CLAIMCHECK: How Grounded are LLM Critiques of Scientific Papers?

Anonymous ACL submission

Abstract

The rapid increase in paper submissions to top AI and ML venues in recent years, in tandem with the development of ever more capable LLMs, has fueled a surge of interest in leveraging these models to automate parts of the peer review process. A core component of the reviewer's task consists of providing *specific* critiques that directly assess the scientific *claims* a paper makes. While it is now relatively easy to automatically generate passable (if generic) reviews, ensuring that these reviews are sound and *grounded* in the papers' claims remains challenging—requiring expert-level domain knowledge, careful reading, and logical reasoning. Furthermore, resources supporting this goal are lacking. To remedy this, and to facilitate benchmarking of LLMs on these objectives, this paper introduces CLAIMCHECK, a dataset of NeurIPS 2023 and 2024 submissions and reviews, annotated by ML experts for weaknesses and the paper claims that they target. We benchmark GPT-4o on three claim-centric tasks supported by CLAIMCHECK and find that even this cutting-edge model exhibits significant weaknesses in these tasks.[1]

1 Introduction

Prior work has highlighted the recent rapid growth in submission rates to academic conferences (Yuan et al., 2022), including at top ones for AI and NLP (Staudinger et al., 2024)[2], resulting in heavy reviewer burdens and a surge of interest in automating parts of the peer review process (Dycke et al., 2023; Drori and Te'eni, 2024). Many tasks and datasets have been proposed targeting different facets of this process, including (meta-)review writing (Wang et al., 2020; Yuan et al., 2022; Shen et al., 2022, *i.a.*), review argument mining and analysis

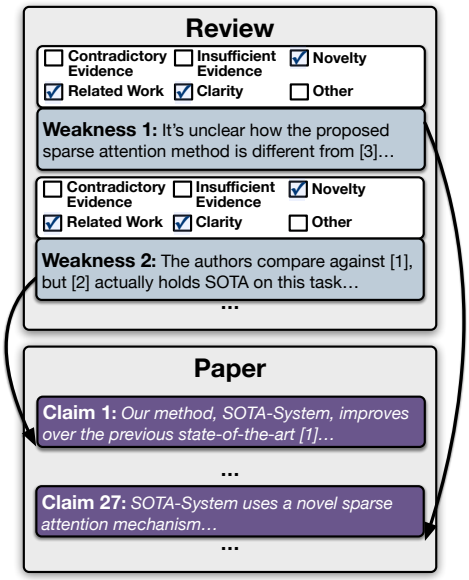

Figure 1: CLAIMCHECK identifies and annotates *weaknesses* in NeurIPS reviews and grounds them to the specific *target claims* that they dispute in the paper. *Grounding* weaknesses in a paper's claims is an essential part of peer review.

(Hua et al., 2019; Fromm et al., 2021; Guo et al., 2023, *i.a.*), determination of review score and acceptance judgments (Kang et al., 2018; Bharti et al., 2021, 2024, *i.a.*), and reviewer-paper assignment (Stelmakh et al., 2021, 2023), among others.

A core component of peer review is the expert critique of the *claims* that a paper makes—about results, theorems, approaches, novelty, etc. And indeed, it is essential to the effectiveness of such critiques that they be clearly *grounded* in the paper's claims. Unfortunately, overly broad and heuristic criticisms by reviewers are as endemic to these fields as they are condemned within them: In its reviewer guidelines, ACL Rolling Review (ARR) features a prominent injunction to "be specific"[3] and NeurIPS admonishes reviewers to "make your review as informative and *substantiated* as possi-

[1] Code & data will be made publicly available upon acceptance

[2] For example, 7x and 8x growth in ACL and NeurIPS submissions from 2014 to 2023 (Staudinger et al., 2024).

[3] https://aclrollingreview.org/reviewerguidelines

ble.[4]" Table 1 shows similar examples.

Curiously, however, the literature on automated peer review has given little attention to the problem of ensuring that reviews are specific and properly anchored to a paper's claims (see §2). As LLMs encroach ever more into intensive knowledge work of all kinds—not only peer review—adequately addressing the challenge of producing grounded generations is paramount.

On the other hand, collecting data for verifying in-the-wild claims from knowledge-intensive documents and grounding them in granular evidence is intrinsically challenging. Existing work tends to alleviate the challenge by narrowing the scope of the claims and/or the evidence pool (Wadden et al., 2020; Lu et al., 2023), making the systems developed based on them difficult to directly adapt to real-world scenarios like claim-grounded peer review. These approaches also tend to focus on the binary factuality of claims, impacting the applicability of datasets to real-world domains in which claims are often *flawed* but not entirely *false* (Estornell et al., 2020; Venkat et al., 2022).

This paper aims to address these challenges by introducing CLAIMCHECK, a novel resource for *automatic, claim-grounded peer review*. CLAIMCHECK is a high-quality multimodal collection of rejected NeurIPS submissions and their reviews, annotated by ML experts for rich information about the weaknesses identified in the reviews, with links to the in-text claims they target (see Figure 1). To our knowledge, CLAIMCHECK is the first resource that jointly tackles technical claim verification and claim-grounded peer review. Claims are sourced directly from papers' full texts (rather than synthetically constructed Thorne et al., 2018a,b) and are rarely *clearly true* or *clearly false*, and review weaknesses are annotated with an informative, multi-label ontology.

Further, we leverage CLAIMCHECK to benchmark GPT-4o[5] on a suite of claim-centric reviewing tasks, and find that even a cutting-edge multimodal LLM of this sort exhibits significant limitations as a reviewing assistant. We summarize our contributions as follows:

1. We introduce CLAIMCHECK, a dataset of real-world scientific papers, *claim-grounded* reviews, and rich expert annotations;

2. We present a novel suite of tasks for *claim-centric* scientific paper review evaluation, enabled by CLAIMCHECK;

3. We report experimental results on these tasks with GPT-4o, demonstrating the shortcomings of contemporary LLMs for automated, claim-grounded peer review.

2 Related Work

Automated Peer Review Automated peer review is a broad and rapidly growing area of research within AI and NLP, encompassing a wide array of tasks and datasets. We refer the reader to Staudinger et al. (2024) for a general overview and highlight more narrowly relevant work below.

In focusing on grounding reviewer weaknesses to targeted claims, we follow several prior works that emphasize the *dialectic* nature of peer review, in which authors and reviewers respond directly to one another. Cheng et al. (2020) introduce the RR (Review-Rebuttal or APE) dataset for mining arguments from reviews and rebuttals of ICLR submissions, and extracting aligned review-rebuttal argument pairs. The ARIES dataset from D'Arcy et al. (2023) features reviewer comments from submissions to several computer science conferences, automatically aligned to paper edits that were made in response. Kumar et al. (2023) study disagreements *among reviewers*, introducing the ContraSciView dataset, which contains pairs of reviews from ICLR and NeurIPS annotated for reviewer contradictions and disagreements. Lastly, Ruggeri et al. (2023) present ArgSciChat, a dataset of information-seeking (not *critical*) dialogues about a small set of NLP papers, curated by having experts trade questions and answers about each paper, with answers linked to rationale passages in the text.

Claim Verification Weaknesses identified by reviewers can be understood as *verifying* the claims that they target, and *claim* (or *fact*) *verification* is its own active research problem. Historically, datasets and shared tasks for claim verification, such as FEVER (Thorne et al., 2018a,b), SCIVER (Wadden and Lo, 2021), COVID-Fact (Saakyan et al., 2021) and AVeriTeC (Schlichtkrull et al., 2024), have tended to emphasize prediction of scalar veracity judgments over written explanations (like weaknesses provide; Dmonte et al., 2024), although a number of more recent works have given more

[4] https://neurips.cc/Conferences/2024/ReviewerGuidelines

[5] https://openai.com/index/hello-gpt-4o/

	Reviewer guideline excerpts advising *claim-centric* criticism
NeurIPS	Quality: Is the submission technically sound? Are claims well supported (e.g., by theoretical analysis or experimental results)?
ICLR	Does the paper support the claims? This includes determining if results, whether theoretical or empirical, are correct and if they are scientifically rigorous.
ARR	Inappropriate scope of the claims: The authors evaluate a sample that does not represent the population about which the claim is made. Hypotheses/speculations presented as conclusions: Every claim that is made has to be based on evidence or arguments (the authors' or from other work), or clearly marked as conjecture/speculation. Misleading or inappropriate framing, overclaiming: E.g., concluding from benchmark evaluation that LLMs generally 'understand' language, without validating that construct

Table 1: Excerpts from reviewer guidelines of top AI/ML/NLP venues that advise specific, claim-centric reviews.

attention to the latter (Yang et al., 2022; Rani et al., 2023; Ma et al., 2024, *i.a.*).

Beyond SCIVER and COVID-Fact, several other claim verification datasets focus on scientific domains. Notable examples include SciFACT (Wadden et al., 2020), which features 1.4k expert-written scientific claims from a variety of fields (e.g. microbiology, public health); SciFACT-Open (Wadden et al., 2022), which builds on SciFACT, with an additional 279 claims from similarly diverse areas; and SciTAB (Lu et al., 2023), which provides a set of 1.2k claims describing table results extracted from arXiv papers on computer science, each requiring compositional reasoning on tables for their verification.[6]

Our Work While CLAIMCHECK draws raw data from similar sources as other works on peer review (*viz.* NeurIPS OpenReview submissions), it is unique in focusing on the relationship between reviewer-identified weaknesses and papers' claims. Further, the suite of tasks we explore in §4 appear to be novel to this domain.

