

# REVIEWSCORE: MISINFORMED PEER REVIEW DETECTION WITH LARGE LANGUAGE MODELS

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Peer review serves as a backbone of academic research, but in most AI conferences, the review quality is degrading as the number of submissions explodes. To reliably detect low-quality reviews, we define *misinformed* review points as either “weaknesses” in a review that contain incorrect premises, or “questions” in a review that can be already answered by the paper. We verify that 15.2% of weaknesses and 26.4% of questions are *misinformed* and introduce REVIEWSCORE indicating if a review point is *misinformed*. To evaluate the factuality of each premise of weaknesses, we propose an automated engine that reconstructs every *explicit* and *implicit premise* from a weakness. We build a human expert-annotated REVIEWSCORE dataset to check the ability of LLMs to automate REVIEWSCORE evaluation. Then, we measure human-model agreements on REVIEWSCORE using eight current state-of-the-art LLMs and verify moderate agreements. We also prove that evaluating premise-level factuality shows significantly higher agreements than evaluating weakness-level factuality. A thorough disagreement analysis further supports a potential of fully automated REVIEWSCORE evaluation.<sup>1</sup>

## 1 INTRODUCTION

Peer review is an evaluation of academic work by experts to assess its originality, significance, and validity before publication (Kelly et al., 2014; Taylor & Francis, 2025). In AI conferences, as the number of submissions is exponentially increasing, required number of reviewers is also exploding. As a result, review quality is degraded, which undermines the integrity and reliability of a peer-review system (Stelmakh et al., 2021; Cortes & Lawrence, 2021; Shah, 2022; Kim et al., 2025). Due to the importance of the issue, previous works propose criteria for evaluating review quality. However, we observe a trade-off between applicability to reviews and specificity of rubrics. Goldberg et al. (2025) and Du et al. (2024) proposes criteria that could apply to nearly every review but those are quite vague and subjective. Guo et al. (2023), Sadallah et al. (2025), Purkayastha et al. (2025), and Ou et al. (2025) introduce specific and objective criteria but those target narrow scope of reviews. If we could reliably measure review quality based on specific yet applicable criteria, we could filter out misinformed reviews or give feedback to reviewers before the reviews are released.

To resolve this issue, we introduce two specific yet applicable criteria of a review quality: *unanswerability of questions* and *factuality of weaknesses*. To select the criteria, we recruited a group of human experts and let them independently analyze a small subset of ICLR reviews<sup>2</sup>. Specifically, each of them decomposed a review into *review points*, which are formally defined as follows.

**Definition 1** (Review Point). A review point is a single, self-contained unit of evaluation or inquiry in a review—either a weakness or a question—that stands on its own semantically.

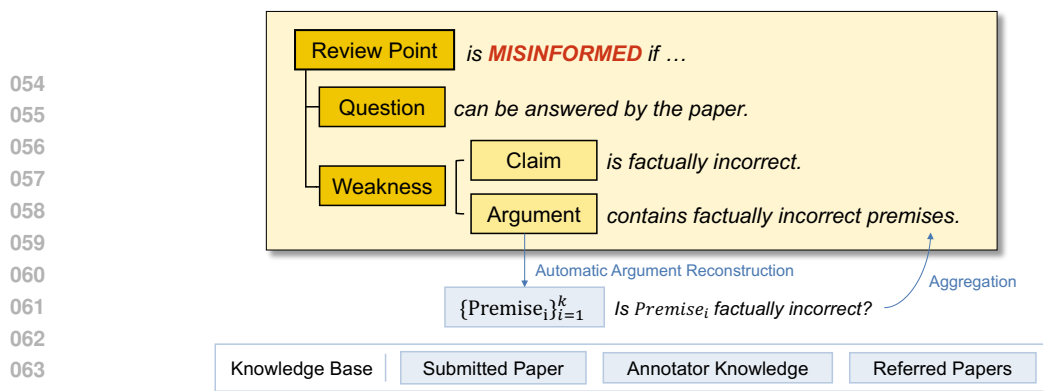
Each human evaluated quality of review points, and they discussed trustworthy criteria to detect low-quality. Based on the discussion, we formally define a *misinformed* review point as follows.

**Definition 2** (Misinformed Review Point). A review point is *misinformed* if and only if

- a question stated in a review can be already answered by the paper, or
- a weakness stated in a review is incorrect or contains incorrect premises regarding the paper.

