al., 2020; Saakyan et al., 2021; Sarrouti et al., 2021; Kotonya and Toni, 2020; Lu et al., 2023; Vladika and Matthes, 2023, 2024) used scientific documents as evidence to assess the veracity of claims. These approaches face challenges with fine-grained differences, such as specificity mismatches (Wadden et al., 2022). Our work bridges the gap between scientific AFC and fallacy detection and sheds light on the abilities of AFC models to reason over claims and misrepresented evidence passages.

Science communication research (Augenstein, 2021) compares claims and the cited evidence document across various dimensions such as claim strength (Li et al., 2017), certainty of the used language (Pei and Jurgens, 2021), sentence-level causal exaggerations (Yu et al., 2020), quantification of the information match (Wright et al., 2022), or combinations of multiple dimensions (Wührl et al., 2024). Our work stands apart by focusing on harmful misinformation and tasks that involve retrieving relevant passages and articulating the fallacious reasoning.

In order to investigate the antiviral properties of chloroquine on SARS-CoV after the initiation of infection, Vero E6 cells were infected with the virus and fresh medium supplemented with various concentrations of chloroquine was added immediately after virus adsorption. Infected cells were incubated for an additional 16-18 h, after which the presence of virus antigens was analyzed by indirect immunofluorescence analysis. When chloroquine was added after the initiation of infection, there was a dramatic dose-dependant decrease in the number of virus antigen-positive cells (Fig. 2A). As little as 0.1-1 muM chloroquine reduced the infection by 50% and up to 90-94% inhibition was observed with 33-100 muM concentrations (Fig. 2B). At concentrations of chloroquine in excess of 1 muM, only a small number of individual cells were initially infected, and the spread of the infection to adjacent cells was all but eliminated. A half-maximal inhibitory effect was estimated to occur at 4.4 ± 1.0 muM chloroquine (Fig. 2C). These data clearly show that addition of chloroquine can effectively reduce the establishment of infection and spread of SARS-CoV if the drug is added immediately following virus adsorption.

Figure 2: A real-world passage (Vincent et al., 2005) communicates the paraphrased content s_1 from MISSCI that *the study used cell cultures for their experiments*.

3 Background

Preliminaries MISSCI (Glockner et al., 2024a) comprises 208 health-related claims that misrepresent biomedical publications. Each claim is modeled as a fallacious logical argument where a claim is the conclusion of its premises. Formally, each argument comprises exactly one accurate premise (p_0) based on which the claim was made, and at least one fallacious premise $\overline{p}_i$ with the fallacy class f_i (cf. §A for details of all nine fallacy classes), that is needed to (falsely) conclude the claim $\overline{c}$ from the study's content. We refer to the pair of fallacious premise and the applied fallacy class ($\overline{p}_i$, f_i) as *fallacy*. Each fallacy in MISSCI is linked to one publication context (s_i) that points to a reasoning gap in the argument, and necessitates the fallacy. Each publication context (and accurate premise p_0) faithfully summarizes parts of the study. MISSCI defines the argument reconstruction task as: Given the publication context (s_i), the incorrect claim ($\overline{c}$) and the accurate premise (p_0), the model must verbalize the fallacious premise and detect the applied fallacy class ($\overline{p}_i$, f_i).

Linked Passages In MISSCI, the accurate premise (p_0) and publication contexts (s_i) were manually extracted from the HFC article and *not* the study itself. Since these HFC articles were specifically written to explain to non-experts why the misrepresented study does not support the claim, the paraphrased publication contexts often reveal the reasoning gaps easily. This severely limits the applicability of MISSCI in the real world, where models must identify relevant content from the misrepresented study and reason through complex scientific text. To address this, MISSCIPLUS links the paraphrased information from MISSCI with actual passages from the misrepresented study. Figure 2 shows a verbatim passage from the misrepresented study in MISSCIPLUS, which is linked (i.e., communicates the same content) to the paraphrased content that "*the study used cell cultures for their experiments*" from MISSCI (see Figure 1). Formally, given a misrepresented study S with its passages $S_j \in S$, we linked the passage S_j to the corresponding paraphrased information (p_0 and s_i) if (parts of) S_j entail the paraphrased information (p_0 or s_i). Any passage S_j linked to p_0 (or s_i) can replace p_0 (or s_i) during the argument reconstruction. The same passage may be linked to multiple paraphrased information and vice versa. We denote a passage S_j as S_j^0 if it links to p_0.

Subtasks We consider three sub-tasks for reconstructing fallacious arguments in the wild. First (§5), the model must retrieve a passage S_j^0, upon which the claim is based. This is crucial for understanding the general reasoning of the claim. Second (§5), the model must retrieve all additional passages S_j required to detect fallacies, i.e., passages linked to any publication context s_i. Lastly (§6), the argument reconstruction task is adapted from MISSCI, but replaces the paraphrased content with the respective linked passages. We always assume oracle input to establish a strong foundation for each sub-task, laying the groundwork for a more robust end-to-end system in the future.

4 Grounding MISSCI with Evidence

We selected all fallacious logical arguments from MISSCI, for which the misrepresented study is available via PMC[2]. This resulted in 118 arguments misrepresenting 100 distinct publications, which we automatically split into the constituent passages (cf. §B.1). We used the IMS model (Wright et al., 2022) to pre-select relevant passages and avoid exhaustively annotating every paraphrased information with every passage. The IMS model quantifies the information match between statements and publications, which aligns well with our needs. For each paraphrased information (p_0 and s_i), we selected the top-ranked passage according to IMS (cf. §B.2) and collected a minimum of six passages per argument in total, if possible. We employed two biology master's students with annotation experience in biomedical misinformation. The annotators

[2] https://www.ncbi.nlm.nih.gov/pmc/

	Model	*closed* (annotated) P@1	MRR	*open* (all) MRR
Baselines	*Random*	0.360	0.566	0.209
	Ordered	0.480	0.658	0.443
	BM25	0.547	0.705	0.539
Embeddings	BioBERT ST	0.547	0.712	0.582
	INSTRUCTOR	0.573	0.738	0.631
	IMS	0.587	0.742	**0.664**
AFC	AFC (SciFact)	0.603	0.748	0.535
	AFC (CovidFact)	0.517	0.691	0.450
	AFC (HealthVer)	0.608	0.765	0.516
	AFC (*all*)	0.608	0.768	0.514
LLMs	Llama2-70B	0.711	0.830	–
	Llama3-8B	0.729	**0.850**	–
	GPT-3.5	0.671	0.815	–
	GPT-4	**0.742**	**0.850**	–

Table 1: Finding a passage S_j^0 based on which the claim was made *(a)* among the annotated passages only (*closed*) or *(b)* across the entire publication (*open*). Averaged over five seeds (three for LLMs).

assigned an entailment label by comparing each paraphrased information (p_0 and s_i) with each selected passage (S_j), determining whether S_j entails (and hence is linked to) the paraphrased information. Following Glockner et al. (2024b), the annotators could express uncertainty if the entailment relation was ambiguous. If the paraphrased information was not linked to any selected passage, a corresponding passage was manually selected from the entire study. We removed four arguments for which no paraphrased information could be linked to any passage, yielding 2,257 double-annotated relations between passages and paraphrased information across 114 arguments. We retained the same instances for validation (30 arguments) and test (84 arguments) splits as in MISSCI. The inter-annotator agreement, measured by Cohen's κ, was 0.602. For details about the annotation process please refer to §B.3. Overall, 400 pieces of paraphrased information (88.6% of the accurate premises p_0; 72.0% of the publication contexts s_i; 76.8% overall) could be linked to at least one passage (analysed in §B.4). Descriptive statistics about the passages are in §C.

5 Retrieving Relevant Passages

The linked passages provide evidence why the claim was made, and why this involves fallacies. They are needed for the fallacious argument reconstruction and for effective debunking, which must explain why a claim was thought to be true and why it actually is not (Lewandowsky et al., 2020).

Finding the Kernel of Truth Given an incorrect claim $\overline{c}$ that misrepresents a publication $S = [S_0, S_1, ..., S_n]$, the model must rank all passages $S_i \in S$ such that the top-ranked passage communicates the accurate premise (p_0) based on which the claim $\overline{c}$ was made (denoted as S_j^0). This task is similar to finding supporting evidence in automated fact-checking (Thorne et al., 2018), but differs as the evidence passage is in a "corrupted" support relationship with the claim, meaning it only supports the claim when a fallacy is involved. The passage S_j^0 explains the basis of the claim and reveals the (broken) rationale behind the claim. We report P@1 and MRR over the subset of annotated passages with comprehensive annotations (*closed*) and over all passages (*open*). The open evaluation is more realistic, but only serves as a lower bound.

As *baselines*, we randomly shuffled passages (*random*) or maintained their original order from the publication (*ordered*). We used BM25 for lexical similarity-based ranking. As semantic *embedding* based approaches, we ranked the passages by their cosine similarity to the claim using sentence embeddings from BioBERT (Lee et al., 2020), fine-tuned by Deka et al. (2022) for evidence selection in scientific AFC, and prompt-based embeddings from INSTRUCTOR (Su et al., 2022). We also report the performance of the IMS (Wright et al., 2022) used during the dataset construction. We trained DeBERTaV3 (He et al., 2022) *AFC* models on three scientific AFC datasets SCIFACT (Wadden et al., 2020), COVIDFACT (Saakyan et al., 2021), HEALTHVER (Sarrouti et al., 2021), and their union (*all*) (cf. §D). Given the claim and the passage, we rank passages based on the predicted label probability for SUPPORTED, which fits closest to the task definition of finding a passage that seemingly supports the claim. We used LLMs only in the closed evaluation due to their computational cost. Here, we implemented PRP (Qin et al., 2023), which reorders passages through pairwise comparisons akin to the early iterations of the bubble sort. We used GPT-4 (OpenAI, 2023) and GPT-3.5[3] as proprietary LLMs, Llama3-8B[4] and Llama2-70B (Touvron et al., 2023) as open-source LLMs. Implementation details are outlined in §E.

Table 1 shows solid performance across all models. The strong *ordered* baseline suggests that claims often rely on early parts of a study. In the open evaluation, embedding-based approaches perform the best. Note that the IMS model preselected the passages for annotation, and its performance

[3] https://platform.openai.com/docs/models/
[4] https://llama.meta.com/llama3/

must be interpreted with caution. AFC models are superior to embedding models in the closed evaluation, but fall behind PRP ranking via LLMs.

Finding Undermining Passages Given an incorrect claim $\bar{c}$ that misrepresents a publication $S = [S_0, S_1, ..., S_n]$, and a passage $S_j^0 \in S$, based on which the claim was made, the model must rank all passages $S_i \in S$ such that the top-ranked passages S_i are linked to reasoning gaps (linked to the publication context s_i) between S and $\bar{c}$. This passage ranking task starkly differs from evidence retrieval in AFC because the model must reason about the relationships between the claim, why it was made (S_j^0), and the content of the passage. For example, "*in vitro experiments*" only indicate a fallacy because the claim relies on the results of these experiments and incorrectly transfers them to humans. This task further seeks content that reveals a fallacy by *disrupting* the reasoning, which is different from retrieval for multi-hop reasoning that connects multiple pieces of information to *establish* a reasoning path. We report the passage-level mean average precision (MAP) as our main metric. We further report P@1 because one detectable fallacy is the minimum requirement to reject a claim. To measure how many distinct reasoning gaps can be detected, we report the fallacy-level recall of the top ten ranked passages (Fall-R@10). This penalizes models that only detect passages linked to the same s_i. We only evaluate on the *open* subset, as most annotated passages are linked with reasoning gaps as per the dataset construction. We slightly adapted the AFC-based ranking to use the sum of the predicted probabilities for the labels SUPPORTED and REFUTED. We further assume oracle outputs from the previous subtask and prepended a randomly sampled gold S_j^0 to the claim in all baselines. These design choices follow our experiments on the validation split (cf. §F).

To solve this task, the model must first understand the (false) rationale behind the claim as expressed in the passage S_j^0 (e.g., communicating that "*chloroquine reduced infection of the coronavirus*"). Then, the model can identify how a different passage interacts with this reasoning to highlight a reasoning gap (e.g., that this was observed in "*in vitro*" experiments). Table 2 shows that cosine-similarity-based ranking outperforms all AFC-based models capable of jointly reasoning over claims and evidence, despite the complexity of the reasoning required. This suggests that lexical or semantic similarities correlate sufficiently strongly with passages that indicate reasoning gaps, while AFC models seem to fail to reason over them.