Within the claim verification literature, our work is distinctive in drawing evidence for disputed claims from reviews and in leveraging complete paper data (text, images, figures, algorithms, captions)—from both the reviewed paper and relevant prior works—for verification.

3 CLAIMCHECK Construction

3.1 Overview

We aim to collect pairs consisting of (1) a *claim-related weakness* and (2) one or more *target claims*, given a paper and a review of that paper. We define a *claim-related weakness* as a contiguous passage from the review that disputes the validity of one or more claims that the paper makes.[7] For each weakness we also collect a detailed set of labels.

We describe the full set of annotation tasks in §3.3 and the actual annotation process in §3.4, but begin with our preprocessing pipeline (§3.2).

3.2 Preprocessing

In selecting papers and reviews for CLAIMCHECK annotation, we sought a corpus that satisfied the following desiderata: 1) *open-access*: the papers and reviews should be publicly available; 2) *domain*: paper topics should align with the expertise of our annotators (primarily NLP); 3) *recency*: the papers should reflect relatively up-to-date research trends in AI and NLP; and 4) *version alignment*: the publicly available versions of the papers should be the *exact* version that the reviews comment on.

After an initial search, we found that rejected OpenReview submissions to NeurIPS 2023 and 2024 met these criteria. We note that, unfortunately, only the camera-ready versions are available for *accepted* papers.

We obtain an initial set of 1,575 publicly available reviews (from 378 rejected papers) from the OpenReview API [8], which is then filtered using a two-step process. First, we subset to reviews that contain at least one of a predefined set of claim-related keywords (see Appendix A). We then further filter this subset to reviews of papers that are broadly related to NLP—our annotators' primary area of expertise—determined by zero-shot prompting GPT-4o. This process yielded a final set of 60 reviews and 41 papers for annotation.

We download the PDFs for all 41 papers and parse the full text using PaperMage (Lo et al., 2023), and further clean the text to mitigate OCR

[6] See also Sarrouti et al. (2021); Wang et al. (2021); Akhtar et al. (2022). We refer the reader to Dmonte et al. (2024) for a good general overview of claim verification.

[7] Claim-related weaknesses can be contrasted with those *not* about (a) specific claim(s) made in the paper, such as those highlighting key omissions or issues with the paper taken as a whole. Such weaknesses are not the focus of our work.

[8] https://docs.openreview.net/reference/api-v2

noises/errors. We then manually extract as images all tables, figures, and algorithms, along with the captions for each. Finally, we automatically extract claims from the full text of the paper. Text cleaning, topic classification, caption extraction, and claim extraction are all done by zero-shot prompting GPT-4o (see Appendix D for prompts).

Finally, a number of the reviews cite related work in connection with the issues they raise. These works may thus provide information critical to assessing the review and the claim(s) it disputes. To ensure that these works are included, we manually read through each review, identifying related works that they cite, and then perform the same preprocessing steps described above on each. This process yielded 56 related work papers.

3.3 Annotation Tasks

CLAIMCHECK annotation consists of three tasks:

1. **Weakness Identification (WI)**: identification of review passages describing *claim-related weaknesses*.
2. **Claim Association (CA)**: Identifying the *target claims* disputed by each weakness.
3. **Weakness Labeling (WL)**: providing a set of informative labels for each weakness.

All three tasks take as input the full paper PDF and a single review of that paper. Further task-specific information is provided depending on the task. The WI task was conducted in one interface and the WL and CA tasks were conducted together in another. Appendix B contains screenshots of the interfaces and other annotation information. We detail each task below.

Weakness Identification (WI) Annotators are shown the full review text and must highlight contiguous passages that describe *claim-related weaknesses* (see §3.1). Passages that raise other issues that are clearly *not* based on a specific claim or result (e.g. unclear exposition, missing related work) are not highlighted. Annotators then provide a *groundedness* confidence label (1-5) for each weakness, indicating the extent to which they believe the weakness to be grounded in an explicit claim in the paper (5), rather than in a broad or speculative claim imputed by the reviewer (1). These labels are not of inherent interest, but rather are collected to help annotators in the CA task.

Claim Association (CA) entails identifying claims in a paper that are *target claims* of the weaknesses identified in WI. We say that a claim c is a *target claim* of a *claim-related weakness* w iff (1) the truth or accuracy of c is clearly disputed by w, and (2) determining this does not require appealing to any other claim(s).[9] Importantly, not all weaknesses have target claims. This is the rationale for the groundedness confidence labels collected in WI: to help CA annotators triage those that are (not) likely to be grounded in an explicit target claim. As additional task input, annotators are given the set of claims automatically extracted from the paper and select the target claims from this set.

Along with paper and review details, the annotation interface shows the weaknesses identified in WI and the set of extracted candidate claims. Annotators toggle through the claims, tentatively indicating for each whether they think it *may* be a target of each weakness. Only after seeing all claims do annotators finalize each weakness's target claims by selecting a (potentially improper) subset of the tentative target claims identified so far. Additionally, if annotators feel that a weakness clearly targets *some* claim in the paper—but one not included in the candidate set (e.g. due to its being missed during automatic extraction)—they are allowed to manually add it here.

Weakness Labeling (WL) asks annotators to provide further labels on the weaknesses, given the (*claim-related weakness*, *target claims*) pairs collected from WI and CA—ones that *are* of inherent interest, in contrast to the groundedness confidence labels. The additional labels include: (1) an ordinal *subjectivity* rating, indicating the extent to which the weakness is based on subjective factors (e.g. interest in the topic) vs. objective facts about paper contents; (2) an ordinal *agreement* rating, indicating the extent to which the annotator agrees with the weakness; and (3) one or more *weakness type* labels, characterizing the issue(s) raised by the weakness towards the claims (*insufficient evidence*, *contradictory evidence*, *novelty*, *clarity*, *related work*, or *other*).[10]

[9] Condition (2) thus restricts target claims to those *most directly implicated* by the weakness; a weakness in one claim may have implications for others, but we do not count these others as *target claims* for our purposes.

[10] Both the weakness taxonomy and the decision to have a multi-label (vs. categorical) scheme were determined by the annotation team through multiple rounds of reading papers and reviews prior to beginning CLAIMCHECK annotation proper.

	Agr	Sub	Con	Ins	Nov	Rel	Cla	Oth
Humans Only	18.2	13.1	17.9	44.6	77.6	52.4	0.0	22.8
Humans + GPT	18.1	9.8	16.4	40.4	78.3	52.4	-1.1	17.5

Table 2: Agreement (α) on the WL pilot task between annotators with (bottom) and without (top) GPT-4o included as an additional annotator. Agreement drops for most labels when GPT-4o is included, suggesting that the model struggles with this task relative to human experts. **Agr** and **Sub** use ordinal α (1-5); the rest use nominal (binary).

3.4 Annotation Process

All annotators are authors of this work and are either Ph.D. students or full-time researchers in AI/NLP. None received monetary compensation.

WI: Pilot Pilot annotations for this task were collected on a set of five (paper, review) pairs. Six annotators completed the WI pilot. We calculate pairwise agreement between annotators on weakness span selection by (1) obtaining alignments between weaknesses by solving a linear sum assignment between their selected spans, using normalized edit distance as the span similarity; then (2) computing micro-average pairwise span F_1 using this same similarity ($F_{1,edit}$), obtaining $F_{1,edit} = 52.4$.[11]

WI: Main All examples in the main annotation were singly annotated. Five of the six annotators from the WI pilot performed this annotation and were instructed to annotate no more than 20 reviews each. In total, we obtain 168 weaknesses across the 60 reviews. Figure 2 shows the distributions of groundedness confidence scores, weakness types, subjectivity scores, and agreement scores for CLAIMCHECK.

WL + CA: Pilot Five annotators completed a pilot for the WL and CA subtasks (both done in the same interface) using the same set of five (paper, review) pairs as in the WI pilot. The input weaknesses were drawn from the WI pilot annotations of the annotator with the highest individual $F_{1,edit}$ agreement. Similar to the above, we report $F_{1,edit}$ on the identified target claims, obtaining a value of 45.8. Since annotators are also largely choosing from among a fixed set of candidate claims (rather than unrestricted span selection, as in WI), we also report exact-match F_1, obtaining $F_{1,exact} = 28.5$.

We report Krippendorff's α (Krippendorff, 1970) for (1) the weakness type labels, (2) weakness subjectivity, and (3) weakness agreement, using the nominal form of the alpha for each label in (1) and the ordinal form for (2) and (3). Results are shown in Table 2. For (1), we observe significant variability in agreement across labels—finding medium-to-high agreement for **Ins**ufficient evidence ($\alpha_{\textbf{Ins}} = 44.6$), **Rel**ated work ($\alpha_{\textbf{Rel}} = 52.4$), and **Nov**elty ($\alpha_{\textbf{Rel}} = 77.6$), but lower agreement on other labels (with $\alpha_{\textbf{Cla}}$ showing chance agreement). For (2) and (3), we find modest agreement ($\alpha_{\textbf{Agr}} = 18.2$, $\alpha_{\textbf{Sub}} = 13.1$). The modest and lower agreements on some of these labels reflect the intrinsic challenges of claim-grounded paper review - even for experts with carefully constructed label taxonomy, "meta-reviewing" reviews with grounding on specific claims remains inevitably subjective to some extent. And even for this subset of labels, we deem our annotations are still helpful in 1) providing insightful expert-level annotation and analysis for this realistic and challenging task; 2) offering informative references for evaluating and comparing LLMs with human experts in scenarios where a significant level of subjectivity judgments are involved.

Papers	41
Reviews	60
Related Work Papers	56
Target Claims	154
Weaknesses	168
→ w/ Target Claims	120

Table 3: Summary statistics for CLAIMCHECK.