<sup>1</sup>We will make the source code and dataset publicly available.

<sup>2</sup>A group of three graduate students studying AI analyzed reviews of ICLR submissions available in OpenReview. Detailed process of this group work is described in Appendix B.



065 Figure 1: Overview of REVIEWSCORE. Each review point in a review is categorized into question  
 066 and weakness. We further categorize weakness into claim and argument by the presence of support-  
 067 ing reasons. Based on an appropriate knowledge base, if a question is answerable by the paper, a  
 068 claim is factually incorrect, or an argument contains factually incorrect premises, then the review  
 069 point is *misinformed*. For arguments, to extract all *explicit* and *implicit premises*, we also introduce  
 070 an automatic argument reconstruction engine.

071 In this work, the human annotation shows that 26.4% of questions and 15.2% of weaknesses are  
 072 *misinformed*, which means that the current AI conference reviews contain considerable amount of  
 073 *misinformed* review points.<sup>3</sup> We also note that these criteria aligns with reviewer guidelines of major  
 074 AI conferences. For instance, ACL 2023 Peer Review Policies indicate that “before writing a  
 075 negative review, check whether your questions are already answered.” (Boyd-Graber et al., 2023),  
 076 and NeurIPS 2025 reviewer guidelines indicate that “reviewers should minimize the chance of mis-  
 077 understandings during the reviewing process” (NeurIPS 2025 Program Chairs, 2025).

078 Based on these observations, we define REVIEWSCORE that indicates if a review point is *misin-*  
 079 *formed* (Section 3.1). First, we define BASE REVIEWSCORE by directly applying the definition of  
 080 *misinformed* review points in a 5-point scale. However, we discovered human-annotated factuality  
 081 often diverges. This is because a weakness often contains both correct and incorrect *explicit premises*  
 082 or incorrect *implicit premises*, which hinders humans to reliably annotate factuality. To resolve this  
 083 issue, we further define ADVANCED REVIEWSCORE as an aggregation of premise factuality scores.  
 084 If a weakness has no supporting reason, we call it as a *claim*, and if a weakness consists of more than  
 085 one premises, we call it as an *argument*. We adopt two aggregation methods, logical conjunction,  
 086 following the literature of logic (Beall et al., 2024), and weighted average, maintaining a 5-point  
 087 scale (Figure 1). Before evaluating ADVANCED REVIEWSCORE, we have to extract all *explicit* and  
 088 *implicit premises* from an argument. This process is called *argument reconstruction* in logic and  
 089 critical thinking (Brun & Betz, 2016; Groarke, 2024; Dowden, 2024), and we construct an engine  
 090 that automatically does this process (Section 3.2). To fully automate *misinformed* review detection  
 091 with LLMs, we should verify if model-estimated REVIEWSCORE align with human-annotated RE-  
 092 VIEWSCORE. To do that, we construct a human expert-annotated REVIEWSCORE dataset based on  
 093 ICLR reviews (Section 3.3). We recruited 15 experienced graduate students and they dedicated total  
 244 hours for trustworthy human annotation.

094 We verify the reliability of automatic evaluation of REVIEWSCORE by measuring human-model  
 095 agreements (Section 4). We use eight current state-of-the-art models, including five proprietary  
 096 and three open-sourced models, and confirm moderate agreements with the human experts. AD-  
 097 VANCED REVIEWSCORE clearly outperforms BASE REVIEWSCORE, which proves the effective-  
 098 ness of premise-level factuality scoring. We also conduct a thorough human-model disagreement  
 099 analysis and observe that models sometimes misunderstand the meaning of review points or predict  
 100 scores that minority human annotators give.

101 To summarize our contributions:

- 102 1. We introduce REVIEWSCORE, a novel evaluation criteria that detects *misinformed* review  
 103 points (i.e., questions that can be answered by the paper or weaknesses with incorrect premises).
- 104 2. To evaluate the factuality of premises, we propose an automatic argument reconstruction engine  
 105 that generates a *valid* and *faithful* set of premises and conclusion.  
 106

107 <sup>3</sup>A group of 15 graduate students annotate reviews of 40 works submitted to ICLR 2021-2023. Detailed  
 process of this human annotation is described in Section 3.3.