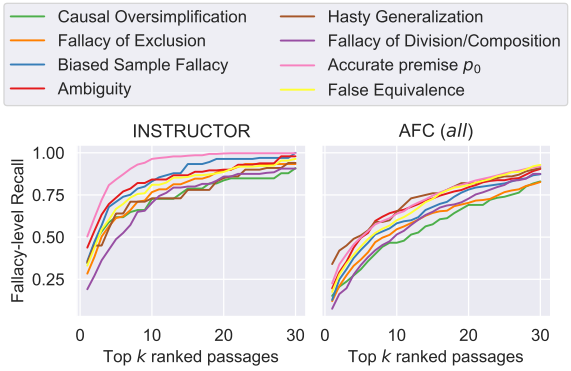

Figure 3: Recall of undermining passages per fallacy class (and accurate premise) over the top k ranked passages. We only list fallacies with ≥ 20 occurrences.

	Model	MAP	P@1	Fall-R@10
Baselines	*Random*	0.205	0.136	0.438
	Ordered	0.286	0.298	0.497
	BM25	0.496	0.617	0.602
Embedding	INSTRUCTOR	**0.541**	**0.652**	**0.613**
	SPICED-IMS	0.524	0.640	0.595
	BioBERT ST	0.491	0.600	0.570
AFC	AFC (SciFact)	0.360	0.326	0.518
	AFC (CovidFact)	0.380	0.457	0.538
	AFC (HealthVer)	0.368	0.410	0.544
	AFC (*all*)	0.306	0.338	0.554

Table 2: Retrieving passages linked to fallacies from the entire publication (*open*). Averaged over five seeds.

We show the recall of detectable fallacy classes over the number of retrieved passages for the INSTRUCTOR and AFC (all) in Figure 3. Overall, within the same model, different fallacy classes follow similar trends. The superior performance of INSTRUCTOR seemingly relies on fallacies that can be detected from S_j^0 passages. For comparison, we visualize the performance without S_j^0 passages in §F.2, Figure 9, which reduces the performance gap between the two ranking models.

6 Fallacious Argument Reconstruction

Given an incorrect claim $\bar{c}$ and relevant selected passages $S_j \in S$ of the misrepresented study, the model must verbalize the fallacious premises $\hat{\bar{p}}_i$ with the fallacy class $\hat{f}_i$ such that all reasoning gaps between S and $\bar{c}$ can be bridged via the generated fallacious premises $\hat{\bar{P}} = [\hat{\bar{p}}_1, \hat{\bar{p}}_2, ..., \hat{\bar{p}}_n]$. We assume oracle output of both passage retrieval tasks and only consider passages linked to a fallacy (with at least one S_j^0 passage as the basis for the claim).

Fallacy class: *Fallacy of Composition*
Definition: Inferring that something is true of the whole from the fact that it is true of some part of the whole
Logical Form: A is part of B. A has property X. Therefore, B has property X.
Example: Hydrogen is not wet. Oxygen is not wet. Therefore, water (H2O) is not wet.

Figure 4: Examples for the (**D**)efinition, (**L**)ogical form and (**E**)xample for the *Fallacy of Composition*, used as supplementary fallacy information in the prompts.

We experimented with GPT-4 Turbo and GPT-3.5 as proprietary LLMs , and Llama3-8B as a representative small open-source LLM with strong leaderboard performance. As in MISSCI, we prompted the LLMs in a zero-shot setting (cf. §G.1 for prompt selection). By default, all prompts include **D**efinition, **L**ogical Form and toy **E**xamples, as shown in Figure 4, from literature (Bennett, 2012; Cook et al., 2018) as supplementary information for each fallacy (cf. §A for all fallacy classes).

Holistic Argument Evaluation The evaluation on MISSCIPLUS faces two challenges: First, unlike MISSCI, no one-to-one mapping exists between the reasoning gaps and passages, since a passage can link to multiple reasoning gaps (s_i) and vice versa. Second, different reasoning gaps that necessitate distinct fallacious reasoning of the same class can be present. For example, the claim that "*hydroxychloroquine is a cure for COVID-19*" in Figure 1 falsely assumes that *i*) hydroxychloroquine will have an equal effect as chloroquine, and that *ii*) SARS-CoV-2 will behave equally to SARS-CoV-1. Both use the *False Equivalence* fallacy, but address distinct gaps. To address both, we evaluated the fallacious argument holistically. Given *all* relevant passages of the misrepresented study, we expect up to five ranked fallacies ($\hat{\overline{p}}_i$, $\hat{f}_i$), where five is the maximum number of distinct reasoning gaps in MISSCI. Following Schlichtkrull et al. (2023), we automatically match the generated fallacies with the gold fallacies at the *argument* level (instead of evaluating fallacies per s_i as in MISSCI). We define a function $\phi : (y, \hat{y}) \rightarrow \{0, 1\}$ which discerns if the predicted fallacy $\hat{y}_i = (\hat{\overline{p}}_i, \hat{f}_i)$ and the gold fallacy $y_k = (\overline{p}_k, f_k)$ match based on two implementations:

- ϕ^{f} outputs 1 if the predicted fallacy class equals the gold fallacy class.
- $\phi^{\text{f+p}}$ outputs 1 if the generated fallacious premise additionally bridges the same reasoning gap as the gold fallacious premise.

The ϕ^{f} is an upper bound, as it does not penalize models for poorly phrased premises. $\phi^{\text{f+p}}$ uses a Llama3-8B model fine-tuned with QLoRA adapters (Dettmers et al., 2023) on the human evaluation data from MISSCI (cf. §G.2). The accuracy via cross-validation is 79.8% (78.8 in F1-macro). We do not perform a human evaluation, which is too complex given the many-to-many relationship between the predicted and the gold fallacies, and is hard to reproduce for future work. Instead, we report evaluation measures based on the two complementary implementations ϕ, which we deem adequate to answer our research questions. As a primary measure, we report the recall of reasoning gaps for which ϕ found a match among the five fallacy predictions (R@5). Following Glockner et al. (2024a), we use precision P@1, to check if the top-ranked fallacy is correct, and Arg@1, which considers an argument as successfully rejected if at least one of the predicted fallacies is correct.

Experiments We compare the performance when prompting LLMs based on each passage S_j individually (*per-passage*) and when including all concatenated passages in a single prompt (*all passages*). The *per-passage* prompts follow MISSCI, but replace the accurate premise (p_0) and publication context (s_i) with the respective real-world passage. We always provided oracle passages and selected the top five ranked fallacy predictions per argument for evaluation using $\phi^{\text{f+p}}$ in Table 4.

Per-passage prompting is superior for both LLMs, but more expensive due to multiple prompts per argument. The higher (or similar) precision suggests that focusing on one passage (per-passage prompting) can be advantageous when identifying the fallacies. Llama3-8B outperforms GPT-3.5 across all measures, but detects none of the annotated fallacies for more than half of the misinformation, according to Arg@1.

Table 3 compares the performance of all three LLMs and ablations over various fallacy information in the prompts (**D**efinition, **L**ogical Form and **E**xplanation) on MISSCI and MISSCIPLUS using per-passage prompting (using the paraphrased s_i and p_0 on MISSCI). GPT-4 performs best across all measures and datasets on this task. We further observe a considerable impact of different fallacy information for each LLM. Yet, none is universally beneficial or harmful for all tested LLMs.

A clear trend of decreasing performance from the paraphrased information in MISSCI to the real-

LLM	Info	MISSCI R@5 (ϕ^{f+p})	MISSCI R@5 (ϕ^{f})	MISSCI Arg@1 (ϕ^{f+p})	MISSCIPLUS R@5 (ϕ^{f+p})	MISSCIPLUS R@5 (ϕ^{f})	MISSCIPLUS Arg@1 (ϕ^{f+p})
Llama3-8B	DLE	**0.277**	**0.514**	**0.552**	**0.226**	**0.477**	**0.476**
	DL	0.241	0.445	0.512	0.195	0.463	0.413
	DE	0.227	0.470	0.480	0.174	0.449	0.389
	LE	0.255	0.469	0.504	0.209	0.439	0.460
GPT-3.5	DLE	0.248	0.491	0.512	**0.165**	**0.428**	**0.361**
	DL	0.232	0.492	0.464	0.146	0.416	0.321
	DE	**0.276**	**0.517**	**0.567**	0.160	0.400	0.333
	LE	0.249	0.478	0.524	0.157	0.410	0.341
GPT-4 Turbo	DLE	**0.332**	0.486	**0.619**	0.224	0.458	0.452
	DL	0.308	0.500	0.583	0.238	0.495	0.488
	DE	0.318	**0.528**	0.595	0.210	0.491	0.440
	LE	0.304	0.505	0.583	**0.252**	**0.519**	**0.500**

Table 3: Argument reconstruction performance using paraphrased information from MISSCI compared to the real passages from MISSCIPLUS across various fallacy information. Results (expect GPT-4) are averaged over 3 seeds. We evaluate different combinations of fallacy (**D**)efinition, (**L**)ogical form and (**E**)xample in the prompt.

LLM	Passages	R@5	P@1	Arg@1
Llama3-8B	per-passage	**0.226**	**0.290**	**0.476**
Llama3-8B	all passages	0.199	0.266	0.425
GPT-3.5	per-passage	0.165	0.190	0.361
GPT-3.5	all passages	0.089	0.194	0.206

Table 4: *All passage* prompting vs. *per passage* prompting. Averaged over three seeds.

Training Data	*Argument level* Sup.	Ref.	Mix.	NEI	*AFC* Acc.
SciFact	55.5	4.5	23.3	16.7	88.9
HealthVer	39.0	**20.0**	27.1	13.8	82.1
CovidFact	**36.9**	12.1	**51.0**	–	90.7
All AFC	48.8	11.9	27.1	12.1	84.1

Table 5: Veracity predictions from scientific AFC models on 84 misinformation claims with 510 evidence passages in MISSCIPLUS. Averaged over five seeds.

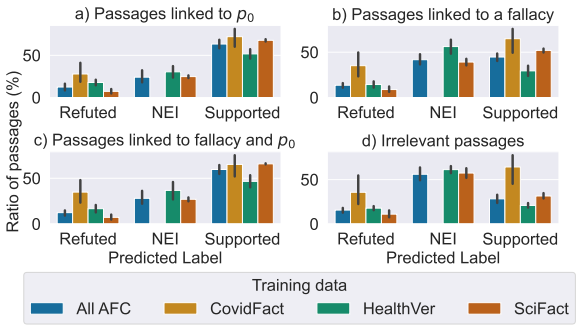

Figure 5: AFC predictions over passages linked to the accurate premise (*top left*), to reasoning gaps (*top right*), to both (*bottom left*) or none (*bottom right*).

world passages in MISSCIPLUS is evident. We note that, the approximate measures via ϕ^{f} and ϕ^{f+p} may underestimate performance as they cannot match valid fallacies not covered by the annotations (Glockner et al., 2024a). The performance from MISSCI to MISSCIPLUS only drops marginally for Llama3-8B and GPT-4 Turbo. Interestingly, GPT-3.5 outperforms Llama3-8B using the paraphrased content. However, its poor performance across all evaluated prompts on MISSCIPLUS corroborates the findings in Table 4 and suggests poor adaptability toward realistic scientific text on this task.

7 Scientific AFC Evaluation

A key novelty of MISSCIPLUS are the claims paired with real-world evidence passages that can be assessed by fallacy detection models *and* evidence-based fact-checking models. This allows to test AFC models on fallacious claims.

Fine-Tuned Models for Scientific AFC Given a claim $\bar{c}$ that misrepresents the publication S, we form n fact-checking instances ($\bar{c}$, S_j) by treating each of the n annotated passages $S_j \in S$, as evidence. We use the scientific AFC models from §5 that predict n veracity labels. Following Schlichtkrull et al. (2023), we assign an overall veracity label for a claim as MIXED if the labels SUPPORTED and REFUTED were predicted. If the model only predicted NOTENOUGHINFORMATION (NEI), the overall verdict was NEI. We label all other cases in which the model predicted SUPPORTED (with optional NEI) or REFUTED (with optional NEI) as SUPPORTED or REFUTED, respectively. The studies in MISSCIPLUS constitute trustworthy evidence that do not support the inaccurate claims – HFC have rated them to misrepresenting these studies. A literate scientific AFC model should equally detect the claims as misinformation.