WL + CA: Main All of the annotators from the CA subtask pilot participated in the CA main annotation and were again instructed to annotate no more than 20 reviews. In total, we obtained 154 target claims across the 60 reviews, where 120/168 weaknesses had at least one target claim. Summary statistics for CLAIMCHECK are shown in Table 3.

[11] We use edit distance rather than exact match for Span F_1 given that annotators may exhibit minor differences in how they determine span extents.

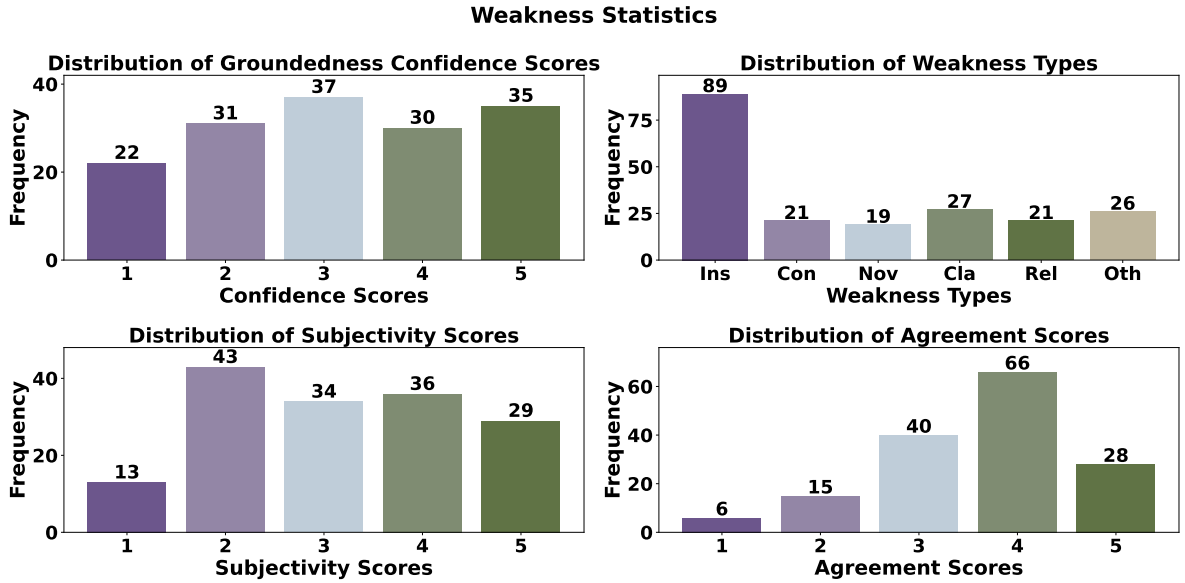

Figure 2: Distribution of the various weakness labels for CLAIMCHECK: groundedness confidence scores (top left), weakness types (top right), subjectivity scores (bottom left), and agreement scores (bottom right).

4 Experiments

To support progress on LLM-based claim-grounded review, our experiments benchmark GPT-4o in the zero-shot setting on three sets of experiments that leverage CLAIMCHECK: Claim Association (CA), Claim Verification (CV), and Weakness Labeling and Editing (WLE). Each task is motivated by a particular peer review/claim verification use case. Results are computed over all examples in CLAIMCHECK, excluding those in the pilot, unless noted otherwise. Hyperparameters and prompts for all experiments are in Appendix C and Appendix D, respectively.

4.1 Claim Association (CA)

First, we evaluate LLMs on the CA task. CA is motivated by a scenario in which a reviewer has written a weakness for a paper and would like an LLM to collect in-text citations to support it. We provide GPT-4o with a single claim-related weakness, the paper contents, and the same set of candidate claims and instructions as were given to annotators (see §3.3), and ask the model to identify up to three target claims for the provided weakness—selecting from the candidate claims and/or supplying a custom target claim as above.

Results We report the same CA metrics from before, obtaining $F_{1,edit} = 32.3$ and $F_{1,exact} = 22.7$. We note first that these values are substantially lower than the level of inter-expert agreement on the CA pilot ($F_{1,edit} = 45.8, F_{1,exact} = 28.5$). However, for a more direct comparison, we also incorporate GPT-4o's predictions on the pilot examples into the CA pilot agreement calculation, finding that (1) it noticeably reduces aggregate agreement ($F_{1,edit} = 45.8 \rightarrow 43.2, F_{1,exact} = 28.5 \rightarrow 23.3$) and (2) GPT-4o exhibits lowest *individual* average pairwise agreement among all annotators ($F_{1,edit} = 32.6$ vs. $F_{1,edit} = 37.1$ for worst human annotator)—suggesting that GPT-4o struggles to identify appropriate target claims.

4.2 Claim Verification (CV)

In reading a paper, a reviewer may come across a claim they find suspect without having to hand evidence to back up their suspicions. In such cases, it would be valuable for an LLM to *verify* the claim by supplying a grounded rationale (weakness) for it. We compare such weaknesses generated by reviewers and by GPT-4o.

Since weaknesses in CLAIMCHECK may have up to three associated target claims, we first have GPT-4o *distill* from the weakness the issues it describes relating to one claim at a time—the *focal* claim. To do this, we provide the model with a prompt containing the original weakness, the focal claim, details about the paper, and details about related work(s) cited by the weakness and focal claim. The prompt asks the model to (1) extract pieces of evidence from the paper(s) needed to assess the focal claim and (2) provide the distilled weakness

Original (Human)	Revised (GPT-4o)
In Table 2, I would like to see the performance of other correction methods (e.g., GPT3.5/4/4o) for a more comprehensive comparison.	In Table 2, the performance of other correction methods (e.g., GPT3.5/4/4o) should be included to provide a more comprehensive comparison and substantiate the claims of scalability and generalizability of the proposed solution.
Line 78: There should be some statistical significance by which the models perform poorly and then you could consider it underperformance and not just lower than actual. Models being non-deterministic, there is always a chance of slight changes in scores.	The paper should include statistical significance testing to determine if the models' underperformance is truly indicative of sandbagging rather than random variation. Given the non-deterministic nature of AI models, this would provide stronger evidence for the claims made.
What evidence (empirical results) can support the claim that 'the combination of both improvements — text splitting and iterated calls, has proven itself to perform the best' (line 146)	The paper lacks empirical evidence to support the claim that the combination of text splitting and iterated calls performs best. It would benefit from experiments or data demonstrating this improvement, such as comparative analysis with other methods or detailed performance metrics.

Table 4: Examples of (original, revised) weakness pairs from the weakness editing task (§4.3) where GPT-4o (**Revised**) fails to improve upon the specificity of the human-written (**Original**) weakness—a common occurrence in our human evaluation (See §4.3).

based on the focal claim and this evidence. We then take (2) as the reviewer's weakness for the focal claim.

To obtain the *model's* weakness for the same focal claim, we provide the claim as input in a separate prompt, along with the details of the paper and related work(s) as above. The prompt asks the model to (1) extract pieces of evidence from the paper(s) needed to assess the focal claim; (2) describe a weakness that targets that claim; and (3) perform weakness labeling on the result.

Finally, given the focal claim-related weaknesses from the reviewer and models, we use an LLM-based evaluation to determine whether these weaknesses describe exactly the *same* issue, merely *similar* issues, or entirely *different* issues with the focal claim. We provide the focal claim and the two weaknesses as input to the evaluation prompt, along with the pieces of evidence extracted for each weakness in the previous steps.

Results We use GPT-4o as the LLM judge. We find that GPT-4o-generated weaknesses for the focal claim overwhelmingly tend to be judged *different* from those identified by the reviewers (73.0%). A smaller portion of these reviews are deemed *similar* to those of the reviewers (20.0%), and an even smaller fraction are considered the same (7.0%). While *different* here does not necessarily mean *wrong*, manual inspection reveals that model-written weaknesses tend to be overly generic in their diagnoses (e.g. *"there is a lack of precise evidence linking GSNR to controlling the generalization gap as claimed"*) and sometimes make more basic errors, such as denying that the paper comments on the claim at all.

4.3 Weakness Labeling and Editing (WLE)

Our final task is motivated by the needs of *meta-reviewers* who must synthesize primary reviews. We envision that an LLM may be used to *enrich* primary reviews by providing weakness labels and by enhancing their specificity, helping the meta-reviewer more efficiently write their own review.

We provide an LLM with the full contents of the reviewed paper (full text, tables, figures, images, and captions), a reviewer-written weakness, its target claims, and the full contents of related work(s) mentioned either by the target claim(s) or the weakness. Given this information, we ask the model to provide WL annotations (weakness types and agreement and subjectivity scores) for the weakness and (if necessary) an *edited* weakness that enhances the specificity of the original.

Results: Labels We first consider the model-predicted WL annotations for weakness type (**Con**tradictory evidence, **Ins**ufficient evidence, **Nov**elty, **Rel**ated Work, **Cla**rity, **Oth**er), **Agr**eement, and **Sub**jectivity, evaluating these against the gold labels in CLAIMCHECK. Table 5 reports agreement using Krippendorff's α.

For the weakness types, we observe strong agreement for **Nov** ($\alpha_{\mathbf{Nov}} = 73.9$) and **Ins** ($\alpha_{\mathbf{Ins}} = 55.0$), consistent with our pilot results (§3.4). This is intuitive, as weaknesses of both kinds are often readily identifiable from common lexical cues (e.g. *novel(ty)*, *convincing*). Further consistent

Agr	Sub	Con	Ins	Nov	Rel	Cla	Oth
21.7	23.2	16.4	55.0	73.9	25.5	32.0	2.1

Table 5: Agreement (α) between GPT-4o and gold CLAIMCHECK agreement scores (**Agr**), subjectivity scores (**Sub**), and weakness type labels on our weakness labeling task (§4.3).