3. We construct a trustworthy human expert-annotated dataset to measure the reliability of automatic evaluation of REVIEWSCORE.
4. We validate the reliability of automatic REVIEWSCORE evaluation by using eight current state-of-the-art LLMs and conducting a comprehensive human-model disagreement analysis.

## 2 RELATED WORKS

**Peer review evaluation.** Previous works studied how to evaluate the quality of peer reviews. Goldberg et al. (2025) introduces four criteria, *understanding*, *substantiation*, *coverage*, and *constructiveness*. However, they focus on evaluating review as a whole, not each individual review point, which makes difficult to clarify which specific points are *misinformed*. They also do not use LLMs to automate evaluation. Guo et al. (2023) specifically focuses on evaluating *substantiation*. However, they evaluate whether each review point has supporting evidence, not whether the supporting pieces of evidence are factually correct, leaving it as a future work. Our work’s *factuality* criteria exactly covers this part. Sadallah et al. (2025) proposes four criteria, *actionability*, *grounding & specificity*, *verifiability*, and *helpfulness*. However, they only focus on evaluating weaknesses, excluding questions, and as in Guo et al. (2023), they only evaluate whether weaknesses are grounded and supported, not whether those are factually correct. Purkayastha et al. (2025) defines *lazy thinking* in the context of NLP paper reviews and introduces heuristic-based rubrics. However, their criteria is only applicable to NLP domain which is hard to be generalized into other domains. [Further comparisons with previous works in terms of dataset statistics and review evaluation methods are described in Appendix I.](#)

**Argument evaluation.** In logic and critical thinking, an argument is a list of statements, one of which is the conclusion and the others are the premises (Dutilh Novaes, 2022; Lau & Chan, 2025). To evaluate an argument, we need to follow two steps. First, we have to identify and reconstruct the argument into a set of premises and conclusion, which is called an argument reconstruction (Brun & Betz, 2016; Groarke, 2024; Dowden, 2024). Then, we evaluate whether each premise is factually correct. An argument reconstruction should both be *valid*, which means premises deductively imply a conclusion, and *faithful*, which means premises and a conclusion accurately and completely represents an original argument (Brun & Betz, 2016; Betz & Richardson, 2021). Previously, Betz & Richardson (2021) trains a T5 model for argument reconstruction. However, the training datasets are either synthetic or polished and the reconstruction do not require any additional context information. In contrast, our work targets peer reviews, which include real-world unpolished arguments, and the reconstruction requires an entire paper to fully understand the context of arguments.

## 3 REVIEWSCORE

We newly define REVIEWSCORE that measures how *misinformed* a review point is (Section 3.1). To evaluate REVIEWSCORE, we also introduce an engine that automatically extracts every *explicit* and *implicit premises* from a weakness (Section 3.2). Lastly, we construct a human expert-annotated dataset that evaluates LLM’s ability to evaluate REVIEWSCORE (Section 3.3).

### 3.1 DEFINITION

Our goal of defining REVIEWSCORE is to detect *misinformed* review points. Following this goal and the review quality criteria discussion in Section 1, we first define BASE REVIEWSCORE as *factuality of weaknesses* and *unanswerability of questions*. The following definition formally describes it.

**Definition 3** (BASE REVIEWSCORE). Let  $x$  be a review point (either a weakness or a question) about a submitted paper  $S$ . Define

$$\text{Factuality}_S : \mathcal{W} \rightarrow \{1, 2, 3, 4, 5\}, \quad \text{Unanswerability}_S : \mathcal{Q} \rightarrow \{1, 2, 3, 4, 5\},$$

where  $\mathcal{W}$  and  $\mathcal{Q}$  are, respectively, the sets of weaknesses and questions appearing in a review of  $S^4$ . We considered score 1–2 as *Misinformed* and score 3–5 as *Not misinformed* for binary classification

<sup>4</sup>Detailed rubric is described in Appendix H.

162 setup. The BASE REVIEWSCORE of  $x$  is:

$$163 \text{ReviewScore}_{\text{base}}(x) = \begin{cases} \text{WScore}(x) = \text{Factuality}_S(x), & \text{if } x \in \mathcal{P} \text{ (weakness),} \\ \text{QScore}(x) = \text{Unanswerability}_S(x), & \text{if } x \in \mathcal{Q} \text{ (question).} \end{cases}$$

167 However, during the group discussion in Section 1, we discovered that human annotators’ evaluations  
168 on factuality diverge if a weakness contains both factual and nonfactual premises or a non-  
169 factual premise is implicitly presumed. It happens since the human annotators implicitly weigh the  
170 importance of underlying premises of a weakness and then decide the final factuality score.