Table 5 shows that all AFC models yielded high in-domain accuracy (82-90%) on the respective AFC dataset. Because MISSCI assumes that the

	LLM	Predicted as True	False	NEI	True	False	NEI
Know (*FC*)	Llama2	1.6	61.1	37.3	34.7	22.3	41.3
	Llama3	0.0	86.9	2.4	20.0	43.3	14.3
	GPT4	0.0	85.3	14.7	59.0	23.0	17.3
	GPT3.5	0.8	71.0	28.2	46.7	17.3	35.7
Know (*Ask*)	Llama2	0.0	100.0	0.0	29.7	69.3	1.0
	Llama3	8.3	88.9	2.4	68.7	26.0	3.0
	GPT4	3.6	68.3	27.8	49.7	6.7	36.3
	GPT3.5	1.6	50.4	48.0	47.3	6.0	45.0
RAG	Llama2	23.8	61.5	12.7	58.7	29.7	10.7
	Llama3	44.4	53.2	2.4	80.3	16.3	3.3
	GPT4	27.4	34.1	38.5	55.0	4.0	41.0
	GPT3.5	38.9	31.7	29.0	78.0	5.3	16.0
		Misinformation			**True claims**		

Table 6: Averaged veracity predictions from LLMs on misinformation from MISSCIPLUS (*left*) and true claims from HEALTHVER and COVIDFACT (*right*).

study provides sufficient evidence *for* the claim, any prediction other than SUPPORTED can be considered as correctly rejecting the claim. Yet, AFC systems falsely predict 37-56% of the claims to be true. The seemingly best COVIDFACT model uses binary classification without NEI and benefits from a substantially increased chance to predict MIXED.

Grounding the verdict of AFC models is critical for their trustworthiness and often is a key part of the evaluation protocol (Thorne et al., 2018; Wadden et al., 2020; Glockner et al., 2024b). To understand whether the AFC models reject the claims based on the undermining evidence that indicates reasoning gaps, we visualize the AFC prediction over different passages in Figure 5. Intuitively, the models mostly predict SUPPORTED over passages based on which the claim was made (*top left*). If the same passage is additionally linked to a reasoning gap (*bottom left*), the predicted distribution does not change much, suggesting unawareness of the fallacy. Based on passages that are linked to a reasoning gap but not to the accurate premise (*top right*), the distribution changes towards NEI. This is similar to completely unrelated passages (*bottom right*), again suggesting unawareness of the scientific distortion. More analysis is in §H.1.

LLMs as Scientific AFC Models We prompted each LLM (*a*) without any evidence to test whether it already has parametric knowledge about the claims from pretraining (Magar and Schwartz, 2022) and (*b*) with the relevant evidence passages to test their reasoning capabilities in a *RAG* style, similarly to Manakul et al. (2023). For parametric knowledge, we let the LLM generate a fact-checking article (*FC*), which is closest to our application, but may bias the model due to the prevalence of misinformation in fact-checking articles. As a neutral prompt, we additionally *asked* the LLM to predict the veracity of the claim to the best of their knowledge. All prompts are listed in §I.

Table 6 shows the results on MISSCIPLUS and 100 randomly sampled correct claims from HEALTHVER and COVIDFACT (50 per dataset). Both datasets contain real-world health-related claims from the internet. The percentages may not add up to 100% if LLMs refuse to answer. All LLMs tend to have parametric knowledge about the veracity of many claims. When providing evidence, the ratio of claims predicted as TRUE increases not only for correct claims but also for misinformation, despite that the *misrepresented* evidence does not truly support the claims. This phenomenon is similar to overruling the internal knowledge with evidence (Wu et al., 2024) and suggests that the LLMs do not identify the reasoning gaps between the claim and the evidence. Similarly to humans, LLMs seem prone to misinterpreting scientific publications. Due to their high persuasiveness (Augenstein et al., 2023; El-Sayed et al., 2024), this can lead to disastrous problems when people rely on LLMs, even if LLMs transparently output the (misrepresented) used evidence. Akin to previous work on fallacy recognition (Jin et al., 2022; Alhindi et al., 2022; Glockner et al., 2024a), the prompts in this work focus on *which* fallacies apply instead of *whether* any applies, and are no panacea either: in preliminary experiments (cf. §H.2), LLMs found fallacies in 85-99% of the claims in MISSCIPLUS but also in 78-99% of the true claims. Overall, MISSCIPLUS unifies fallacy detection with AFC and provides a resource for studying how to handle fallacious arguments and true claims alike.

8 Conclusion

We introduced MISSCIPLUS, a new dataset for reconstructing fallacious arguments based on realistic evidence. We showed that existing ranking models and LLMs struggle to reconstruct fallacious arguments using the real-world evidence that reveals reasoning gaps. Moreover, fine-tuned AFC models and LLMs failed to refute claims in MISSCIPLUS when presented with misrepresented evidence, highlighting the dangers of persuasive LLMs. Future work may improve models using synthetic data generation approaches, or extend the task definition over multiple documents.

Limitations

MISSCIPLUS assumes that the claims are based on a single publication only and that each publication is inherently trustworthy, i.e., that the only error was done in the reasoning from the publication to the claim. In the real world, finding complementary credible evidence is critical. We observed that sometimes single passages are insufficient to ground each fallacy in evidence. Currently MISSCIPLUS does not address grounding fallacies in multi-modal content, cases that require evidence from multiple passages, or conclusions that can only be drawn from the analysis of the entire study. We leave these challenges for future research.

Our evaluation is based on automatic matching, which is inevitably imperfect. Models may detect fallacies that are valid but not contained in the original annotations in MISSCI. We compensate for this with two complementary matching strategies and experiments over different seeds and prompt variations to confirm the robustness of our observations. By focusing on recall, we further do not penalize models for predicting fallacies that are not within our annotations while still requiring the models to detect the most prominent ones as highlighted by the annotators based on the HFC articles in MISSCI.

Our results are reported over two representative state-of-the-art LLMs at the time of writing (Llama3 8B as an easy-to-run open-source LLM and GPT4 as a proprietary LLM), and our claims are bound to these models. We opted for extensive tuning of these models to establish strong baselines for the novel tasks rather than providing comprehensive comparisons across various LLMs.

While our focus lies in grounding the individual argument constituents in the real-world misrepresented study, which contains 2,257 annotated links and 400 argument constituents, MISSCIPLUS is only based on 114 fallacious arguments. This may lead to variance and biases in the experiments, which must be interpreted with caution. We note that creating high-quality fallacy datasets with complex fallacious arguments requires suitable, professionally fact-checked claims, for which data is scarce. Future work could explore synthetic data generation to help bridge this gap.

Ethics Statement

The research questions targeted in this work aim to improve the detection of claims that distort scientific publications, which is ethically uncritical. Ethical concerns are bound to cases in which the content of this study are used in unintended ways.

Dual Use False interpretations of health-related claims can have disastrous effect. Any output of models derived from MISSCIPLUS only serves research purposes to detect such misinformation, but under no circumstances must be considered accurate without consulting experts in the field. Our work poses dangers for dual use, particularly verbalizing the fallacious reasoning to draw incorrect conclusions from real-world studies. While generating (parts of) misinformation always poses a risk, it is unavoidable to build resilience against real-world misinformation, as demonstrated in previous work (Zellers et al., 2019; Huang et al., 2023; Alhindi et al., 2023; Glockner et al., 2024a).

Data Collection All publications used in MISSCIPLUS have been published by the respective authors and we did not anonymize their work. All publications used in MISSCIPLUS are openly available and are part of the public discourse; in fact, they have even been distorted by misinformation. Hence, similarly to other scientific corpora (Lehman et al., 2019; Lo et al., 2020) that rely on such publications as evidence, we did not ask for explicit permission from the authors of each study to use their work.

References

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

Tariq Alhindi, 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.

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.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2023. Factuality challenges in the era of large language models. *ArXiv preprint*, abs/2310.05189.

Bo Bennett. 2012. *Logically fallacious: the ultimate collection of over 300 logical fallacies (Academic Edition)*. eBookIt. com.

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

John Cook, Peter Ellerton, and David Kinkead. 2018. Deconstructing climate misinformation to identify reasoning errors. *Environmental Research Letters*, 13(2):024018.

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.

Pritam Deka, Anna Jurek-Loughrey, and Deepak P. 2022. Evidence extraction to validate medical claims in fake news detection. In *International Conference on Health Information Science*, pages 3–15. Springer.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit Optimizers via Block-wise Quantization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Seliem El-Sayed, Canfer Akbulut, Amanda McCroskery, Geoff Keeling, Zachary Kenton, Zaria Jalan, Nahema Marchal, Arianna Manzini, Toby Shevlane, Shannon Vallor, et al. 2024. A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI. *ArXiv preprint*, abs/2404.15058.

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.

Max Glockner, Yufang Hou, Preslav Nakov, and Iryna Gurevych. 2024a. MISSCI: Reconstructing Fallacies in Misrepresented Science. *ArXiv preprint*, abs/2406.03181.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024b. AmbiFC: Fact-Checking Ambiguous Claims with Evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

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. 2018. 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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *The Eleventh International Conference on Learning Representations*.

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.

Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking for Public Health Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring Which Medical Treatments Work from Reports of Clinical Trials. 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 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

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*.

Yingya Li, Jieke Zhang, and Bei Yu. 2017. An NLP Analysis of Exaggerated Claims in Science News. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 106–111, Copenhagen, Denmark. Association for Computational Linguistics.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-Alignment Pretraining for Biomedical Entity Representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

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. *ArXiv preprint*, abs/2305.13186.

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.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. 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):20563051221150407.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models

for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 Technical Report. *Preprint*, arXiv: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.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, et al. 2023. Large language models are effective text rankers with pairwise ranking prompting. *ArXiv preprint*, abs/2306.17563.

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.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. 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 2116–2129, Online. Association for Computational Linguistics.

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 Umar Salman, Asif Hanif, Shady Shehata, and Preslav Nakov. 2023. Detecting Propaganda Techniques in Code-Switched Social Media Text. *ArXiv preprint*, abs/2305.14534.

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 Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. 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*.

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.

Hongjin Su, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, Tao Yu, et al. 2022. One embedder, any task: Instruction-finetuned text embeddings. *ArXiv preprint*, abs/2212.09741.

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

Martin J Vincent, Eric Bergeron, Suzanne Benjannet, Bobbie R Erickson, Pierre E Rollin, Thomas G Ksiazek, Nabil G Seidah, and Stuart T Nichol. 2005. Chloroquine is a potent inhibitor of SARS coronavirus infection and spread. *Virology journal*, 2:1–10.

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.

Juraj Vladika and Florian Matthes. 2024. Comparing knowledge sources for open-domain scientific claim verification. *ArXiv preprint*, abs/2402.02844.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):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.

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.

Kevin Wu, Eric Wu, and James Zou. 2024. How faithful are RAG models? Quantifying the tug-of-war between RAG and LLMs' internal prior. *ArXiv preprint*, abs/2404.10198.

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.

John Zarocostas. 2020. How to fight an infodemic. *The Lancet*, 395(10225):676.

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.

A Fallacies in MISSCIPLUS

We provide the definitions, logical form, and examples from literature taken from Glockner et al. (2024a) in Tables 7-8. MISSCI aggregates the fallacies into nine different classes:

1. **Ambiguity:** Combines *Ambiguity* and *Equivocation*.
2. **Impossible Expectations**: Uses *Impossible Expectations* and its alternative names.
3. **False Equivalence**: Uses *False Equivalence*.
4. **False Dilemma**: Uses *False Dilemma*.
5. **Biased Sample Fallacy**: Uses *Biased Sample Fallacy*.
6. **Hasty Generalization**: Uses *Hasty Generalization*.
7. **Causal Simplification**: Combines *False Cause Fallacy* and *Single Cause Fallacy*.
8. **Fallacy of Composition**: Combines *Fallacy of Composition* and *Fallacy of Division*.
9. **Fallacy of Exclusion**: Uses *Fallacy of Exclusion* and its alternative names.

B Dataset Construction

B.1 Passage Extraction

When documents are available in full-text via the PMC API[5], we retrieve the XML document and collect text passages enclosed within <p> tags within the <body> node. If the full-text document is accessible as HTML but not through the API, we extract the HTML content from the Wayback Machine[6] and separate the passages based on HTML structure. We make the data collection script publicly available.

B.2 Passage Selection

We segment each passage into constituent sentences using SciSpacy (Neumann et al., 2019) and compute the passage-level IMS (Wright et al., 2022) by max-pooling the IMS of all sentences. We retain the passage with the highest max-pooled IMS for each argument paraphrased information that needs a match (p_0, s_i). More passages (if available) are selected based on the highest IMS until we have six passages per argument. The IMS is particularly well-suited for our problem as it quantifies the information of the scientific findings rather than focusing on semantic similarity

B.3 Annotation

We use Surge AI[7] as the annotation platform and employed two M.Sc. students in biology paid 12.26 EUR per hour. We provide a screenshot of the anno-

[5] See https://www.ncbi.nlm.nih.gov/books/NBK25499/#chapter4.EFetch

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

[7] https://www.surgehq.ai/

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 7: Fallacy Overview. Definition and logical form taken from Bennett (2012) and Cook et al. (2018). Table as provided by Glockner et al. (2024a).

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 8: Fallacy Examples (taken from Bennett (2012)). Table as provided by Glockner et al. (2024a).