Win%	Tie%	R	BS
20.0	47.8	57.2	92.2

Table 6: Results on the weakness editing task (§4.3). WR denotes specificity win-rate: % of cases in which a human judge deemed the model-revised weakness more *specific* in its feedback than the original human one. R=ROUGE-1 F_1 w.r.t. the original weakness. BS=BERTScore F_1. Results are based on a single run.

with the pilot, GPT-4o struggles to identify when a weakness directly *contradicts* a target claim ($\alpha_{\textbf{Con}} = 16.4$) and to determine when a weakness raises issues not matching one of the primary types ($\alpha_{\textbf{Oth}} = 2.1$). Contrasts with the pilot include much lower agreement on **Rel** ($\alpha_{\textbf{Cla}} = 52.4 \rightarrow 25.5$) and substantially higher agreement on **Cla** ($\alpha_{\textbf{Cla}} = 0 \rightarrow 32.0$). For **Agr**eement and **Sub**jectivity scores, we observe somewhat higher though still modest agreement compared to the pilot ($\alpha_{\textbf{Agr}} = 18.2 \rightarrow 21.7$, $\alpha_{\textbf{Agr}} = 18.1 \rightarrow 23.2$).

Results: Edited Weaknesses Next, we compare the *texts* of the revised weaknesses with those of the original, again subsetting to weaknesses with at least one target claim. Table 6 shows results from GPT-4o. We report ROUGE-1 F_1 (Lin, 2004) and BERTScore F_1 (Zhang et al., 2019) of the revised weaknesses relative to the originals as approximate indicators of the degree of lexical (**R**) and semantic (**BS**) similarity. We observe relatively high scores on both metrics, suggesting that the revised weaknesses tend to hew fairly closely to the original texts along these two dimensions.

To evaluate specificity, we provide a human judge (one of the authors) with an (original, revised) weakness pair, along with the associated target claim(s), and ask the judge to indicate which weakness in the pair provides more specific feedback on the paper, with ties permitted. To minimize bias in the responses, we omit the provenance of each weakness (human or LLM) and also randomize the presentation order across examples.

The first column of Table 6 reports win rates of the revised weaknesses from GPT-4o against the human originals and the second column reports rates of ties. GPT-4o tends to struggle substantially to improve upon the specificity of the original weaknesses, achieving a win rate of only 20%. Empirically, we find that the model tends to make revisions that render the tone of the review more polite (e.g. by moving from first- to third-person), or that verbalize a suggestion already strongly implied in the original review, without actually providing more concrete feedback (Table 4, *top*)—an observation further reflected in the high rate of ties (47.8%). Worse, we find that both models often strip out helpful textual anchors, such as line numbers and quotation marks (Table 4, *middle*, *bottom*), making it more difficult to locate the disputed claim, and thus making the revised weakness *less* specific.

5 Conclusion

This work has introduced CLAIMCHECK—a benchmark of reviewer-identified weaknesses in NeurIPS 2023 and 2024 submissions, richly annotated with descriptive labels by experts and grounded in the *claims* that they dispute in the reviewed papers. Further, we benchmark GPT-4o on three novel tasks enabled by CLAIMCHECK—Weakness Labeling and Editing (WLE), Claim Association (CA), and Claim Verification (CV)—all aimed at assisting reviewers during the peer review process. Across these tasks, we find that GPT-4o struggles to provide specific, grounded reviews and to identify the specific claims targeted by those reviews. We release CLAIMCHECK to support further research in this direction.

Limitations

CLAIMCHECK focuses on reviewer-identified weaknesses that are plausible *claim-related*, meaning that they take issue with a particular claim or claim(s) a paper makes. While we believe this kind of weakness is among the most valuable in the peer review process, other kinds can be valuable as well. For example, weaknesses that identify important experiments or related work that were *omitted* can provide valuable feedback. Weaknesses of this sort are arguably even harder to identify than our claim-related weaknesses, and we think that empowering models to do this is an interesting direction for future work.

Moreover, the CLAIMCHECK is limited in its scale due to 1) the limited sources that satisfy all

the criteria; and 2) the intrinsic challenges in annotation even for expert-level annotators. And CLAIMCHECK is intended as purely as an *evaluation* benchmark for LLMs and LLM-based models for peer review and is likely not large enough for meaningful supervised fine-tuning.

Ethics

We do not believe this work raises any significant ethical concerns. In collecting CLAIMCHECK, we have complied with OpenReview licensing and terms of use. Further, since both the papers and the reviews in CLAIMCHECK are anonymized, there is little concern about leakage of personally identifiable information (PII).

References

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2022. PubHealthTab: A public health table-based dataset for evidence-based fact checking. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1–16, Seattle, United States. Association for Computational Linguistics.

Prabhat Kumar Bharti, Tirthankar Ghosal, Mayank Agarwal, and Asif Ekbal. 2024. Peerrec: An ai-based approach to automatically generate recommendations and predict decisions in peer review. *International Journal on Digital Libraries*, 25(1):55–72.

Prabhat Kumar Bharti, Shashi Ranjan, Tirthankar Ghosal, Mayank Agrawal, and Asif Ekbal. 2021. Peerassist: Leveraging on paper-review interactions to predict peer review decisions. In *International Conference on Asian Digital Libraries*, pages 421–435.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. APE: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7000–7011, Online. Association for Computational Linguistics.

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2023. Aries: A corpus of scientific paper edits made in response to peer reviews. *arXiv preprint arXiv:2306.12587*.

Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2024. Claim verification in the age of large language models: A survey. *arXiv preprint arXiv:2408.14317*.

Iddo Drori and Dov Te'eni. 2024. Human-in-the-loop ai reviewing: Feasibility, opportunities, and risks. *Journal of the Association for Information Systems*, 25(1):98–109.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. NLPeer: A unified resource for the computational study of peer review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.

Andrew Estornell, Sanmay Das, and Yevgeniy Vorobeychik. 2020. Deception through half-truths. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10110–10117.

Michael Fromm, Evgeniy Faerman, Max Berrendorf, Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, Yang Mao, and Thomas Seidl. 2021. Argument mining driven analysis of peer-reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4758–4766.

Yanzhu Guo, Guokan Shang, Virgile Rennard, Michalis Vazirgiannis, and Chloé Clavel. 2023. Automatic analysis of substantiation in scientific peer reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10198–10216, Singapore. Association for Computational Linguistics.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. 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 2131–2137, Minneapolis, Minnesota. Association for Computational Linguistics.

Guo-qing Jiang, Jinlong Liu, Zixiang Ding, Lin Guo, and Wei Lin. 2023. Accelerating large batch training via gradient signal to noise ratio (gsnr). *arXiv preprint arXiv:2309.13681*.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. 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 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and psychological measurement*, 30(1):61–70.

Sandeep Kumar, Tirthankar Ghosal, and Asif Ekbal. 2023. When reviewers lock horns: Finding disagreements in scientific peer reviews. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16693–16704, Singapore. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kyle Lo, Zejiang Shen, Benjamin Newman, Joseph Chang, Russell Authur, Erin Bransom, Stefan Candra, Yoganand Chandrasekhar, Regan Huff, Bailey Kuehl, Amanpreet Singh, Chris Wilhelm, Angele Zamarron, Marti A. Hearst, Daniel Weld, Doug Downey, and Luca Soldaini. 2023. PaperMage: A unified toolkit for processing, representing, and manipulating visually-rich scientific documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 495–507, Singapore. 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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7787–7813, Singapore. Association for Computational Linguistics.

Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang Wang, Qiang Liu, Shu Wu, and Liang Wang. 2024. EX-FEVER: A dataset for multi-hop explainable fact verification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9340–9353, Bangkok, Thailand. Association for Computational Linguistics.

Anku Rani, S.M Towhidul Islam Tonmoy, Dwip Dalal, Shreya Gautam, Megha Chakraborty, Aman Chadha, Amit Sheth, and Amitava Das. 2023. FACTIFY-5WQA: 5W aspect-based fact verification through question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10421–10440, Toronto, Canada. Association for Computational Linguistics.

Federico Ruggeri, Mohsen Mesgar, and Iryna Gurevych. 2023. A dataset of argumentative dialogues on scientific papers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7684–7699, Toronto, Canada. 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.

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

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2024. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022. MReD: A meta-review dataset for structure-controllable text generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2521–2535, Dublin, Ireland. Association for Computational Linguistics.

Moritz Staudinger, Wojciech Kusa, Florina Piroi, and Allan Hanbury. 2024. An analysis of tasks and datasets in peer reviewing. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 257–268, Bangkok, Thailand. Association for Computational Linguistics.

Ivan Stelmakh, Nihar B Shah, and Aarti Singh. 2021. Catch me if i can: Detecting strategic behaviour in peer assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4794–4802.

Ivan Stelmakh, John Wieting, Graham Neubig, and Nihar B Shah. 2023. A gold standard dataset for the reviewer assignment problem. *arXiv preprint arXiv:2303.16750*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. 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.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Sharanya Venkat, Richa, Gaurang Rao, and Bhaskarjyoti Das. 2022. Liarx: A partial fact fake news data set with label distribution approach for fake news detection. In *Innovations in Computational Intelligence and Computer Vision: Proceedings of ICICV 2021*, pages 221–229. Springer.

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 and Kyle Lo. 2021. Overview and insights from the SCIVER shared task on scientific claim verification. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 124–129, 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.