171 To resolve this issue, we further define ADVANCED REVIEWSCORE. We categorize weaknesses  
172 into arguments and claims based on whether there are supporting reasons or not. We keep the defini-  
173 tion of WScore to evaluate claims, but further develops a finer-grained score to evaluate arguments.  
174 Following the literature of critical thinking (Brun & Betz, 2016; Groarke, 2024; Dowden, 2024),  
175 we reconstruct an argument into a premise-conclusion structure and then define ADVANCED RE-  
176 VIEWSCORE for arguments as (an aggregation of) *factuality of premises*. The following definition  
177 formally describes it.

178 **Definition 4** (ADVANCED REVIEWSCORE). Let  $x$  be a review point about a submitted paper  $S$ . Let  
179  $\mathcal{C}$ ,  $\mathcal{A}$ , and  $\mathcal{Q}$  denote, respectively, the sets of claims, arguments, and questions in a review of  $S$ . For  
180  $x \in \mathcal{A}$ , let its (*explicit* and *implicit*) premises be  $\mathcal{P}(x) = \{p_1, \dots, p_k\}$  with  $\{p_1, \dots, p_k\} \vdash C$  for  
181 the conclusion  $C$  of  $x$ . Let  $\mathcal{K}$  be the set of knowledge bases available for factuality judgments (i.e.,  
182  $S$ , annotator knowledge, referred papers), and let

$$183 \text{Factuality} : \mathcal{U} \times \mathcal{K} \rightarrow \{1, 2, 3, 4, 5\}, \quad \mathcal{U} := \mathcal{C} \cup \bigcup_{x \in \mathcal{A}} \mathcal{P}(x),$$

185 be a 5-point scoring function for claims and premises given a knowledge base. Define a selector  
186  $\text{KB} : \mathcal{U} \rightarrow \mathcal{K}$  that chooses the knowledge base used for each item (for claims  $x \in \mathcal{C}$ ,  $\text{KB}(x) = S$ ;  
187 for premises  $p_i$ ,  $\text{KB}(p_i) = \text{KB}_i \in \mathcal{K}$ ). For a given  $\text{KB}(x) \in \mathcal{K}$ , we notate the factuality function  
188 as  $\text{Factuality}_{\text{KB}(x)}(\cdot)$ .  $\text{Agg}$  is an operator that aggregates a list of scores to a single score. The  
189 ADVANCED REVIEWSCORE of  $x$  is:

$$190 \text{ReviewScore}_{\text{adv}}(x) = \begin{cases} \text{ClaimScore}(x) = \text{Factuality}_S(x), & \text{if } x \in \mathcal{C} \text{ (claim),} \\ \text{ArgScore}(x) = \text{Agg}(\text{Factuality}_{\text{KB}_i}(p_i)_{p_i \in \mathcal{P}(x)}), & \text{if } x \in \mathcal{A} \text{ (argument),} \\ \text{QScore}(x) = \text{Unanswerability}_S(x), & \text{if } x \in \mathcal{Q} \text{ (question).} \end{cases}$$

195 **Aggregation methods.** To aggregate premise factuality scores as a single ArgScore, we introduce  
196 two aggregation methods: logical conjunction and weighted average. Following the literature of  
197 logic, an argument is *true* if and only if all premises are *true*. We define a premise is *true* if and  
198 only if it has factuality score 3–5, and otherwise, it is *false*. We dubbed this aggregation as logical  
199 conjunction, which follows the binary classification setup. However, if an annotator mistakenly  
200 evaluates one of premises, then the error propagates to the entire argument. To alleviate this issue,  
201 we also aggregate by a weighted average. Since it is difficult to measure the importance of premises,  
202 we instead weighted scores by untrivialness of premises (with a scale of 0–2). This is intended to  
203 simply filter out trivially true premises by measuring their importance as 0.<sup>5</sup>

### 205 3.2 AUTOMATIC ARGUMENT RECONSTRUCTION

206  
207 To evaluate ArgScore, we have to extract (*explicit* and *implicit*) premises  $\mathcal{P}(x)$  from an argument  
208  $x \in \mathcal{A}$ . Since human experts require significant amount of time and costs to do this, we automate  
209 it using LLMs. First, we check if a model could directly reconstruct an argument by giving detailed  
210 instructions. To preserve the context of an argument, we also give the model a submitted paper  $S$ .  
211 However, it mostly fails to generate *valid* and *faithful* reconstructions<sup>6</sup>. To resolve this issue, we add  
212 two feedback loops to ensure *validity* and *faithfulness* of an argument reconstruction (Figure 2a).