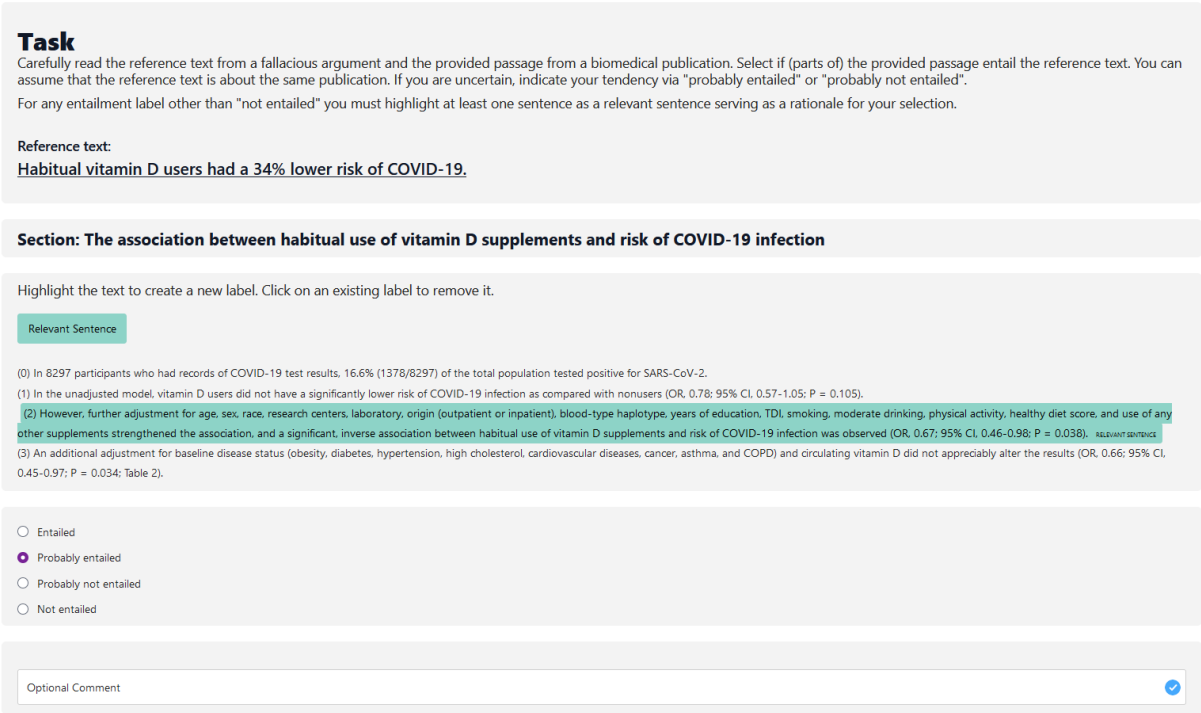

Figure 6: Annotation Interface from Surge AI to assign the entailment labels between the a passage and the paraphrased information.

Stage	Args	Passages	Relations	Mapped
1st Round	118	681	2,191	64.3%
2nd Round	118	719	2,334	67.9%
Consolid.	118	719	2,334	74.2%
Cleaned	114	694	2,257	76.8%

Table 9: Stages of the argument mapping annotation, including how many of the paraphrased information are linked to passages.

tation interface in Figure 6 and construct MISSCI-PLUS in multiple rounds. Table 9 provides statistics for each round of dataset construction.

Passage Linking First, every combination of paraphrased information (s_i or p_0) and selected passage S_j is double annotated by our annotators, deciding whether (parts of) the passage entail the paraphrased information. To account for cases without a clear entailment label, we follow Glockner et al. (2024b) and allow annotators to express their uncertainty. Specifically, annotators can choose one of the labels "entailed", "probably entailed", "probably not entailed" or "not entailed". We conservatively only consider the link between paraphrased information and passaged and as *entailed* if at least one of both annotators labeled it as "entailed" and the other labeled it as either "entailed" or "probably entailed". In the first round, we annotated 2,191 relations with 681 distinct passages. This annotation round linked 266 paraphrased content to at least one passage. In a second round, one annotator manually identified (if possible) a corresponding passage from all remaining passages for each missing link. These new passages were then double annotated with every paraphrased information (s_i, p_0) as outlined before. This annotation round contained 38 additional passages with 143 new relations and increased the number of paraphrased information linked to a passage to 281. The agreement for the fine-grained labels is 0.445. When merging "entailed" with "probably entailed" and "not entailed" with "probably not entailed", the inter-annotator agreement rises to 0.602. This suggests that some of the disagreement stems from uncertainty regarding a definite label.

Consolidation To account for possible false negatives due to our conservative label aggregation, we consolidate annotations where the two annotators reached no unanimous label. We consider instances as not unanimous if they include instances labeled as "probably entailed" by both annotators or "entailed" by one and "probably not entailed" or "not entailed" by the other. To finalize the overall label, we provided the annotator with the entire fallacious argument during consolidation. This

#	Sentence
(1)	The study did not include a group that did not wear masks at all. (*negated*)
(2)	The experiments were done on concentrations that are different from concentrations found in patients or vaccinated people. (*scope*)
(3)	Chloroquine diphosphate and hydroxychloroquine sulfate show antiviral activity against MERS-CoV and SARS-CoV. (*multi-hop*)
(4)	The average amount of spike protein in the blood was about 30 to 40 picograms/mL after receiving the Moderna vaccine. (*multi-modal*)

Table 10: Examples of paraphrased information with no linked passage.

additional context allowed for a more contextual understanding of the role of the passage in reconstructing the fallacious argument. Our primary annotator, who had extensive experience in fallacy annotation and was involved in all our pilot studies, handled consolidation. Cases in which our consolidator could not clearly identify as "entailed" were labeled as "not entailed". In total, 245 relations needed consolidation, 58 of which were consolidated as "entailed", leading to 26 previously unlinked paraphrased information. Four arguments were not linked to any passages and were subsequently removed. This likely happened when the link to the misrepresented publication was falsely selected in MISSCI.

B.4 Missing Passage Link Analysis

During dataset construction, we assumed that each paraphrased information from MISSCI could be linked to a single passage. This was feasible in 400 cases only (76.8%). Accurate premises were more frequently linked to at least one passage (88.6%), than publication contexts (72.0%). We list representative reasons where no passages could not be found in Table 10. The most common reason, accounting for 41.6%, was the presence of negation in the paraphrased information, discussing information *not* present in the study. For instance, a claim questioning mask effectiveness would require a study with a control group without masks. However, if the study did not focus on mask effectiveness in general, there is no need for such a control group or to explicitly mention its absence. Negated sentences were more prevalent in undermining publication contexts (47.4%) and less frequent in accurate premises (7.7%). Among the remaining instances, we identified (2) scope mismatches between the claim and the study, (3) information spread across multiple passages, and (4) non-textual components like tables or figures.

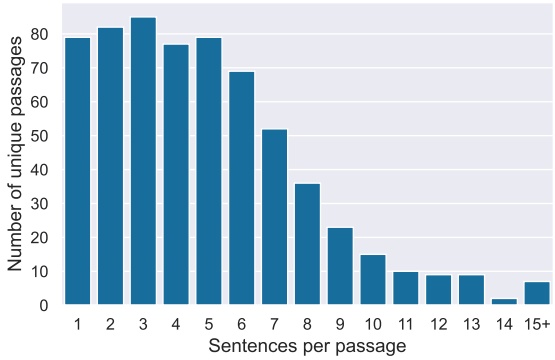

Figure 7: Number of sentences per passage over all *annotated* passages in MISSCIPLUS. Passages are only considered once (if used by multiple arguments).

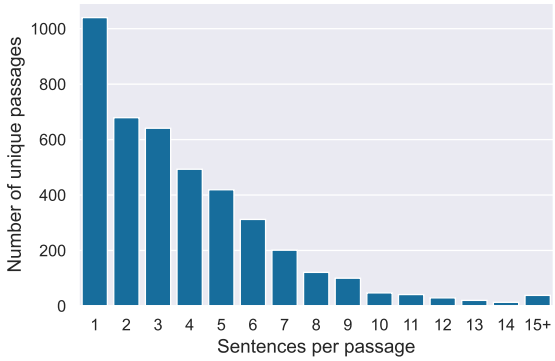

Figure 8: Number of sentences per passage over *all* passages of the entire publication in MISSCIPLUS. Passages are only considered once (if used by multiple arguments).

While these fields could not be linked to a single passage, it is theoretically possible to reconstruct the argument using the complete publication.

B.5 Licenses

MISSCI extends the MISSCI dataset, published under the Apache 2.0 license. MISSCIPLUS aligns directly with their intended use, which is to outline the fallacious reasoning of distorted science transparently, and only improved its applicability in the real world. Our collected annotations and all scripts to collect and preprocess the scientific publications will be made publicly available under an open-source license. All fallacious logical argument and all publications in MISSCIPLUS are in English.

C Length of Passages in MISSCIPLUS

The number of annotated passages per argument ranges from 1-8 passages (mean: 6.1; std: 1.2;

median: 6.0). The number of all annotations per argument ranges from 1-114 passages (mean: 43.7; std: 25.5; median: 43.0). Annotated passages vary in length, with 1 to 28 sentences. These passages contain 4.9 sentences on average (median: 4). When considering all passages of the entire misrepresented publication, each passage averages 3.8 sentences, with a median of 3. Figure 7 displays the distribution of sentence counts in annotated passages, while Figure 8 shows the same for all passages.

D Scientific AFC Models

We train four DeBERTaV3 (He et al., 2022) models for scientific AFC using SciFact (Wadden et al., 2020), CovidFact (Saakyan et al., 2021), HealthVer (Sarrouti et al., 2021) and the combination of all.

D.1 Scientific AFC datasets

SciFact (Wadden et al., 2020) covers multiple scientific domains using scientific abstracts as evidence documents. Citations were used to collect claims supported by the evidence. Refuted claims were generated via careful paraphrasing. While the original task includes the selection of evidence documents and rationale sentences, we only train models to predict the stance towards the claim given the full abstract as evidence. We used the official validation split as our test split (due to the hidden official test split). We randomly selected 200 instances from the official training set as the validation set and trained on the remaining instances. CovidFact (Saakyan et al., 2021) covers scientific claims related to COVID-19. Pairs of claims and evidence originate from a strictly moderated Reddit forum where every claim must provide an evidence document. Refuted claims were created by automatically changing the true claims. Unlike other AFC datasets, this dataset only distinguishes between "supported" and "refuted" claims, missing a label for "not enough information". HealthVer (Sarrouti et al., 2021) used COVID-19-related questions as queries and collected claims from resulting web pages via a search engine. Each claim was annotated against retrieved scientific abstracts as evidence using the three labels Supports, Refutes, and Neutral.

D.2 Hyperparameter search

We fix the number of epochs to 5 and the seed to 1 and perform a hyperparameter search over the learning rates (1e-7, 5e-7, 1e-6, 5e-6, 1e-5, 5e-5) and batch sizes (4, 8, 16), evaluated on the respective validation set. When training on multiple datasets, the overall performance considers each instance equally (instead of averaging dataset scores). The results are shown in Tables 11-14. We train five models using different seeds (1-5) based on the best-performing hyperparameters and report the averaged test performance in Table 15.

Model	lr	batch-size	Acc.	F1-score
deberta-v3-large	1e-07	4	49.5	24.7
deberta-v3-large	1e-07	8	49.5	24.7
deberta-v3-large	1e-07	16	49.5	24.7
deberta-v3-large	5e-07	4	50.5	26.2
deberta-v3-large	5e-07	8	49.5	24.7
deberta-v3-large	5e-07	16	49.0	22.1
deberta-v3-large	1e-06	4	58.5	43.7
deberta-v3-large	1e-06	8	53.0	31.6
deberta-v3-large	1e-06	16	50.0	22.2
deberta-v3-large	5e-06	4	90.5	88.6
deberta-v3-large	5e-06	8	88.5	86.4
deberta-v3-large	5e-06	16	82.5	79.6
deberta-v3-large	1e-05	4	**91.5**	**90.4**
deberta-v3-large	1e-05	8	**91.5**	90.2
deberta-v3-large	1e-05	16	88.5	86.8
deberta-v3-large	5e-05	4	50.0	22.2
deberta-v3-large	5e-05	8	50.0	22.2
deberta-v3-large	5e-05	16	82.5	80.7

Table 11: Hyperparameter search reported on our SciFact validation split. We fix the random seed to 1 and number of epochs to 5.

E Details on Finding the Kernel of Truth

E.1 Implementation Details

E.1.1 Baselines

As baselines, we randomly shuffle all passages (*random*) using five different seeds (1-5) or order passages as they appear within the original publication (*ordered*). We rank passages via BM25[8] for ranking based on lexical similarity. For preprocessing, we use SpaCy[9] for tokenization and converting all tokens to lowercase.