Nancy X. R. Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. SemEval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (SEM-TAB-FACTS). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 317–326, Online. Association for Computational Linguistics.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. ReviewRobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 384–397, Dublin, Ireland. Association for Computational Linguistics.

Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2608–2621, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

A Dataset Details

A.1 Licensing and Terms of Use

The papers and reviews included in CLAIMCHECK are all obtained from OpenReview and our use of them is consistent with the OpenReview terms of use: https://openreview.net/legal/terms. Upon paper acceptance, we will release CLAIMCHECK under a [CC-BY 4.0] license, which is also consistent with these terms.

A.2 Data Preprocessing

We use `GPT-4o-2024-08-06` with zero-shot prompting and `temperature=1.0` for full-text extraction, text cleaning, caption extraction and topic classification. See Appendix D for the respective prompts.

We filter reviews to contain at least one claim-related keywords from the list: (see **Claim-related Keywords** on the next page.)

B Annotation Details

B.1 Annotator Demographics

A total of six annotators were involved in the annotation process. Five are Ph.D. students in AI/NLP and one is a full-time NLP research scientist–all fluent speakers of English. None of these individuals received compensation beyond their recognition as co-authors of this work.

B.2 Annotation Interface

B.3 Further Annotation Details

This section provides some additional details about the annotation process. Annotation instructions are included in the supplementary materials.

Weakness Groundedness Labels Below are descriptions of each value on the ordinal groundedness labeling scale used during the WI annotation subtask.

0. Not an actual scale value (DO NOT USE); included only for reference. This value is reserved for spans of text you aren't even inclined to highlight as potential claim-related weaknesses in the first place. This would include weaknesses that very clearly do not target a claim or result (e.g. those that call out poor style or unclear exposition) or other spans that don't describe a weakness at all (e.g. spans that summarize related work or that pose a clarifying question).

1. The weakness *seems* to be responding to some claim or result in the paper (and thus is not a 0), but it's unlikely ($< 25\%$ chance) you'd be able to find actual claims in the paper that you would consider at all targeted by this weakness. This could be because the weakness is highly subjective or because the reviewer makes lots of inferences not grounded in the paper's contents.

2. Like (1), but you think it's somewhat likelier (25-50% chance) that you'd be able to find at least some claim or result in the paper targeted by this weakness.

3. The weakness makes reference to a claim that is plausibly grounded in the paper, but that is not an explicit quote or not an obvious paraphrase. You would likely (50-75% chance) be able to find a claim or claims targeted by this weakness in the paper, but the actual claims discussed in the weakness might reflect a modest amount of interpretation on the part of the reviewer, and, further, might be made on the basis of figures, tables, or numerical results rather than claims *per se*.

4. Like (3), but you are *quite* confident ($> 75\%$ chance) that you would be able to find target claims for this weakness in the paper. The claims referenced in the weakness involve minimal interpretation on the part of the reviewer and are very closely grounded either in claims from the paper and/or in figures, tables, or numerical results.

5. The weakness explicitly (partially) quotes or otherwise makes explicit reference (e.g. via paraphrase) to a specific *claim*—not figure, table, or raw numerical result—that is almost certainly made in the paper (assuming the reviewer is not a blatant liar). These spans may start with (e.g.) "the paper claims that..." or "the authors state that...", or may refer to specific line numbers that contain the claim of interest.

Weakness Objectivity Labels The objectivity score is an ordinal score (1-5) for how objective the criticism raised by a particular weakness is. Below are the interpretations of scores 1, 3, and 5 as given to annotators, where scores of 2 and 4 are to be interpolated on the basis of these descriptions.

Claim-related Keywords

["overclaim", "over-claim", "over claim", "claim", "claims", "claiming", "claimed", "supported", "fully support", "fully supporting", "fully supportive", "supported", "support", "supporting", "substantiate", "substantiating", "substantiated", "convincing", "convince", "convincingly", "convinces", "supportive", "unsubstantiated", "unsubstantiated", "unsupported", "unverified", "unverified", "unverifiable", "unverifiable",]

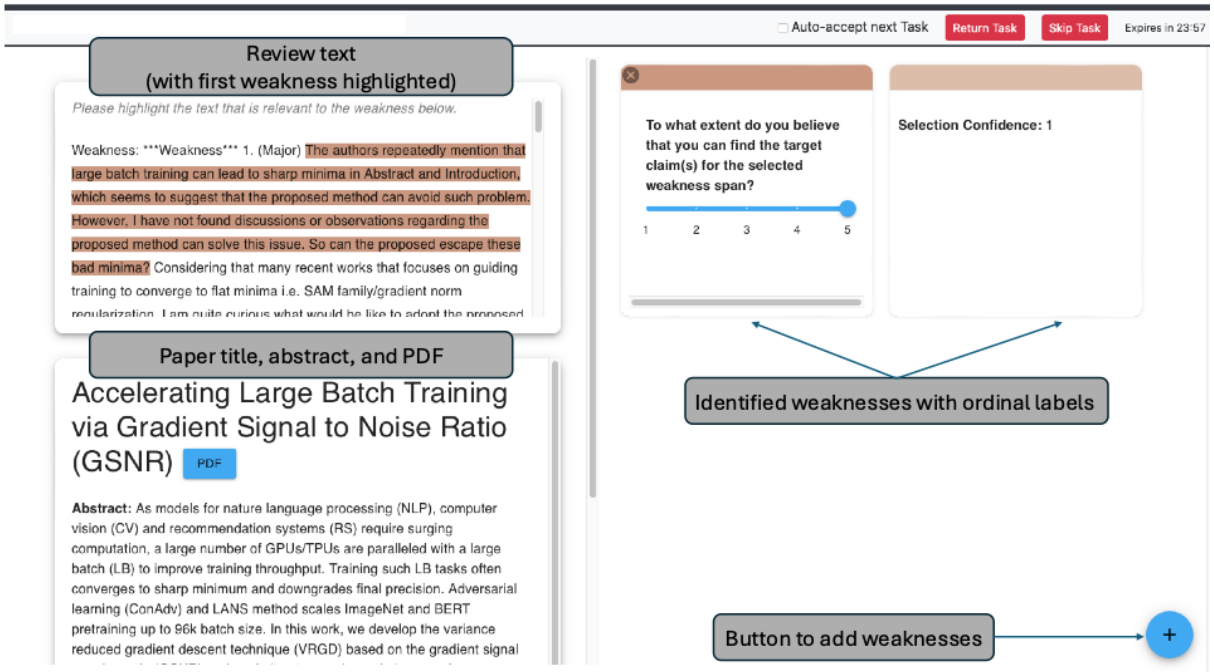

Figure 3: Annotation interface for the **Weakness Identification (WI)** subtask. Annotators select contiguous spans from from the review text (top left), each describing a weakness raised by the reviewer. For each weakness, annotators supply a Likert-scale judgment (top right) indicating the extent to which they believe the weakness targets a *specific claim* made in the paper (bottom left). Annotators select as many weaknesses as they can find in the review that plausibly target *some* claim. The paper in this example (and in Figures 4-6) is Jiang et al. (2023).

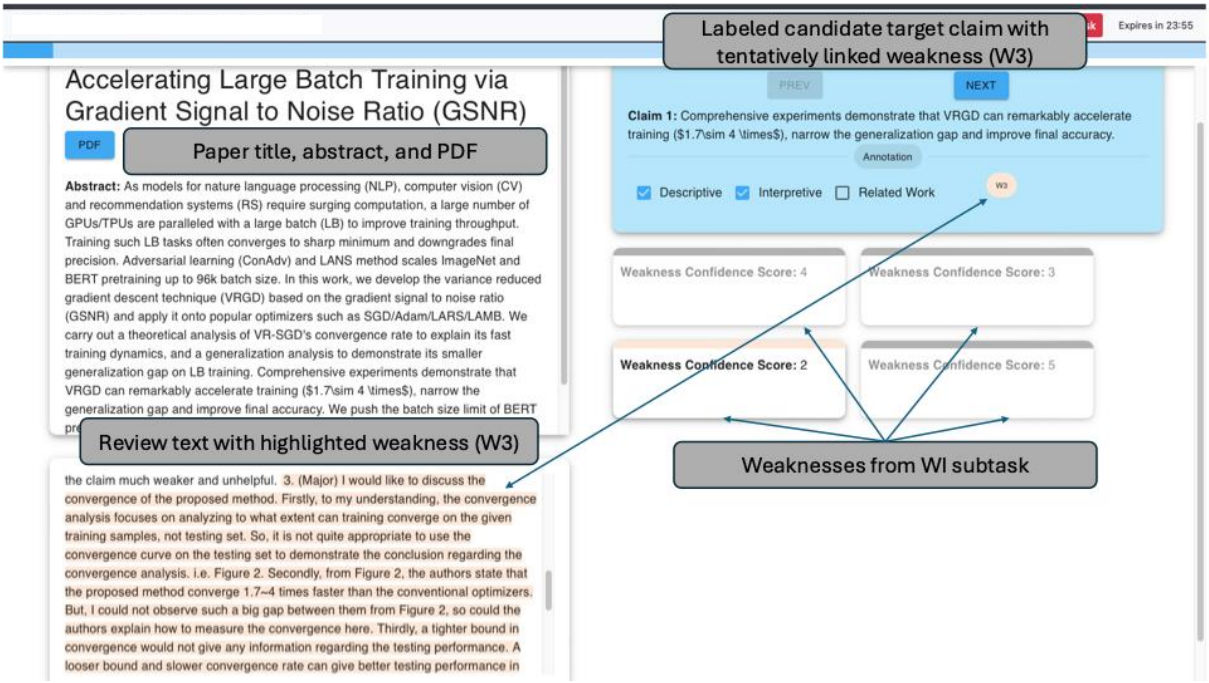

Figure 4: Annotation interface showing part of the **Claim Association (CA)** subtasks. Given (1) the weaknesses identified for a given review during the Weakness Identification (WI) subtask (Figure 3) and (2) a set of candidate claims extracted by GPT-4o, annotators must determine which of these claims are targeted by each weakness (if any). Although during the annotation we also ask annotators to provide type labels for each candidate target claim, we find these labels do not provide necessary information for other annotation subtasks or for LLM reasoning and decide to drop it from the final dataset/evaluation.