213 <sup>5</sup>Since we reconstruct every argument as valid (i.e., a set of premises deductively implies a conclusion),  
214 there are often conditional premises that make an argument valid but are trivially true. Detailed rubric of  
215 untrivialness is described in Appendix H.

<sup>6</sup>Detailed numerical results and qualitative analysis are reported in Appendix C.

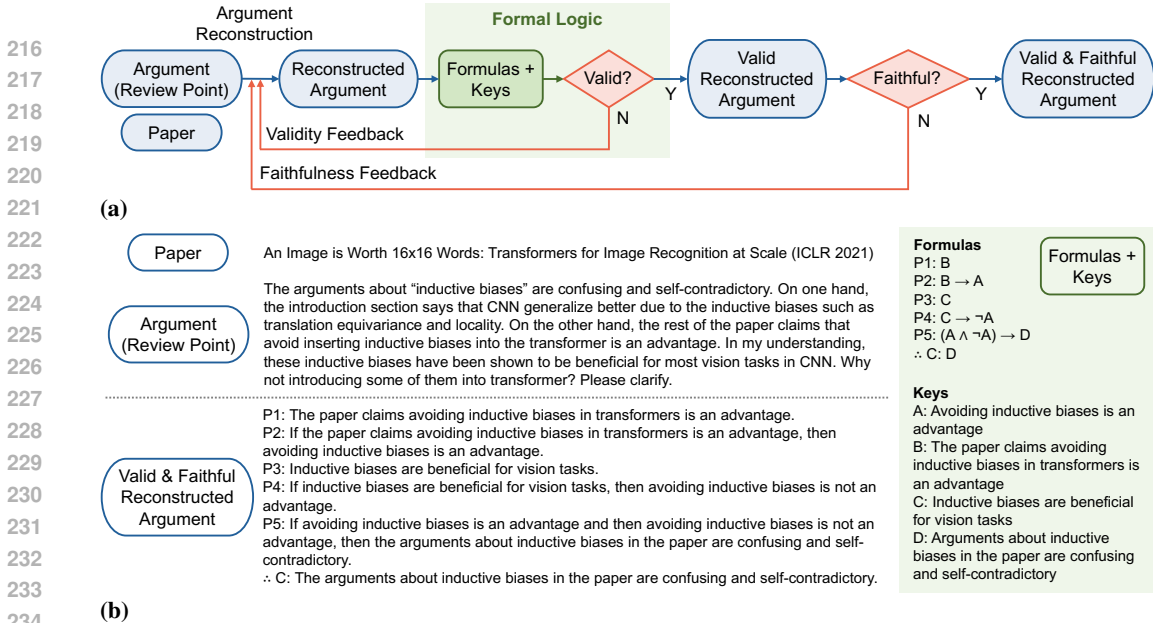


Figure 2: (a) Overview of an automatic argument reconstruction. Given an argumentative review point with a paper, a model first generates a reconstructed argument (i.e., a set of premises and conclusion). To check its *validity*, a model translates a NL reconstructed argument into FOL formulas, and then a SAT solver judges if it is *valid*. To check its *faithfulness*, a model translate FOL formulas back into the NL domain, and a model judges if the reconstruction is *faithful*. If one of two criteria does not met, then corresponding NL feedback is given to the generator model. (b) A representative example. We sample a review point of Dosovitskiy et al. (2020) and its reconstruction along with corresponding formulas and keys.

An LLM alone often falls short in ensuring *validity*, so we include a SAT solver which could automatically judges the *validity* of a set of premises and a conclusion without any logical errors. To do that, an LLM translates a set of natural language (NL) premises and conclusion into first-order logic (FOL) formulas. Then, a SAT solver decides whether the premises deductively implies the conclusion. If the reconstruction is *valid*, then an LLM translates formalized premises and conclusion back into NL domain. This process is called *logical streamlining*, which means to rephrase NL premises or conclusion in order to make their logico-semantic structure more transparent (Bowell & Kemp, 2014; Brun & Betz, 2016; Betz & Richardson, 2021). We then pass these *streamlined* NL premises and conclusion to the subsequent faithfulness feedback loop. However, if the reconstruction is *invalid*, then we feed a rule-based reward signal to the argument reconstructor to regenerate the reconstruction. There are two types of reward signals, one is a naive signal that says the formalized premises do not imply the conclusion, and the other one tells that the proof is circular.