E.1.2 Embedding-based approaches

We experiment with various sentence-embedding models (in addition to those reported in Table 1). This includes SBERT[10] (Reimers and Gurevych, 2019), and sentence transformers fine-tuned on

[8] https://pypi.org/project/rank-bm25/
[9] https://spacy.io/
[10] all-mpnet-base-v2

Model	lr	batch-size	Acc.	F1-score
deberta-v3-large	1e-07	4	69.0	40.8
deberta-v3-large	1e-07	8	69.0	40.8
deberta-v3-large	1e-07	16	69.0	40.8
deberta-v3-large	5e-07	4	75.4	61.8
deberta-v3-large	5e-07	8	69.0	40.8
deberta-v3-large	5e-07	16	69.0	40.8
deberta-v3-large	1e-06	4	85.0	83.2
deberta-v3-large	1e-06	8	85.4	83.6
deberta-v3-large	1e-06	16	69.0	40.8
deberta-v3-large	5e-06	4	90.5	89.1
deberta-v3-large	5e-06	8	90.2	88.8
deberta-v3-large	5e-06	16	89.0	87.4
deberta-v3-large	1e-05	4	88.8	87.4
deberta-v3-large	1e-05	8	**90.9**	**89.3**
deberta-v3-large	1e-05	16	88.8	87.1
deberta-v3-large	5e-05	4	69.0	40.8
deberta-v3-large	5e-05	8	69.0	40.8
deberta-v3-large	5e-05	16	89.7	88.3

Table 12: Hyperparameter search reported on the COVIDFACT validation split. We fix the random seed to 1 and number of epochs to 5.

Model	lr	batch-size	Acc.	F1-score
deberta-v3-large	1e-07	4	51.6	29.4
deberta-v3-large	1e-07	8	49.8	29.3
deberta-v3-large	1e-07	16	46.7	30.1
deberta-v3-large	5e-07	4	72.3	65.6
deberta-v3-large	5e-07	8	66.2	50.2
deberta-v3-large	5e-07	16	64.8	47.5
deberta-v3-large	1e-06	4	80.1	78.2
deberta-v3-large	1e-06	8	76.5	72.0
deberta-v3-large	1e-06	16	69.1	56.0
deberta-v3-large	5e-06	4	**86.6**	**84.6**
deberta-v3-large	5e-06	8	85.4	83.4
deberta-v3-large	5e-06	16	86.3	84.5
deberta-v3-large	1e-05	4	85.6	83.6
deberta-v3-large	1e-05	8	85.7	83.4
deberta-v3-large	1e-05	16	85.8	83.6
deberta-v3-large	5e-05	4	51.8	22.7
deberta-v3-large	5e-05	8	51.8	22.7
deberta-v3-large	5e-05	16	51.8	22.7

Table 13: Hyperparameter search reported on the HEALTHVER validation split. We fix the random seed to 1 and number of epochs to 5.

Model	lr	batch-size	Acc.	F1-score
deberta-v3-large	1e-07	4	54.6	33.5
deberta-v3-large	1e-07	8	52.9	28.1
deberta-v3-large	1e-07	16	51.7	27.9
deberta-v3-large	5e-07	4	69.6	62.5
deberta-v3-large	5e-07	8	65.1	46.6
deberta-v3-large	5e-07	16	64.0	46.0
deberta-v3-large	1e-06	4	79.4	78.0
deberta-v3-large	1e-06	8	76.5	73.6
deberta-v3-large	1e-06	16	69.7	59.7
deberta-v3-large	5e-06	4	**85.2**	83.6
deberta-v3-large	5e-06	8	84.7	83.1
deberta-v3-large	5e-06	16	84.9	83.6
deberta-v3-large	1e-05	4	52.3	22.9
deberta-v3-large	1e-05	8	84.8	83.1
deberta-v3-large	1e-05	16	85.1	**83.8**
deberta-v3-large	5e-05	4	52.3	22.9
deberta-v3-large	5e-05	8	52.3	22.9
deberta-v3-large	5e-05	16	52.3	22.9

Table 14: Hyperparameter search reported on the SCIFACT, COVIDFACT, and HEALTHVER validation split. We fix the random seed to 1 and number of epochs to 5.

	Overall		*Per Label (F1)*		
Datasets	**Acc.**	**F1**	**S**	**N**	**R**
SCIFACT	88.9	87.9	92.0	86.0	85.8
HEALTHVER	82.1	81.4	80.1	86.6	77.4
COVIDFACT	90.7	89.5	86.0	–	93.1
All	84.1	83.3	83.8	86.7	79.5

Table 15: Evaluation of the scientific AFC models on the test set of the respective dataset used for training. The F1-score is macro averaged. The results are averaged over five seeds

Model	Agg.	Sent	MRR	P@1
BioBERT-ST	concat	all	0.657	0.462
BioBERT-ST	mean	1	0.676	**0.500**
BioBERT-ST	mean	2	0.664	**0.500**
BioBERT-ST	mean	3	**0.678**	**0.500**
PubMedBERT-ST	concat	all	**0.764**	**0.654**
PubMedBERT-ST	mean	1	0.620	0.423
PubMedBERT-ST	mean	2	0.661	0.462
PubMedBERT-ST	mean	3	0.664	0.462
SapBERT-ST	concat	all	**0.701**	**0.538**
SapBERT-ST	mean	1	0.664	0.500
SapBERT-ST	mean	2	0.680	0.462
SapBERT-ST	mean	3	0.663	0.462
SBERT	concat	all	**0.663**	**0.500**
SBERT	mean	1	0.620	0.423
SBERT	mean	2	0.655	0.462
SBERT	mean	3	0.602	0.385
SPICED IMS	concat	all	0.699	0.538
SPICED IMS	mean	1	0.598	0.346
SPICED IMS	mean	2	**0.717**	**0.577**
SPICED IMS	mean	3	0.673	0.500
INSTRUCTOR	concat	all	0.726	0.577
INSTRUCTOR	mean	1	0.671	0.500
INSTRUCTOR	mean	2	**0.763**	**0.654**
INSTRUCTOR	mean	3	0.747	0.615

Table 16: Hyper-parameter search on the validation split for sentence-embedding models to select passages $S_j^0 \Rightarrow p_0$.

BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2021), and SapBERT (Liu et al., 2021) as provided by Deka et al. (2022). For the INSTRUCTOR, we use the following prompts for the INSTRUCTOR (Su et al., 2022), which follows the official templates:

- **Prompt (Claim):** *"Represent the Scientific claim for retrieving supporting sentences: "*
- **Prompt (Passage Sentences):** *"Represent the Scientific sentence for retrieval: "*

For all embedding models, we ranked the passages according to the cosine similarity to the embedded claim. For SPICED-IMS, we use the model[11] provided by Wright et al. (2022) and rank passages according to their IMS between claim and passage. We treat the scientific AFC models from §D as cross-encoder models to jointly encode each claim-passage pair and re-rank passages based on the predicted probability for the label SUPPORTED.

Hyperparameter Tuning We compare two perspectives on how to represent passages. First, the model embeds the entire passage by concatenating all sentences (denoted as *concat*). Second, we compute the cosine-similarity (or IMS) between the claim and each sentence of the passage individually and rerank passages based on the mean score of the top k sentences with the highest score (denoted as *mean*). The intuition is that only fractions of a passage may be relevant for the claim, and focusing on only parts of the passage can benefit the ranking performance. We report the performance on the validation set in Table 16 and select the best hyperparameters based on the P@1.

E.1.3 LLMs with PRP

For PRP, we use the claim and two passages as input to the LLM and prompt it to output which passage should be ranked first. Following Qin et al. (2023), we use these outputs to re-rank passages akin to bubble sort. The bubble sort algorithm ensures that the top k elements are ranked at the top after k iterations. We evaluate four prompts on the validation split in Table 17, starting with the prompt provided by Qin et al. (2023) and three deviations tailored towards our task. We further assess whether passages should include or exclude the title and found that including the title generally improves performance. During prompt selection in Table 17 we randomly shuffle all passages prior to reranking them to avoid the impact of the strong positional bias on the validation split (P@1: 0.577). Since our metrics are only sensitive to the top-ranked results, PRP can only be completed over a few iterations to reduce costs without affecting the performance much. Following our experiments on the validation split, we only ran PRP for three iterations on the test set. Prompts are listed in §I.1.

E.2 Exhaustive Results on MISSCIPLUS

Table 18 provides a list of all evaluated models, including additional sentence embedding models, tasked to find the passage S^0 based on which the claim was made. We additionally report HasPositives@3 (Shaar et al., 2020) as a more relaxed measure that allows an appropriate passage to be ranked within the top three results.

F Details on Finding Undermining Passages

We only evaluate finding undermining passages in the open subset. Hence, the results represent a lower bound of the system's performance. Per our hyper-parameter search (§F.1.1 and §F.1.2), we

[11] https://github.com/copenlu/scientific-information-change

Prompt	Section Title	P@1	*MRR* after iteration *i*			
			i=1	i=2	i=3	All
Qin et al. (2023)	*no*	0.628	0.750	0.778	0.783	0.783
Qin et al. (2023) (claim)	*no*	0.705	0.806	0.833	0.835	0.837
Qin et al. (2023) (convincing)	*no*	0.731	0.825	0.849	0.848	0.848
Qin et al. (2023) (support)	*no*	0.615	0.739	0.758	0.758	0.760
Qin et al. (2023)	*yes*	0.692	0.798	0.824	0.824	0.825
Qin et al. (2023) (claim)	*yes*	0.731	0.822	0.844	0.841	0.843
Qin et al. (2023) (convincing)	*yes*	**0.744**	**0.831**	**0.858**	**0.858**	**0.858**
Qin et al. (2023) (support)	*yes*	0.667	0.769	0.789	0.791	0.793

Table 17: Evaluation of different prompts for PRP with Llama2-70B (8bit quantization), with random initialization. Averaged over three seeds.

Model	Annotated Passages (*closed*)		All Passages *(open)*		
	P@1	MRR	P@1	MRR	HasPos@3
Random	0.360	0.566	0.096	0.209	0.213
Ordered	0.480	0.658	0.320	0.443	0.507
BM25	0.547	0.705	0.387	0.539	0.627
SBERT (Reimers and Gurevych, 2019)	0.400	0.631	0.280	0.460	0.547
PubMedBERT ST (Deka et al., 2022)	0.440	0.652	0.240	0.442	0.573
BioBERT ST (Deka et al., 2022)	0.547	0.712	0.427	0.582	0.680
SapBERT ST (Deka et al., 2022)	0.480	0.672	0.333	0.514	0.627
INSTRUCTOR (Su et al., 2022)	0.573	0.738	0.480	0.631	0.733
SPICED-IMS (Wright et al., 2022)	0.587	0.742	**0.533**	**0.664**	**0.760**
DeBERTa$_{v3}$ SciFact (Wadden et al., 2020)	0.603	0.748	0.389	0.535	0.627
DeBERTa$_{v3}$ CovidFact (Saakyan et al., 2021)	0.517	0.691	0.307	0.450	0.507
DeBERTa$_{v3}$ HealthVer (Sarrouti et al., 2021)	0.608	0.765	0.347	0.516	0.629
DeBERTa$_{v3}$ Scienfic AFC (*all*)	0.608	0.768	0.349	0.514	0.600
Llama2-70B (Touvron et al., 2023) PRP (it=3)	0.711	0.830	–	–	–
Llama3-8B PRP (it=3)	0.729	**0.850**	–	–	–
GPT3.5 PRP (it=3)	0.671	0.815	–	–	–
GPT4 (OpenAI, 2023) PRP (it=3)	**0.742**	**0.850**	–	–	–

Table 18: Ranking performance to find the passages based on which the claim was made (S_j^0) over all 75 test instances where p_0 was linked to at least one passage.

Train data	Strategy	*mAP when S_j^0 included as* No S_j^0	Claim	Evid.
SciFact	*Support*	0.304	0.389	0.193
SciFact	*Refute*	0.258	0.251	0.230
SciFact	*Both*	0.348	**0.404**	0.208
CovidFact	*Support*	0.236	**0.369**	0.206
CovidFact	*Refute*	0.161	0.150	0.214
HealthVer	*Support*	0.270	0.357	0.235
HealthVer	*Refute*	0.234	0.269	0.241
HealthVer	*Both*	0.273	**0.375**	0.245
All AFC	*Support*	0.308	0.410	0.188
All AFC	*Refute*	0.270	0.183	0.192
All AFC	*Both*	0.355	**0.420**	0.191

Table 19: Evaluation of ranking passages based on the predicted probability of *supported*, *refuted* or their sum *(both)* on the validation split.

always provide one randomly sampled passage S^0 to the model, which is necessary to understand the rationale (and reasoning gaps) behind the claim. If no passage S^0 exists, we use the paraphrased accurate premise from MISSCI instead. All experiments are averaged over five seeds (1-5).