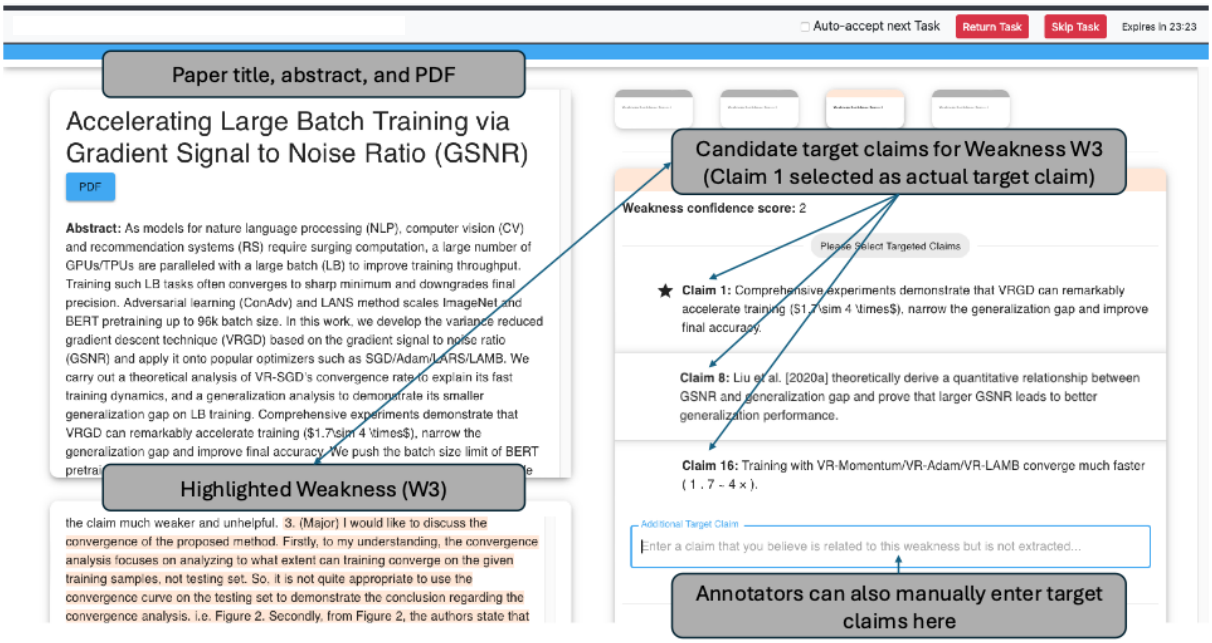

Figure 5: Annotation interface for the final part of the **Claim Association (CA)** subtask. After selecting a set of *tentative* target claims for each weakness (Figure 4), annotators then *finalize* their selections by starring a (potentially improper) subset of these claims (here, **Claim 1**). Additionally, they may manually add a target claim from the text if it was not among the extracted candidate claims (bottom right).

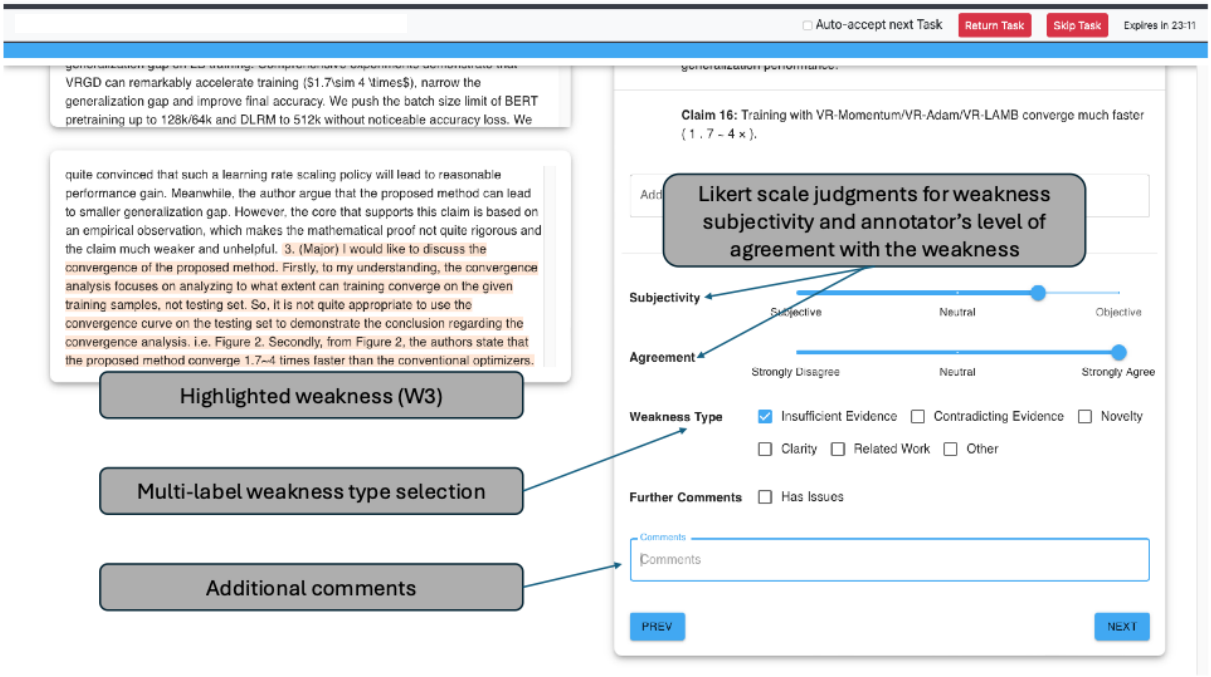

Figure 6: Annotation interface for the **Weakness Labeling (WL)** subtask. After finalizing the set of target claims for a given weakness (Figure 5), annotators label these weaknesses by providing: (1) a *subjectivity* rating, indicating how subjective the annotator believes the weakness to be; (2) an *agreement* rating, indicating the extent to which the annotator agrees that the weakness is valid; and (3) a multi-label set of *weakness types*, indicating the kind of weakness this is. Annotators may also leave further comments about the weakness in the text box at the bottom.

1. The claim-related weakness depends almost exclusively on subjective judgments about one or more aspects of the paper, such as how significant or exciting its contributions are, its novelty, likely impact, ethical implications, etc.

3. The claim-related weakness depends on objective observations or judgments but also includes some subjective interpretations of, or opinions about, those observations and their implications.

5. The claim-related weakness depends almost exclusively on objective observations (possibly in conjunction with valid commonsense, mathematical, logical, or statistical reasoning), with limited or no appeal to subjective interpretation of the paper's claims or contributions.

Weakness Agreement Labels the agreement score is an ordinal score (1-5) for a weakness that represents the the extent to which an annotator agrees that the issue raised by the weakness is a problem for the paper. As with the objectivity labels, we provided annotators with descriptions for scores of 1, 3, and 5, with the interpretations of scores of 2 and 4 to be interpolated on the basis of these descriptions.

1. The claim-related weakness makes no sense, is ill-founded, or simply does not apply to any claims made in the paper.

3. The claim-related weakness is somewhat convincing and/or partially applicable to the target claims.

5. The claim-related weakness is fully convincing and directly applicable to the target claims. The target claims would need to be heavily revised or even jettisoned entirely in response to the weakness.

Weakness Type Labels Below are the descriptions of the multi-label weakness types as provided to annotators. As with the claim types (see above), our preliminary investigations revealed that a substantial fraction of weaknesses were adequately characterized only by two or more of these labels (e.g. weaknesses that call the *novelty* of some method into question based on very similar proposals in uncited *related work*). Thus, we were similarly motivated to implement a multi-label typing scheme here.

- *Insufficient Evidence*: The weakness argues that the paper provides insufficient evidence for some claim(s)—e.g. due to lack of statistical significance testing, missing experiments, weak baselines, inappropriate choice of datasets, etc.

- *Contradictory Evidence*: The weakness provides evidence that some claim(s) in the paper are not only insufficiently supported but are in fact false—e.g. due to numerical or methodological errors or results in another paper (see *related work*) that undermine the paper's claims of state-of-the-art performance.

- *Novelty*: The weakness claims that the paper is not novel in one or more important respects.

- *Clarity*: The weakness highlights difficulties in understanding the paper itself—possibly due to poor writing or paper organization.

- *Related Work*: The weakness calls attention to other work related to the paper that was uncited or otherwise given inadequate consideration or treatment.

- *Other*: The weakness identifies some issue with the paper that does not clearly belong to one of the other categories described above.

C Experimental Details

C.1 Model Details and Hyperparameters

We run all the experiments (WLE, CA, CV, and reviewer-written weakness grounding for WIS evaluation) with `GPT-4o-2024-08-06` zero-shot prompting. We use `temperature=0.9` for CA and `temperature=0.3` for all the other experiments. All the experiments are repeated 3 times with `seeds=[0,42,2025]` and we report the average results across the three runs.

We provide the prompts for all the experiments in Appendix D

D Prompts

Prompts used in data preprocessing and experiments.