Although the reconstruction is now *valid*, it should also *faithfully* represents the original argument. To ensure that, we prompt an LLM to judge whether the reconstruction is *faithful* or not and justify its decision. If the reconstruction is *faithful*, then we stop iterating the loop. If the reconstruction is *unfaithful*, then we feed LLM judge’s justification to the argument reconstructor to regenerate the reconstruction. To minimize model calls in practice, we connect these two loops in series so that only *valid* reconstructions are judged for their *faithfulness*. Also, we limit the total number of loop iteration to 10 and return the last reconstructed argument if the loop fails to generate *valid* and *faithful* reconstruction. We provide an example reconstruction in Figure 2b, and details on feedback loops and model prompts are described in Appendix C and H, respectively.

**Quality of argument reconstruction.** We measure *validity* and *faithfulness* of reconstructed arguments using a SAT solver and human annotators, respectively. With `claude-sonnet-3.7` as a base LLM, every reconstruction is *valid*, and average *faithfulness* score is 4.5 / 5. Detailed analysis including comparison with direct reconstruction is described in Appendix C.

### 3.3 DATASET CONSTRUCTION

Our final destination of proposing REVIEWSCORE is to *automatically* filter out *misinformed* review points using LLMs. To verify LLM’s ability to do that, we build a human expert-annotated dataset to measure an agreement between humans and LLMs on REVIEWSCORE evaluation. Our dataset contains total 657 annotated review points, consisting of 143 questions, 92 claims, and 422 arguments. Specifically, 1,748 premises of the arguments are manually annotated<sup>7</sup>.

**Human-review matching.** We recruit 15 graduate students studying AI as human annotators, and they annotate total 40 papers submitted to ICLR 2021–2023<sup>8</sup>. Specifically, we first make five groups by their research interests. For each group, three human annotators discuss which papers to annotate and select eight papers that are relevant with all three. Then, each human annotates selected eight papers’ review points which are preprocessed from OpenReview.

**Data curation process.** We collaboratively use an LLM and humans for REVIEWSCORE data curation, where an LLM preprocesses reviews and then humans annotate those. Given a review, an LLM extracts independent review points. For each review point, an LLM automatically annotates the type (i.e., claim, argument, or question) and human verifies it. If the review point is a question, then human scores if the question is (un)answerable by the paper in a 5-point scale and justifies it if needed. If the review point is a claim, then human scores if the claim is true in a 5-point scale and justifies it if needed. If the review point is an argument, then human scores the argument’s factuality same as in claims (i.e., BASE REVIEWSCORE). To annotate ADVANCED REVIEWSCORE for arguments, we first run automatic argument reconstruction engine (Section 3.2) to extract underlying premises of the argument. After that, human scores if the reconstruction is *faithful* in a 5-point scale. If the *faithfulness* score is less than 4 (i.e., faithful, but one or two minor changes recommended), then they skip the subsequent annotations. Otherwise, they judge the factuality of premises. For each premise, they first select a knowledge base (i.e., submitted paper, annotator knowledge, or referred papers), score if the premise is true based on the knowledge base in a 5-point scale, and score if the premise is (un)trivial in a 3-point scale. They justify any of three decisions if needed.

**Trustworthiness of human annotation.** Three humans independently annotate reviews (i.e., we take a median value as a final human label), and they are allowed to use any related materials or tools (e.g., discussion between authors and reviewers in OpenReview, web search, etc.). We also provide annotators detailed guidelines and an hour-long online orientation session. Furthermore, we highly encourage them to communicate with us through a group chat so that we could give them instant responses to their questions and share with all, which builds a global consensus among human annotators. Thanks to these efforts, despite the difficulty of the work, we obtain median 0.301 and highest 0.428 inter-annotator agreement in Krippendorff’s Alpha (Hughes, 2021) across different annotator groups. Detailed results are described in Appendix E.