F.1 Implementation Details

F.1.1 AFC as Rankers

A key challenge in retrieving passages that indicate reasoning gaps in a zero-shot setting is the absence of directly comparable tasks. For instance, passages containing content that undermines an argument may convey a supporting stance (e.g., if the claim is confirmed based on a small sample size, as in *Hasty Generalization*). When re-purposing AFC models to predict whether S_j points to a reasoning gap in the argument, we experiment with various strategies that aggregate the stance prediction labels from AFC models. Specifically, we measure the predicted probability mass for the labels SUPPORTED, REFUTED, or their sum[12] (*both*). For each strategy, we experiment with the role of the passage S^0 based on which the claim was made:

- **No S^0:** The AFC model sees only $\overline{c}$ and S_j.
- **Claim:** The claim is reformulated as an argument by incorporating S^0 through the expression "S^0 `Therefore: claim`".
- **Evid.:** S^0 is added at the beginning of the evidence passage (S_j) that is subject to ranking.

Table 19 shows the validation split results. We select the best-performing strategy according to MAP, which always concatenates a randomly sampled passage S^0 with the claim.

[12]Except when trained on CovidFact, which only exhibits two labels.

Model	MAP *no S_j^0*	S_j^0
BM25	0.283	**0.554**
SBERT (*concat*)	0.328	**0.522**
SBERT (*mean-1*)	0.333	0.366
SBERT (*mean-2*)	0.353	0.354
SBERT (*mean-3*)	0.308	0.342
INSTRUCTOR (*concat*)	0.388	**0.582**
INSTRUCTOR (*mean-1*)	0.404	0.480
INSTRUCTOR (*mean-2*)	0.389	0.516
INSTRUCTOR (*mean-3*)	0.368	0.479
SPICED-IMS (*concat*)	0.384	**0.562**
SPICED-IMS (*mean-1*)	0.379	0.429
SPICED-IMS (*mean-2*)	0.380	0.408
SPICED-IMS (*mean-3*)	0.323	0.380
PubMedBERT ST (*concat*)	0.283	0.490
PubMedBERT ST (*mean-1*)	0.337	**0.529**
PubMedBERT ST (*mean-2*)	0.318	0.513
PubMedBERT ST (*mean-3*)	0.324	0.490
BioBERT ST (*concat*)	0.270	0.516
BioBERT ST (*mean-1*)	0.313	0.531
BioBERT ST (*mean-2*)	0.311	**0.536**
BioBERT ST (*mean-3*)	0.280	0.521
SapBERT ST (*concat*)	0.323	0.484
SapBERT ST (*mean-1*)	0.312	0.497
SapBERT ST (*mean-2*)	0.325	**0.500**
SapBERT ST (*mean-3*)	0.323	**0.500**

Table 20: Ranking performance of embedding-based models on the validation instances to select passages linked to fallacies. We select the best model based on the highest MAP.

F.1.2 Dense embedding rankers

Using the same ranking models via embeddings or BM25 as in §E, we assess the impact of prepending a passage S^0 to the claim on the validation split in Table 20. We select the best configuration according to mAP for the test split. For the INSTRUCTOR, we modify the prompt to encode the claim by looking for "refuting" instead of "supporting" sentences (cf. §E.1.2), which resulted in a higher MAP on the validation set.

F.2 Impact of S_j^0 Passages

We artificially remove all passages from the ranked results if the passage communicates the content based on which the claim was made (i.e., S_j^0 passages) in Figure 9. Among these passages only, both rankers exhibit different strengths. INSTRUCTOR prefers passages based on which the *Biased Sample Fallacy* or the *Fallacy of Exclusion* can be detected. The AFC ranker prioritizes fallacies

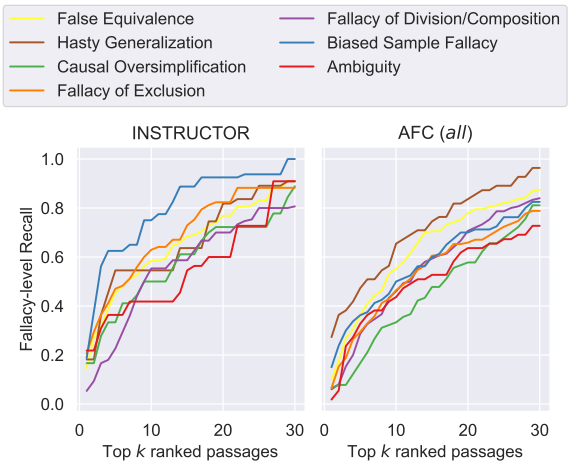

Figure 9: Recall of undermining passages per fallacy class over the top k ranked passages when removing all S^0 passages (based on which the claim was made). Only listing fallacies with ≥ 20 occurrences.

based on which *Hasty Generalization* and *False Equivalence* can be detected. An intuitive explanation is that both fallacies are similar to the support relationship and, hence, to the predicted probability for the label SUPPORTED.

F.3 Exhaustive Results on MISSCIPLUS

We list all results of detecting undermining passages on the MISSCIPLUS test split in Table 21.

G Details on Argument Reconstruction

G.1 LLM argument reconstruction

We provide the LLM with all passages S_j that were linked to at least one reasoning gap (or fallacy), and one passage S^0 that communicates the content based on which the claim was made. If multiple candidates for S^0 exist, we randomly select one. Experiments over MISSCI use the paraphrased p_0 instead of a randomly sampled passage S^0, and the publication context s_i instead of the linked passages S_j. We run each experiment over three different seeds (1-3). For Llama3-8B we use a temperature of 0.3. For a fair assessment given the changed requirements for the LLM, we perform extensive prompt-search over all six prompts evaluated for MISSCI and an additional new prompts (cf. §I.3). During prompt search, we provide the complete fallacy information consisting of the definition, the logical form, and an example (cf. §A; Tables 7-8) within the instructions. All prompts task LLMs to output a ranked list of verbalized fallacious premises and the applied fallacy class. If possible, fallacies outside our inventory were converted; otherwise, they were removed. The results are listed in Table 22. We additionally report the average number of predicted fallacies per argument. This number excludes fallacies that the LLM hallucinated and are invalid (e.g., *Contextomizer*, *Accident*, *Conflict of Interest*, *Fallacy of Conclusion*) or that are outside of our fallacy inventory (e.g., *Fear Mongering*, *Non Sequitur*, *Ad Hominem*, *False Consensus*).

G.2 Llama3 Judge ϕ

We experiment with four different prompts (cf. §I.2) with the Llama3-8B-Instruct model as a binary classifier ϕ that determines whether two fallacious premises exhibit fallacious reasoning that bridges the identical gap. We use the human evaluation data provided by Glockner et al. (2024a) for training data. We discard all trivially invalid instances where the generated premise $\hat{\bar{p}}_i$ was almost identical to the claim or the paraphrased publication context s_i by discarding all premise pairs if

$$\min\Big[\mathrm{lev}(\hat{\bar{\mathrm{p}}}_\mathrm{i}, \bar{\mathrm{c}}), \mathrm{lev}(\hat{\bar{\mathrm{p}}}_\mathrm{i}, \bar{\mathrm{s}}_\mathrm{i})\Big] \leq t$$

where lev is the Levensthein distance and the threshold $t = 2$. This yielded a total of 168/240 manually annotated premise pairs. To adapt Llama3-8B-Instruct for the task, we perform (1) zero-shot experiments, (2) in-context learning (Brown et al., 2020) (ICL) experiments and (3) supervised fine-tuning (SFT) using QLoRA (Dettmers et al., 2023) and 8bit quantization (Dettmers et al., 2022). We validate each approach using five-fold cross-validation where folds are separated by the arguments, ensuring the LLM is evaluated with premises from unseen arguments. We set the temperature to zero and evaluate all ICL and SFT experiments over three random seeds (1,2,3) to account for different (ordering) of seen instances. The results are listed in Table 23. As a baseline, we report the performance of always predicting the *majority* label. We further report a baseline using a univariate logistic regression (*LR*) on top of the automatic *NLI-S* (Glockner et al., 2024a), which showed the highest correlation with human judgment in MISSCI. NLI-S uses the predicted entailment probability of a T5 (Raffel et al., 2020) model fine-tuned by Honovich et al. (2022). Rather than only considering the entailment score given the reference text as a premise and generated text as a hypothesis, NLI-S swaps its roles to avoid

Model	mAP	*Passage-wise* P@1	P@3	P@10	*Fallacy-wise* R@1	R@3	R@10
Random	0.205	0.136	0.130	0.124	0.168	0.252	0.438
Ordered	0.286	0.298	0.210	0.145	0.266	0.374	0.497
BM25	0.496	0.617	0.347	0.212	0.413	0.485	0.602
SBERT (Reimers and Gurevych, 2019)	0.520	0.640	0.381	**0.221**	0.423	0.503	0.609
INSTRUCTOR (Su et al., 2022)	**0.541**	**0.652**	**0.409**	**0.222**	**0.439**	0.529	**0.613**
SPICED-IMS (Wright et al., 2022)	0.524	0.640	0.388	**0.219**	0.423	0.516	0.595
PubMedBERT ST (Deka et al., 2022)	0.489	0.588	0.360	0.207	0.387	**0.548**	0.605
BioBERT ST (Deka et al., 2022)	0.491	0.600	0.377	0.200	0.400	0.495	0.570
SapBERT ST (Deka et al., 2022)	0.504	0.619	0.379	0.211	0.411	0.547	**0.615**
DeBERTaV3 SciFact (Wadden et al., 2020)	0.360	0.326	0.266	0.172	0.264	0.393	0.518
DeBERTaV3 CovidFact (Saakyan et al., 2021)	0.380	0.457	0.260	0.165	0.326	0.407	0.538
DeBERTaV3 HealthVer (Sarrouti et al., 2021)	0.368	0.410	0.284	0.176	0.321	0.426	0.544
DeBERTaV3 Scientific AFC (*all*)	0.306	0.338	0.263	0.177	0.267	0.400	0.554

Table 21: Ranking results for finding passages linked to fallacies on the test split, using all passages of the misrepresented publication.

Prompt	Passages	*Fallacy Level (premise+class)* R@5	P@5	P@1	*Arg. Level* F1@5	Arg@1	*Count* Pred/Arg
p1-basic	per-passage	0.079	0.046	0.133	0.306	0.167	2.6
p2-support	per-passage	0.083	0.048	0.078	0.296	0.200	2.7
p3-undermine	per-passage	0.079	0.047	0.089	0.302	0.178	2.4
p4-connect	per-passage	0.176	**0.134**	**0.222**	0.359	0.333	1.4
p5-auto	per-passage	0.083	0.066	0.144	0.263	0.178	1.4
p6-auto-connect	per-passage	0.111	0.100	0.156	0.272	0.256	1.1
p7-passage	per-passage	0.093	0.047	0.078	0.357	0.222	4.2
p8-passage-assumptions	per-passage	0.162	0.084	0.122	0.379	0.356	4.2
p9-passage-detective	per-passage	0.157	0.146	0.222	0.270	0.344	1.1
p10-passage-evaluate	per-passage	0.120	0.060	0.111	0.337	0.289	4.8
p11-passage-evaluate-2	per-passage	0.176	0.087	0.178	0.353	0.367	4.6
p12-passage-detective-2	per-passage	**0.218**	0.113	0.211	**0.360**	**0.444**	3.9
p13-passage-evaluate-3	per-passage	0.167	0.083	0.167	0.359	0.300	4.4
p1-basic	all passages	0.060	0.053	0.067	0.294	0.144	2.8
p2-support	all passages	0.065	0.058	0.044	0.303	0.156	2.8
p3-undermine	all passages	0.046	0.040	0.044	0.280	0.111	2.9
p4-connect	all passages	0.093	**0.162**	**0.178**	0.239	0.211	1.4
p5-auto	all passages	0.056	0.070	0.089	0.247	0.133	2.0
p6-auto-connect	all passages	0.046	0.110	0.111	0.163	0.111	1.0
p7-passage	all passages	0.042	0.024	0.022	0.333	0.100	4.4
p8-passage-assumptions	all passages	0.144	0.088	0.089	0.376	0.289	4.2
p9-passage-detective	all passages	0.069	0.153	0.133	0.218	0.167	1.1
p10-passage-evaluate	all passages	0.130	0.069	0.167	0.377	0.233	4.7
p11-passage-evaluate-2	all passages	0.144	0.077	0.089	0.371	0.256	5.1
p12-passage-detective-2	all passages	**0.171**	0.105	0.144	0.379	0.400	3.9
p13-passage-evaluate-3	all passages	0.162	0.094	0.124	**0.402**	**0.344**	4.4

Table 22: Argument reconstruction prompt-tuning using Llama3-8B-Instruct on the validation split.