Claim Extraction

You are an experienced AI and NLP researcher that is going to review a paper. Given the title, abstract, and a chunk of text in the paper, your first task is to extract all the scientific claims the authors make in this chunk. The claims should be a consecutive span of text from the sections and consists of one or more sentences. Make sure to extract the exact original claims from the text, without any paraphrasing. When extracting claims, focus on claims that are with respect to the findings/contributions/results/relation with related work of the research, skip all other claims, especially ignore any descriptions of the ideas, methods, and experiment setup. If the chunk contains no claim satisfies the criteria, simply output an empty list. There might be some noisy text in the chunk, such as ocr text from figures, references, due to the noise in parsing the paper pdf.Ignore and only ignore the noisy text, extract the claims from the rest of the text. You can determine if a part of the chunk is noisy by referring to the title and abstract. Output your results as a JSON object with the following format: {Claims: ['Claim 1', 'Claim 2', ...]}, where the claims are listed in the order they appear in the text.

Caption Extraction

Given an image of a table/figure/algorithm from a paper, your task is to extract the caption of the image.The is caption usually located above or below the image, and starts with 'Table X:', 'Figure X:', or 'Algorithm X:', where X is the index of the image.Output your results as a JSON object with the following format: {"Caption": "The caption of the image"}

Text Cleaning

You are an expert in AI/NLP. Given a paragraph extracted from an AI/NLP paper using OCR, your task is to clean the text by removing OCR noises. Specifically, the paper are extracted from NeurIPS2023/2024 anonymized submissions, so OCR will identify the line numbers and embed them in the content text. Additionally, there might be text from tables/figures/captions that are accidentally included in the main text due to OCR error. Your task is to clean these noise strings from the text. Keep the substring such as "' that represents "s'. And for all the numbers encoded in brackets, e.g. [20] are in-line citation, only remove them if they are within the span that you determine are wrong extraction from table/figure/captions. Use your knowledge to determine which parts are noise and which parts are original text, based on fluency and coherence. Especially when determining when mentioning tables/figures/captions is intended in the main content or are OCR errors. Do not modify any of the original text, instead, copy them faithfully. Output the cleaned text.Output your results as a JSON object with the following format: {'cleaned_text': 'The cleaned text'}

NLP Topic Classification

You are an experienced AI and NLP researcher that is going to serve as the program chair for a top AI conference. Given a paper title and abstract, and list of keywords, you job is to determine if the paper is broadly relevant to natural language processing (NLP) research. A paper is broadly related to NLP if it's any part of its topic/subject matter/methods/techniques/data and resource use/evaluation is related to any subfield of NLP. Output your results as a JSON object with the following format: {"NLP": "YES/NO"}, where YES indicates the paper is broadly related to NLP, NO indicates the paper is not related to NLP. .

Weakness Labeling and Editing (WLE)

You are an experienced AI and NLP researcher that is going to give meta reviews. You are provided with:
1. The full text, as well as its tables, figures and algorithms as images with captions (if any) of the main paper 2. (Optionally) The full text, as well as its tables, figures and algorithms as images with captions of one or more related work 3. A review that comments on some weaknesses of the paper. 4. A span of text extracted from the review that is potentially a **claim-related weakness**. 5. One or more claims from the paper that are **target claim(s)** of the claim-related weakness.
A claim-related weakness is a span of text in the provided review that specifically comments on shortcomings of the paper, usually with reference to particular claims the paper makes. A claim is said to be a target claim of a claim-related weakness if:
1. The weakness clearly disputes the truth or accuracy of that claim. 2. Making this determination does not require appealing to any other claim(s).
Your tasks are to:
1. Give an ***objectivity score*** for the claim-related weakness. 2. Give an ***agreement score*** for the claim-related weakness. 3. Assign one or more ***weakness type label(s)*** to the claim-related weakness. 4. If needed, rewrite the claim-related weakness to make it more sound based on your understanding of (a) the paper and optionally related work, (b) the target claim(s), and (c) the original claim-related weakness.
The objectivity score is an ordinal score (1-5) for the claim-related weakness that represents the degree of objectivity involved in the judgments of the agreement annotation. The interpretations of the values 1, 3, and 5 on this scale are as follows:
1: The claim-related weakness depends almost exclusively on subjective judgments about one or more aspects of the paper, such as how significant or exciting its contributions are, its novelty, likely impact, ethical implications, etc. 3: The claim-related weakness depends on objective observations or judgments but also includes some subjective interpretations of, or opinions about, those observations and their implications. 5: The claim-related weakness depends almost exclusively on objective observations (possibly in conjunction with valid commonsense,mathematical, logical, or statistical reasoning), with limited or no appeal to subjective interpretation of the paper's claims or contributions.
A score of 2 should be based on an "interpolation" between the descriptions for 1 and 3 above and a score of 4 should be based an "interpolation" between the descriptions for 3 and 5 above.
Next, the agreement score is an ordinal score (1-5) for the claim-related weakness that represents the the extent to which you would agree with its content if you were the meta-reviewer for the paper. The interpretations of the values, 1, 3, and 5 on this scale are as follows:
1: The claim-related weakness makes no sense, is ill-founded, or simply does not apply to any claims made in the paper. 3: The claim-related weakness is somewhat convincing and/or partially applicable to the target claims. The associated target claims would need to be qualified or rephrased in response to the weakness. 5: The claim-related weakness is fully convincing and directly applicable to the target claims. The target claims would need to be heavily revised or even jettisoned entirely in response to the weakness.
As with the objectivity score, a score of 2 should be based on an "interpolation" between the descriptions for 1 and 3 directly above and a score of 4 should be based an "interpolation" between the descriptions for 3 and 5 directly above.

Weakness Labeling and Editing (WLE) (Continued)

Finally, the weakness type labels characterize the kind of claim-related weakness we are dealing with. Multiple labels may apply and you must select at least one. The labels are as follows:
- Insufficient Evidence (insufficient): The weakness argues that the paper provides insufficient evidence for some claim(s)—e.g. due to lack of statistical significance testing, missing experiments, weak baselines, inappropriate choice of datasets, etc. - Contradictory Evidence (contradictory): The weakness provides evidence that some claim(s) in the paper are not only insufficiently supported but are in fact false—e.g. due to numerical or methodological errors or results in another paper that undermine the paper's claims of state-of-the-art performance. - Novelty (novelty): The weakness claims that the paper is not novel in one or more important respects. - Clarity (clarity): The weakness highlights difficulties in understanding the paper itself—possibly due to poor writing or paper organization. - (Missing) Related Work (related_work): The weakness calls attention to other work related to the paper that was uncited or otherwise given inadequate consideration or treatment. - Other (other): The weakness identifies some issue with the paper that does not clearly belong to one of the other categories described above.
Your output must be a JSON object with the following format: {"Reasoning Objectivity": "Your reasoning for the objectivity score", "Objectivity Score": "The objectivity score", "Reasoning Agreement": "Your reasoning for the agreement score", "Agreement Score": "The agreement score", "Reasoning Weakness Type": "Your reasoning for the weakness type label(s)", "Weakness Types": {"insufficient": true/false, "contradictory": true/false, "novelty": true/false, "clarity": true/false, "related_work": true/false, "other": true/false}}"Reasoning Rewritten Weakness": "Your reasoning for if the claim-related weakness span needs to be rewritten and how", "Rewritten Weakness": "The claim-related rewritten weakness span"}

Claim Association (CA)

You are an experienced AI and NLP researcher that is going to write meta-reviews. You are provided with:
1. The full paper text and a numbered list of claims that have been extracted from the paper. 2. A review that comments on some weaknesses of the paper. 3. A span of text extracted from the review that is potentially a **claim-related weakness**. A claim-related weakness is a span of text in the above review that specifically comments on shortcomings of the paper, usually with reference to particular claims the paper makes. 4. A weakness confidence score: An ordinal label (1-5) indicating how likely you think it is that the claim-related weakness has at least one **target claim** in the paper.
Your tasks is to : Select a subset of claims from the provided claim list that are **target claims** of the claim-related weakness. A claim is said to be a target claim of a claim-related weakness if: 1. The weakness clearly disputes the truth or accuracy of the claim. 2. Making this determination does not require appealing to any other claim(s).
Concerning point (2), a weakness, if true, can clearly have implications for the truth or accuracy of multiple claims made by a paper. But for our purposes, we want to focus only on the claims that are most immediately disputed, which is why we stipulate (2) above. We might therefore distinguish ***direct target*** claims from ***indirect target*** claims—claims whose truth or accuracy is affected by some weakness (if true), but only in virtue of other claims. We illustrate this distinction with the example below.
Example 1: ———- Weakness 1: The paper's claim that method X demonstrates superior performance over all baselines is not convincing since the confidence intervals of X's performances largely overlap with many of the baselines' confidence intervals. Claim 1: The results in Table 2 demonstrate the superior performance of proposed method X over all the existing baselines on dataset A. Claim 2: Findings from Table 1, 2, and 3 showcase the effectiveness of the proposed method X on task T.
Explanation: Here, Claim 1 is a direct target of Weakness 1, since Claim 1's veracity is directly disputed by Weakness 1, and one need not appeal to any other claims to see that this is so. In contrast, Claim 2 is an indirect target of Weakness 1, since Weakness 1 undermines Claim 2, but only by virtue of Claim 1. You should therefore annotate only Claim 1 as a (direct) target claim.
Another important distinction in target claim association annotation is the one between ***direct target*** claims and merely ***relevant*** claims. You should ***NOT*** associate claims that are merely relevant to some weakness. The following example illustrates this second distinction.
Example 2: ———- Weakness 2: While the paper claims the introduced module Y enhances the robustness of model M under realistic types of noise, the only datasets that the paper experiments on—i.e. B and C—are either synthetic or make heavily simplifying assumptions about the noise distribution. More realistic datasets like D should also be considered. Claim 3: Experimental results demonstrate the effectiveness of proposed module Y that renders model M more robust against realistic noise. Claim 4: As shown in Figure 3 and 4, adding Y to M helps improve the robustness of M under various kinds of noise presented in dataset B and C.
Explanation: Here, Claim 3 is clearly a direct target of Weakness 2. But Claim 4, although topically relevant to Weakness 2, is not a direct target. Even though it refers to datasets B and C, which are mentioned in Weakness 2, Claim 4 is not undermined by Weakness 2 and therefore should not be associated with it.