## 4 RELIABILITY OF AUTOMATIC EVALUATION OF REVIEWSCORE

To evaluate LLM’s ability to evaluate REVIEWSCORE, we describe an experimental setup (Section 4.1) and show human-model agreement results (Section 4.2). We also compare the effectiveness of BASE and ADVANCED REVIEWSCORE (Section 4.3). We further analyze human-model disagreements and the effect of providing authors response to models (Section 4.4).

### 4.1 SETUP

Given score rubrics in a 5-point scale, an LLM evaluates REVIEWSCORE according to a knowledge base it selects. We only provide a submitted paper to a model since we assume that the model has a pretrained human-level internal knowledge and has a general understanding of referred papers. Detailed model prompts are described in Appendix H.

<sup>7</sup>Details of the dataset construction and statistics are described in Appendix D.

<sup>8</sup>Since ICLR 2024–2025 submission drafts are not opened to public, we exclude these years.

Table 1: Human-model agreement on REVIEWSCORE evaluation.

Model	ClaimScore		ArgScore		QScore		REVIEWSCORE	
	F1	Kappa	F1	Kappa	F1	Kappa	F1	Kappa
<i>Proprietary models</i>								
claude-sonnet-3.7	<b>0.160</b>	<b>0.137</b>	<b>0.458</b>	<b>0.380</b>	<b>0.542</b>	<b>0.387</b>	<b>0.448</b>	0.355
claude-sonnet-4	0.125	0.130	0.329	0.266	0.507	0.357	0.367	0.333
gpt-4o	0.000	0.119	0.329	0.239	0.538	0.340	0.405	<b>0.362</b>
gpt-5	0.125	-0.004	0.426	0.334	<b>0.543</b>	0.300	0.441	0.310
gemini-2.5-flash	0.091	0.106	0.447	0.366	0.512	0.225	0.439	0.318
<i>Open-sourced models</i>								
qwen3-235b-a22b	<b>0.231</b>	0.094	<b>0.403</b>	0.149	0.531	0.234	<b>0.436</b>	0.243
llama-3.3	0.133	0.153	0.250	0.107	0.551	<b>0.284</b>	0.409	<b>0.355</b>
deepseek-v3	0.000	<b>0.165</b>	0.276	<b>0.196</b>	<b>0.597</b>	0.228	0.382	0.310

**Language models.** To measure LLM’s ability to automatically evaluate REVIEWSCORE, we perform experiments on eight current state-of-the-art LLMs that achieve significantly high alignments with humans. We include five proprietary models, claude-sonnet-3.7 (Anthropic, 2025a), claude-sonnet-4 (Anthropic, 2025b), gpt-4o (OpenAI, 2024), gpt-5 (OpenAI, 2025), and gemini-2.5-flash (Gemini Team, Google, 2025), and three open-sourced models, qwen3-235b-a22b (Yang et al., 2025), deepseek-v3 (DeepSeek-AI et al., 2024), and llama-3.3 (AI@Meta, 2024)<sup>9</sup>. To get consistent and reliable scores from LLM judges, we set a low temperature (i.e., 0) and select the highest probability response from the model (Liang et al., 2022; Liu et al., 2023; Gu et al., 2024). We exclude reasoning models as LLM judges since REVIEWSCORE mostly depends on grounding and evidence, not longer or smarter chains of thought.

**Evaluation metrics.** We use different sets of metrics for two types of problem formulations, binary classification and 5-point scale scoring. For the binary classification, since majority of human-annotated labels are *Not misinformed*, we mainly use F1 score which is robust to class imbalance. For the 5-point scale scoring, since majority of human-annotated scores are 4 and 5, we mainly use Quadratic Weighted Kappa (Warrens, 2012), a variant of Cohen’s Kappa (Cohen, 1960), that is robust to the skewed data distribution. Hereinafter, we call this metric Kappa for simplicity. To provide more comprehensive results, we additionally use Precision and Recall for the binary classification and Pearson rank correlation and Gwet’s AC2 (Gwet, 2001) for the 5-point scale scoring.

## 4.2 MAIN RESULTS

We empirically validate the alignment of human-annotated and model-estimated REVIEWSCORE using different models and evaluation metrics in Table 1. Most models show 0.4–0.5 F1 score and 0.3–0.4 Kappa score, which indicates moderate agreement between humans and models on REVIEWSCORE evaluation. However, there are differences in human-model agreements for three types of review points. Regardless of the models, questions show the highest agreement, arguments follow subsequently, and claims show the lowest agreement. Specifically, for claims, some models show zero F1 score or negative Kappa score. We analyze human-model disagreements thoroughly in Section 4.4 and conclude that since claims lack supporting evidence and are often value-laden, models often misinterpret the intended meaning of the claims or judge differently than humans.