Method	Prompt	Extra	F1	Acc.
Majority	–	–	0.372	0.593
NLI-S + LR	–	–	0.642	0.695
Zeroshot	p1	0-shot	0.570	0.581
Zeroshot	p2	0-shot	0.636	0.647
Zeroshot	p3	0-shot	0.636	0.641
Zeroshot	p4	0-shot	0.626	0.629
Zeroshot	p5	0-shot	0.562	0.569
ICL	p1	8-shot	0.623	0.625
ICL	p2	8-shot	0.635	0.639
ICL	p3	8-shot	0.620	0.623
ICL	p4	8-shot	0.604	0.607
ICL	p5	8-shot	0.612	0.617
ICL	p1	16-shot	0.656	0.663
ICL	p2	16-shot	0.640	0.645
ICL	p3	16-shot	0.645	0.651
ICL	p4	16-shot	0.656	0.661
ICL	p5	16-shot	0.652	0.657
SFT	p4	1-epochs	0.671	0.687
SFT	p4	2-epochs	0.690	0.711
SFT	p4	3-epochs	0.725	0.749
SFT	p4	4-epochs	0.751	0.770
SFT	p4	5-epochs	**0.788**	**0.798**
SFT	p4	6-epochs	0.761	0.776

Table 23: Cross-validation evaluation for implementations of the $\phi^{\text{p+f}}$ judge. All prompt-based approaches use Llama3-8B-Instruct as backend.

penalizing the model if it generates more specific text. Without SFT, ICL with 16 shots and the template *p4* reached the best performance measured via the F1-score. We consequentially performed SFT using the same prompt, which led to the overall best model with fine-tuning for 5 epochs with a *linear* schedule, a learning rate of $5e-4$, a batch-size of 4 and $\alpha = 16$, $r = 64$, $\text{dropout} = 0.2$ for QLoRA. We use these hyperparameters to fine-tune Llama3-8B-Instruct on the entire data and use it as the backend model for $\phi^{\text{f+p}}$.

G.2.1 Prompting Strategies

Prompts using *per-passage* prompting follow the prompting scheme of MISSCI and have a dedicated field for the content based on which the claim was made (p_0 or S_j^0) and for the publication context necessary to detect the fallacy (s_i or S_j). This prompting technique requires n separate prompts for n passages. Specifically, we create $n-1$ prompts using a randomly sampled passage S^0 (based on which the claim was made) together with each other passage linked to a fallacy, separately. In MISSCI-PLUS, the passage S^0 itself may be linked to fallacies. Hence, we prompt the model again with only the passage S^0. When selecting the top k results of multiple prompts for the same argument, combine all results, consisting of the fallacious premise and fallacy class, while keeping their ranking information. We then return the top k results based on their prompt-specific rank. We prefer results with different fallacy classes when multiple fallacies share the same rank to avoid sampling the same fallacious reasoning numerous times. Prompts that concatenate *all passages* within a single prompt follow the holistic view of arguments. We sort all passages based on the order in which they occur in the scientific publication and only prompt the LLM once per argument.

G.2.2 Performance on MISSCI with only linked publication context

We report the argument reconstruction of the evaluated LLMs in Table 24. This evaluation only considers a prediction of an LLM based on the paraphrased information s_i if a passage S_j that communicates the same information exists, and the same fallacy could have been detected based on S_j. This serves as a complementary comparison between the performance on MISSCI and MISSCI-PLUS, which removes benefits over MISSCI due to additional information.

H AFC applied on MISSCIPLUS

H.1 Stance Predictions per Fallacy Class

We aim to understand how passages that point to different reasoning gaps (and hence different fallacy classes) affect the stance prediction of scientific AFC models. Figure 10 visualizes for each AFC model and fallacy class the distribution over the predicted veracity labels of SUPPORTED, REFUTED and NEI. Regardless of the fallacy class, rarely any passage is identified as refuting the claim. Most models vary between the labels SUPPORTED and NEI with only minor clear trends among different fallacy classes. Interestingly, the model trained on COVIDFACT, which had the best misinformation detection rate according to Table 5 (as it does not know the label NEI) almost always tends to predict SUPPORTED more frequently than REFUTED over these passages that exhibit fallacious reasoning behind the claim.

H.2 Binary Fallacy Detection by LLMs

We adjusted our prompt (cf. §I.4) to allow LLMs to select no fallacy if they consider a claim correct and empirically evaluate three LLMs in how well they detect misinformation in MISSCIPLUS

LLM	Info	R@5 (ϕ^{f+p})	R@5 (ϕ^{f})	Arg@1 (ϕ^{f+p})
Llama3-8B	DLE	0.255	0.520	0.516
	DL	0.237	0.453	0.520
	DE	0.220	0.470	0.468
	LE	0.251	0.492	0.504
GPT-3.5	DLE	0.217	0.456	0.464
	DL	0.198	0.438	0.413
	DE	0.229	0.461	0.500
	LE	0.213	0.417	0.464
GPT-4 Turbo	DLE	0.327	0.472	0.619
	DL	0.280	0.458	0.560
	DE	0.294	0.500	0.571
	LE	0.299	0.514	0.583

Table 24: Argument reconstruction performance only over fallacies that could be detected based on linked passages (or based on the accurate premise alone).

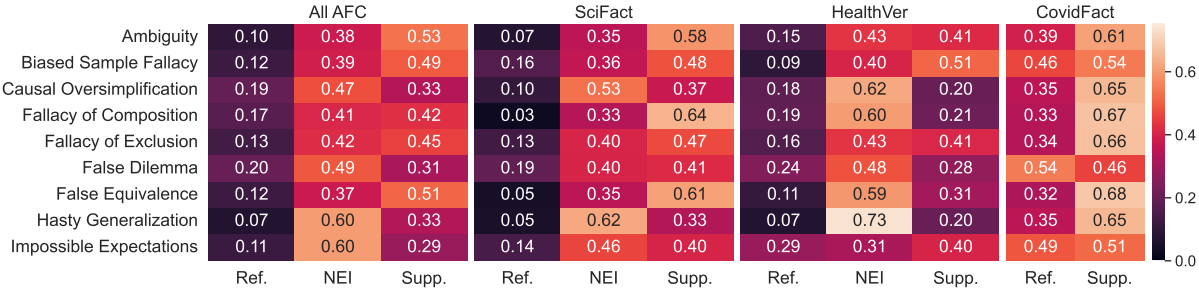

Figure 10: Distribution over predicted fact-checking labels per fallacy class, which was linked to the passage used as evidence, when predicting the veracity label. Results are averaged over five seeds for each fine-tuned AFC model.

	Detected a fallacy	
LLM	MISSCIPLUS	True Claims
Llama2-70B	0.988	0.993
Llama3-8B	0.877	0.783
GPT-3.5	0.853	0.923

Table 25: Binary fallacy evaluation over misinformation from MISSCIPLUS and correct claims from HEALTHVER and COVIDFACT.

or correct information in the 100 selected claims from HEALTHVER and COVIDFACT. Since true claims only come with one evidence passage, we also prompt the LLMs once only for misinformation using the concatenated relevant passages (cf. *all passages* in §6). Table 25 reports the ratio of claims identified as fallacious (the LLM did not specifically say that "no fallacy" exists *and* found at least one fallacy) across both datasets averaged over three seeds. We do not discern *which* fallacy was detected. Using the prompts to reconstruct fallacious arguments, the LLMs tend to find fallacies in all claims regardless of whether they are correct. We note that our claims are designed to correctly verbalize and identify all applied fallacies in MISSCIPLUS. This follows previous works on fallacy detection (Da San Martino et al., 2019; Jin et al., 2022; Alhindi et al., 2022) that focused on which fallacy was applied rather than determining *if* a fallacy was applied. Future work may explore how to detect fallacies and correct claims better.

H.3 LLMs may decline to provide a veracity prediction.

The percentages of the label classification from LLMs as fact-checking models using their parametric knowledge or RAG evidence in Table 6 may not sum up to 100% if the LLM declines to answer. We provide an extended Table that includes the percentage when the LLM declines to answer in Table 26.

I Prompts

I.1 PRP Prompts

The prompt template for PRP ranking is shown in Figure 11 with the different `[INSTRUCTION]` listed in Table 27.

	LLM	*Predicted as* True	False	NEI	No Answer	True	False	NEI	No Answer
Know (*FC*)	Llama2	1.6	61.1	37.3	0.0	34.7	22.3	41.3	1.7
	Llama3	0.0	86.9	2.4	10.7	20.0	43.3	14.3	22.3
	GPT4	0.0	85.3	14.7	0.0	59.0	23.0	17.3	0.7
	GPT3.5	0.8	71.0	28.2	0.0	46.7	17.3	35.7	0.3
Know (*Ask*)	Llama2	0.0	100.0	0.0	0.0	29.7	69.3	1.0	0.0
	Llama3	8.3	88.9	2.4	0.4	68.7	26.0	3.0	2.3
	GPT4	3.6	68.3	27.8	0.4	49.7	6.7	36.3	7.3
	GPT3.5	1.6	50.4	48.0	0.0	47.3	6.0	45.0	1.7
RAG	Llama2	23.8	61.5	12.7	2.0	58.7	29.7	10.7	1.0
	Llama3	44.4	53.2	2.4	0.0	80.3	16.3	3.3	0.0
	GPT4	27.4	34.1	38.5	0.0	55.0	4.0	41.0	0.0
	GPT3.5	38.9	31.7	29.0	0.4	78.0	5.3	16.0	0.7
		Misinformation				**True claims**			

Table 26: Averaged veracity predictions from LLMs on misinformation from MISSCIPLUS (*left*) and true claims from HEALTHVER and COVIDFACT (*right*).

Name	Instructions
Qin et al. (2023)	Given a query [CLAIM], which of the following two passages is more relevant to the query?
Qin et al. (2023) (claim)	Given a claim [CLAIM], which of the following two passages is more relevant to the claim?
Qin et al. (2023) (convincing)	Given a claim [CLAIM], which of the following two passages constitutes more convincing evidence for the claim?
Qin et al. (2023) (support)	Given a claim [CLAIM], is it more likely that someone made the claim based on passage A or based on passage B?

Table 27: Evaluated instructions used for PRP prompting.

Name	Instructions
p1	You are given two premises that exhibit some reasoning of a larger argument. Both premises apply a fallacy. Your task is to select whether the reasoning on an abstract level is identical. If one premise is more specific, they can still apply the same false reasoning. Provide your answer in the first line of your response. Answer with "match" if both premises apply the same false reasoning. Answer with "no-match" if they apply different false reasoning.
p2	I'll present you with two premises, each containing a fallacy in their reasoning. Analyze both statements: Your task: Determine if the core flawed logic behind the fallacies in both statements is identical. Respond with "match" if the underlying reasoning is the same, even if the specifics differ. Respond with "no-match" if they represent different fallacies.
p3	Task: Analyze both premises and determine if they commit the same type of fallacy. Answer: (match / no-match)
p4	Determine whether two premises exhibit identical false reasoning, regardless of their specificity. Provide your answer in the first line of your response. Answer with "match" if both premises apply the same false reasoning. Answer with "no-match" if they apply different false reasoning.
p5	Task: Analyze both premises and determine if they apply a similar reasoning regardless of specificity. Answer: (match / no-match)

Table 28: Instructions used for the judge $\phi^{\text{p+f}}$.

```
[INSTRUCTIONS]

Passage A: [PASSAGE-A]
Passage B: [PASSAGE-B]

Output Passage A or Passage B:
```

Figure 11: Prompt template for PRP prompting.

```
[INSTRUCTIONS]

Premises:
1: "[GEN-PREMISE]"
2: "[GOLD-PREMISE]"

Question: Do both premises use the same flawed reasoning (fallacy)?
```

Figure 12: Prompt for the judge $\phi^{\text{p+f}}$.

I.2 Premise judge $\phi^{\text{p+f}}$

The prompt template for the judge $\phi^{\text{p+f}}$ with ICL and SFT approaches is shown in Figure 12. The instructions are listed in Table 28.

I.3 Argument reconstruction prompts

We take the p1-p6 prompts from MISSCI as-is. For *all-passages* prompting we concatenate all passages (except for one S_j^0 passage), treating them as *publication context* in the MISSCI prompt, and treating the left-out S_j^0 as the accurate premise. The *per-passage* prompt templates for p7-p13 are shown in Figures 13-19., where we always replace `[PASSAGE S0]` with a randomly sampled passage S_j^0 and `[PASSAGE Sj]` with a passage S_j linked to a fallacy. For all, we present the fallacy information consisting of the fallacy definition, logical form and example from literature identically to MISSCI prompts. The *all-passage* prompts are minimally edited from these prompts and have one dedicated field for all selected passages. All used prompts are provided within our repository.