Claim Association (CA) (Continued)

For cases where a weakness quotes or mentions a particular claim (principally, weaknesses with a label of 5), the target claim will generally be quite easy to identify. Beyond this, target claims can be trickier to identify, but here are some general principles:
- Take your cue from what the weakness is about. If the weakness is about novelty, an appropriate target claim really ought to be one that makes some assertion about, or else strongly implies, novelty. Or if the weakness is about the superiority of a proposed method relative to existing methods, you ought to be able to find a claim to that effect (or one that strongly implies that superiority) in the paper—not just a table with results. This is a fairly basic point, but the moral is that if the paper doesn't actually make the claim imputed to it by the weakness, then that weakness might just not have a target claim. Don't go scrounging for target claims that aren't there. - Relatedly, if the weakness is very broad or vague (typically, these will have a label of 1 or 2), then they probably don't have a target claim either. - However, if you think the claim-related weakness should have a target claim but you cannot find one in the list of claims, you may copy up to one additional target claim from the paper text (called a **custom target claim**). You should always use this option if a weakness quotes or mentions a claim in the paper that does not appear anywhere in the list of candidate claims. - Additionally, even if the weakness does *not* explicitly quote or mention a specific claim, you may still be able to find a target claim in the paper. You should use a ***custom target claim*** in this situation as well—especially if the ordinal label score is relatively high (3-5) for the weakness but you are struggling with finding a proper target claim in the list of claims. You should select ***AT MOST 3*** target claims for the weakness, including the custom claim (if you use one).
Your output must be a JSON object with the following format: {"Reasoning": "Your reasoning about why the selected claim(s) are the target(s) of the weakness span, and their labels" "Target claims extracted": ["Claim X: ...", "Claim Y: ..."] (the target claim extracted from the list of claims, if any. Copy the original claim text and the claim number. Leave empty if no target claim is identified.), "Custom target claim": "The custom target claim you extract (if you extract one)}",

Claim Verification (CV)

You are an experienced AI and NLP researcher that is going to give meta reviews. You are provided with:
1. The full text, as well as its tables, figures and algorithms as images with captions (if any) of the main paper 2. (Optionally) The full text, as well as its tables, figures, and algorithms as images with captions (if any) of one or more related work 3. One target claim from the paper.
Your task is to: Determine whether the target claim exhibits one or more of the following types of weakness:
- Insufficient Evidence (insufficient): the paper provides insufficient evidence for the target claim, e.g. due to lack of statistical significance tests, missing experiments, weak baselines, inappropriate choice of datasets, etc. - Contradictory Evidence (contradictory): the target claim in the paper is not only insufficiently supported but is in fact false, e.g. due to numerical or methodological errors or results in another paper. - Novelty (novelty): novelty asserted in the target claim is not valid in one or more important respects. - Clarity (clarity): the claim is difficult to understand, possibly due to poor writing or paper organization. - Related Work (related_work): the claim fails to take into account critical prior work related to the claim. - Other (other): there are some other weakness(es) in the target claim that are not covered by any of the above categories.
The target claim definitely exhibits AT LEAST ONE of these types of weakness and may exhibit multiple.You will then need to extract all relevant pieces of evidence from the paper (and related work if any), which may include statements in the text, figures, tables, or algorithms. You must assess ONLY the target claim. DO NOT try to asses any other claims in the paper for weaknesses. This means that when you extract evidence, you must focus ONLY on pieces of evidence that are relevant to the target claim. You must also label each piece of evidence you extract as follows:
- Label each piece of textual evidence as <text_0>, < text_1>, ..., based on their order of occurrence in the paper. - Label each piece of figure/table/algorithm evidence as <figure_x>, <table_y>, <algorithm_z> ..., where x, y and z are the indices of the figures/tables/algorithms as given in their captions.
In your output, you must also explain your REASONING for the types of weakness you think the target claim exhibits. When explaining your reasoning, you should explicitly cite relevant pieces of evidence that you extracted. For example: 'Based on ... in <text_0> and ... in <figure_1> and ... in <algorithm_3>, ... the target claim exhibits...' Please output your results as a JSON object with the following format:
{"Main Paper Evidence": {"Text Evidence": {"text_0": "The piece of evidence text", "text_1": "The piece of evidence text", ...}, "Figure Evidence": ["figure_x", ...,], "Table Evidence": ["table_y", ...,], "Algorithm Evidence": ["algorithm_z", ...,]}, "Related Work 1 Evidence": {same as above, if provided and needed }"Related Work 2 Evidence": {same as above, if provided and needed }"Reasoning": "Your reasoning about the weaknesses exhibited by the target claim, specifically stating what part(s) of the target claim exhibit weakness(es) and why", "Weaknesses": {"insufficient": true/false, "contradictory": true/false, "novelty": true/false, "clarity": true/false, "related_work": true/false, "other": true/false}}

Claim Verification (CV) Distill Reviewer-Written Claim-Related Weakness

You are an experienced AI and NLP researcher that is going to give meta reviews. You are provided with:
1. The full text, as well as its tables, figures and algorithms as images with captions (if any) of the main paper 2. (Optionally) The full text, as well as its tables, figures, and algorithms as images with captions of one or more related work 3. A review that comments on some weaknesses of the paper. 4. A span of text extracted from the review that is potentially a **claim-related weakness**. 5. The types for the claim-related weakness listed above, which might be one or more of the following: - Insufficient Evidence (insufficient): The weakness argues that the paper provides insufficient evidence for some claim(s)—e.g. due to lack of statistical significance testing, missing experiments, weak baselines, inappropriate choice of datasets, etc. - Contradictory Evidence (contradictory): The weakness provides evidence that some claim(s) in the paper are not only insufficiently supported but are in fact false—e.g. due to numerical or methodological errors or results in another paper that undermine the paper's claims of state-of-the-art performance. - Novelty (novelty): The weakness claims that the paper is not novel in one or more important respects. - Clarity (clarity): The weakness highlights difficulties in understanding the paper itself—possibly due to poor writing or paper organization. - (Missing) Related Work (related_work): The weakness calls attention to other work related to the paper that was uncited or otherwise given inadequate consideration or treatment. - Other (other): there are some other weakness(es) in the target claim that are not covered by any of the above categories. 6. One claim from the paper that is the **target claim** of the claim-related weakness.
A claim-related weakness is a span of text in the provided review that specifically comments on shortcomings of the paper, usually with reference to particular claims the paper makes. A claim is said to be a target claim of a claim-related weakness if:
1. The weakness clearly disputes the truth or accuracy of that claim. 2. Making this determination does not require appealing to any other claim(s).
Your task is to:
Provide the underlying REASONING of the claim-related weakness and its type by grounding it to pieces of evidence from the main and related work (if any), Specifically, explain your REASONING for the types of weakness you think the target claim exhibits, based on the claim-related weakness.based on your understanding of the main paper and optionally related work, (b) the target claim, and (c) the original claim-related weakness and the types of weaknesses the claim-related weakness exhibits. You will first need to extract all relevant pieces of evidence from the paper (and related work if any), which may include statements in the text, figures, tables, or algorithms. You must center your REASONING ONLY around the (target claim, claim-related weakness) pair. DO NOT try to asses any other claims in the paper for weaknesses, or try to justify any other parts of the review. This means that when you extract evidence, you must focus ONLY on pieces of evidence that are relevant to the (target claim, claim-related weakness) pair. You can and should elaborate more in your reasoning through grounding the claim-related weakness in evidence, especially if the original claim-related weakness is too broad with detailed reasoning omitted. In the meantime you must try your best to reflect the original meaning conveyed by the claim-related weakness. But you must try not to directly quote or copy the claim-related weakness, essentially, the REASONING should be a standalone justification of the claim-related weakness.

Claim Verification (CV) Distill Reviewer-Written Claim-Related Weakness (Continued)

You must also label each piece of evidence you extract as follows:
- Label each piece of textual evidence as <text_0>, < text_1>, ..., based on their order of occurrence in the paper. - Label each piece of figure/table/algorithm evidence as <figure_x>, <table_y>, <algorithm_z> ..., where x, y and z are the indices of the figures/tables/algorithms as given in their captions.
When explaining your reasoning, you should explicitly cite relevant pieces of evidence that you extracted. For example: 'Based on ... in <text_0> and ... in <figure_1>, ... the target claim exhibits...' Please output your results as a JSON object with the following format:
{"Main Paper Evidence": {"Text Evidence": {"text_0": "The piece of evidence text", "text_1": "The piece of evidence text", ...}, "Figure Evidence": ["figure_x", ...,], "Table Evidence": ["table_y", ...,], }"Algorithm Evidence": ["algorithm_z", ...,]}, "Related Work 1 Evidence": {same as above, if provided and needed }"Related Work 2 Evidence": {same as above, if provided and needed }"Reasoning": "Your reasoning about the weaknesses exhibited by the target claim based on the claim-related weakness provided, specifically stating what part(s) of the target claim exhibit weakness(es) and why",