I.4 LLM as AFC prompts

The prompts used to directly predict the veracity of claims using LLMs are shown in Figures 20-22. We replace `[EVIDENCE]` with the concatenated evidence passages. The adapted fallacy generation prompt which allows to output that no fallacy exists, is shown in Figure 23.

J Reproducibility

All experiments with Llama2 or Llama3 used the instruction-tuned LLM and were performed on 80GB A100 GPUs. We always used the 70B and 8B models for Llama2 and Llama3, respectively. All experiments with LLMs are averaged over three seeds. The only exception is the experiments with GPT-4 Turbo in Table 3, which we only ran once due to computational costs. We used the API version `2023-10-01-preview` for GPT-3.5 (model version: `0613`)[13], GPT-4 (model version: `0613`) and GPT-4 Turbo (model version: `1106-Preview`). We used Grammarly[14] in writing this paper.

[13] We used `gpt-35-turbo-16k` in the *all passages* prompting experiments and `gpt-35-turbo` in all other experiments

[14] https://app.grammarly.com/

[FALLACY INFORMATION]

Task:
Examine the following fallacious argument:

Passage 1: "[PASSAGE S0]"
Passage 2: "[PASSAGE Sj]"
Claim: "[CLAIM]"

Passages 1 and 2 are sourced from the same credible scientific document. The claim is based on the content of both passages. Your task is to identify and verbalize the fallacious reasoning as the fallacious premise necessary to support the claim given the content of passages 1 and 2. This reasoning should effectively support the claim, ensuring that the passages do 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 multiple applicable fallacies exist, list all fallacious premises with the applied fallacy classes in order of relevance (most to least relevant).

Figure 13: Prompt template for argument reconstruction *p7-passage*.

[FALLACY INFORMATION]

Task:
Examine the following fallacious argument:

Passage 1: "[PASSAGE S0]"
Passage 2: "[PASSAGE Sj]"
Claim: "[CLAIM]"

Passages 1 and 2 are sourced from the same credible scientific document. The claim is based on the content of both passages. Your task is to identify and verbalize the fallacious reasoning (the hidden assumptions) as the fallacious premise necessary to support the claim given the content of passages 1 and 2. This reasoning should effectively support the claim, ensuring that the passages do 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 multiple applicable fallacies exist, list all fallacious premises with the applied fallacy classes in order of relevance (most to least relevant).

Figure 14: Prompt template for argument reconstruction *p8-passage-assumptions*.

[FALLACY INFORMATION]

Challenge:
You've been assigned a detective mission in the world of scientific arguments!
The Case:
An argument has been made based on two passages from a credible scientific document. However, there's a hidden flaw in the reasoning. Your job is to uncover this hidden assumption – the "fallacious premise" – that makes the argument illogical.
The Evidence:
Passage 1 (This is the first piece of information from the scientific document.): "[PASSAGE S0]" Passage 2 (This provides additional details from the same document.): "[PASSAGE Sj]" Claim (This is the conclusion drawn from both passages): "[CLAIM]"
The Tools:
Fallacy Inventory: You have access to a list of common fallacies (errors in reasoning). Only use these fallacies!
Your Mission:
1. Analyze the passages and the claim.
2. Identify the hidden assumption that's needed for the claim to follow logically from the passages.
3. Formulate this hidden assumption as a clear "fallacious premise."
4. Identify the specific type of fallacy from the fallacy inventory that best explains this hidden assumption.
Deliverables:
Present your findings in this format:
"Fallacious Premise: <fallacious premise>; Applied Fallacy Class: <applied fallacy class>."
If multiple applicable fallacies exist, list all fallacious premises with the applied fallacy classes in order of relevance (most to least relevant). Only consider fallacies from the provided fallacy inventory.
Remember:
The passages come from a credible scientific document, so the information itself is likely true. The fallacy lies in how the information is used to support the claim. Focus on the hidden assumption(s) needed to bridge the gap between the passages and the claim.

Figure 15: Prompt template for argument reconstruction *p9-passage-detective*.

[FALLACY INFORMATION]

Task:
This activity focuses on identifying weaknesses in scientific arguments based on source materials.
Materials:
Passage 1: "[PASSAGE S0]"
Passage 2: "[PASSAGE Sj]"
Claim: "[CLAIM]"
The claim is derived from the content of both passages. Both passages stem from the same credible scientific publication.
Objective:
Analyze the relationship between the provided passages and the claim. Identify any assumptions or gaps in reasoning that might weaken the argument's validity.
Instructions:
1. Read Passage 1, Passage 2, and the claim carefully.
2. Consider how the information in the passages connects to the claim.
3. Identify any missing information or hidden assumptions that would be necessary for the claim to logically follow from the passages.
4. Formulate these missing pieces as clear "fallacious premises."
5. Using the provided fallacy list, identify the type of fallacy associated with each fallacious premise.
Output:
Present your findings in this format:
"Fallacious Premise: <fallacious premise>; Applied Fallacy Class: <applied fallacy class>."
If multiple applicable fallacies exist, list them in order of relevance (most to least relevant). Only consider fallacies from the provided fallacy inventory.
Important Note:
The passages are sourced from a credible scientific document, so the information itself is most likely accurate. The focus is on how the information is used to support the claim, not questioning the scientific content.

Figure 16: Prompt template for argument reconstruction *p10-passage-evaluate*.

[FALLACY INFORMATION]

Task:
This activity focuses on identifying weaknesses in scientific arguments based on source materials.
Materials:
Passage 1: "[PASSAGE S0]"
Passage 2: "[PASSAGE Sj]"
Claim: "[CLAIM]"
The claim is derived from the content of both passages. Both passages stem from the same credible scientific publication.
Objective:
Analyze the relationship between the provided passages and the claim. Identify any assumptions or gaps in reasoning that might weaken the argument's validity.
Instructions:
1. Read Passage 1, Passage 2, and the claim carefully.
2. Consider how the information in the passages connects to the claim.
3. Identify any missing information or hidden assumptions that would be necessary for the claim to logically follow from the passages.
4. Formulate these missing pieces as clear "fallacious premises."
5. Using the provided fallacy list, identify the type of fallacy associated with each fallacious premise.
Output:
Present your findings in this format: "Fallacious Premise: <fallacious premise>; Applied Fallacy Class: <applied fallacy class>." If multiple applicable fallacies exist, list all fallacious premises with the applied fallacy classes in order of relevance (most to least relevant). Only consider fallacies from the provided fallacy inventory.
Important Note:
The passages are sourced from a credible scientific document, so the information itself is most likely accurate. The focus is on how the information is used to support the claim, not questioning the scientific content.

Figure 17: Prompt template for argument reconstruction *p11-passage-evaluate-2*.

[FALLACY INFORMATION]
Challenge:
You've been assigned a detective mission in the world of scientific arguments!
The Case:
An argument has been made based on two passages from a credible scientific document. However, there's a hidden flaw in the reasoning. Your job is to uncover this hidden assumption – the "fallacious premise" – that makes the argument illogical.
The Evidence:
Passage 1 (This is the first piece of information from the scientific document.): "[PASSAGE S0]"
Passage 2 (This provides additional details from the same document.): "[PASSAGE Sj]"
Claim (This is the conclusion drawn from both passages): "[CLAIM]"
The Tools:
Fallacy Inventory: You have access to a list of common fallacies (errors in reasoning). Only use these fallacies!
Your Mission:
1. Analyze the passages and the claim.
2. Identify the hidden assumption that's needed for the claim to follow logically from the passages.
3. FOR EACH identified flaw:
a. Formulate this hidden assumption as a clear "fallacious premise."
b. Using the provided fallacy inventory, identify the specific fallacy class from the provided fallacy inventory that best explains this fallacious premise.
c. Output each identified flaw in a separate line following this format: "Fallacious Premise: <fallacious premise>; Applied Fallacy Class: <applied fallacy class>."
Deliverables:
Present your findings in this format:
"Fallacious Premise: <fallacious premise>; Applied Fallacy Class: <applied fallacy class>."
If multiple applicable fallacies exist, list all fallacious premises with the applied fallacy classes in order of relevance (most to least relevant). Only consider fallacies from the provided fallacy inventory.
Remember:
The passages come from a credible scientific document, so the information itself is likely true. The fallacy lies in how the information is used to support the claim. Focus on the hidden assumption(s) needed to bridge the gap between the passages and the claim.

Figure 18: Prompt template for argument reconstruction *p12-passage-detective-2*.

[FALLACY INFORMATION]
Task:
This activity focuses on identifying weaknesses in scientific arguments based on source materials.
Materials:
Passage 1: "[PASSAGE S0]"
Passage 2: "[PASSAGE Sj]"
Claim: "[CLAIM]"
The claim is derived from the content of both passages. Both passages stem from the same credible scientific publication.
Objective:
Analyze the relationship between the provided passages and the claim. Identify any assumptions or gaps in reasoning that might weaken the argument's validity.
Your Mission:
1. Analyze the passages and the claim.
2. Identify the hidden assumption that's needed for the claim to follow logically from the passages.
3. FOR EACH identified flaw:
a. Formulate this hidden assumption as a clear "fallacious premise."
b. Using the provided fallacy inventory, identify the specific fallacy class from the provided fallacy inventory that best explains this fallacious premise.
c. Output each identified flaw in a separate line following this format: "Fallacious Premise: <fallacious premise>; Applied Fallacy Class: <applied fallacy class>."
Output:
Present your findings in this format:
"Fallacious Premise: <fallacious premise>; Applied Fallacy Class: <applied fallacy class>."
If multiple applicable fallacies exist, list all fallacious premises with the applied fallacy classes in order of relevance (most to least relevant). Only consider fallacies from the provided fallacy inventory.
Important Note:
The passages are sourced from a credible scientific document, so the information itself is most likely accurate. The focus is on how the information is used to support the claim, not questioning the scientific content.

Figure 19: Prompt template for argument reconstruction *p13-passage-detective-3*.

Write a fact-checking article about the claim that "[CLAIM]"

Conclude your fact-checking article with a verdict in a single line: *"Verdict: (your verdict label)"*. Choose one of the following verdicts: True, Unknown, False.

Figure 20: Prompt to generate a fact-checking article given no evidence.

Claim: "[CLAIM]"

To the best of your knowledge, what is the veracity of the claim? Provide a thorough explanation supporting your decision, select one of the answers (True, False, Unknown) and output the veracity of the claim in a single line: *"Veracity: (your veracity label)"*.

Figure 21: Prompt to ask for the veracity of the claim given no evidence.

Study:
[EVIDENCE]

Rate if the claim "[CLAIM]" is correct based on the study. Answer with *"Verdict: (Correct / Incorrect / Not Enough Information)"*

Figure 22: Prompt to ask for the veracity of the claim given evidence.

[FALLACY INFORMATION]

Challenge:
You've been assigned a detective mission in the world of scientific arguments!
The Case:
An argument has been made based on two passages from a credible scientific document. Your job is to assess the logical reasoning behind this argument.
The Evidence:
Passages (The used information from the scientific document): "[EVIDENCE]" Claim (This is the conclusion drawn from the passages): "[CLAIM]"
The Tools:
Fallacy Inventory: You have access to a list of common fallacies (errors in reasoning). Only use these fallacies!
Your Mission:
1. Analyze the passages and the claim.
2. Evaluate whether the claim logically follows from the information presented in the passages.
3. IF you identify flaws in the reasoning, pinpoint all hidden assumptions that's needed for the claim to follow logically.
4. FOR EACH identified flaw:
a. Formulate this hidden assumption as a clear "fallacious premise."
b. Using the provided fallacy inventory, identify the specific fallacy class from the provided fallacy inventory that best explains this fallacious premise.
c. Output each identified flaw in a separate line following this format: "Fallacious Premise: <fallacious premise>; Applied Fallacy Class: <applied fallacy class>."
5. IF the claim logically follows from the passages (no fallacy present), then simply output: "No Fallacy"
Deliverables:
If the argument is sound (if no fallacy or only minor fallacies exist), clearly state "No Fallacy" in the output.
If a fallacy is present, present your findings in this format: "Fallacious Premise: <fallacious premise>; Applied Fallacy Class: <applied fallacy class>."
If multiple applicable fallacies exist, list all fallacious premises with the applied fallacy classes in order of relevance (most to least relevant). Only consider fallacies from the provided fallacy inventory.
Remember:
The passages come from a credible scientific document, so the information itself is likely true. The fallacy lies in how the information is used to support the claim. Focus on the hidden assumption(s) needed to bridge the gap between the passages and the claim. If the scientific document supports the claim, no critical fallacy is applied.

Figure 23: Fallacy generation prompt that can predict "no fallacy".