For proprietary models, claude-sonnet-3.7 shows the highest human-model agreements across different types of review points. One thing to note is claude-sonnet-4 consistently shows lower agreement than its preceding model. We find out that it’s because claude-sonnet-4 sometimes overthink and more frequently predicts the minority human-annotated scores than claude-sonnet-3.7 does. In contrast, comparing F1 scores of gpt-4o and gpt-5, the advanced model shows higher agreement than the preceding model.

For open-sourced models, qwen3-235b-a22b shows best human-model agreements in terms of F1 score. Open-sourced models show lower human-model agreements than proprietary models, but the gap is not significant. Open-sourced models sometimes show better agreements than proprietary models, for instance, llama-3.3 outperforms claude-sonnet-4 on REVIEWSCORE evalua-

<sup>9</sup>Details of model specifications are described in Appendix G

Table 2: Comparison of human-model agreement of BASE vs. ADVANCED REVIEWSCORE.

Model	Metric	Base	Advanced w/o Agg	Advanced
claude-sonnet-3.7	F1	0.185	0.262	<b>0.458</b>
	Kappa	0.153	0.280	<b>0.380</b>
gpt-5	F1	0.243	0.316	<b>0.426</b>
	Kappa	0.223	0.315	<b>0.334</b>
gemini-2.5-flash	F1	0.167	0.349	<b>0.447</b>
	Kappa	0.138	0.330	<b>0.366</b>
deepseek-v3	F1	0.203	0.163	<b>0.276</b>
	Kappa	0.120	<b>0.202</b>	0.196

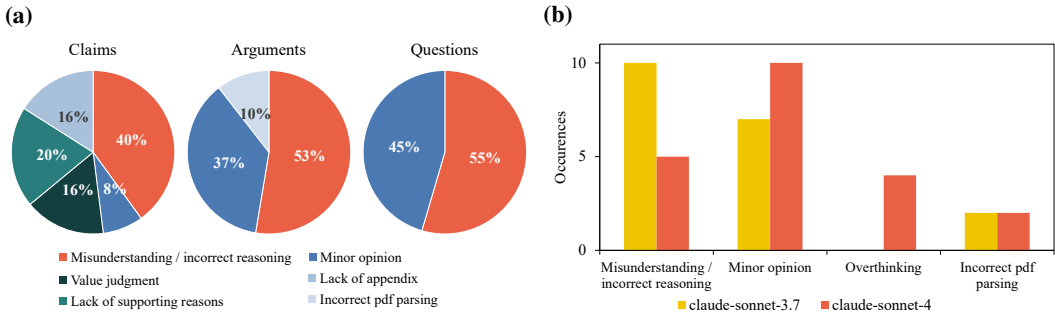


Figure 3: Human-model disagreement analysis. (a) Types of human-model disagreements. (b) Comparison of disagreement types of claude-sonnet-3.7 and claude-sonnet-4.

tion. However, open-sourced models consistently show lower agreements on evaluating ArgScore. More results and qualitative analysis of model evaluation are described in Appendix F.

### 4.3 BASE VS. ADVANCED REVIEWSCORE

We compare human-model agreements of BASE and ADVANCED REVIEWSCORE using different models and evaluation metrics. Since the only difference is in defining scores for arguments, we compare REVIEWSCORE for arguments in Table 2. To verify the effectiveness of aggregation methods, we additionally include ADVANCED REVIEWSCORE for arguments without aggregation (i.e., factuality of premises). Regardless of models, ADVANCED REVIEWSCORE clearly shows higher agreements than BASE REVIEWSCORE. For claude-sonnet-3.7, ADVANCED REVIEWSCORE performs 2.48x higher F1 score and 2.48x higher Kappa score than the BASE REVIEWSCORE. The aggregation methods also contribute to higher agreements. For claude-sonnet-3.7, ADVANCED REVIEWSCORE performs 1.75x higher F1 score and 1.36x higher Kappa score than the one without aggregation.

### 4.4 ANALYSIS

**REVIEWSCORE Distribution.** We present model-estimated REVIEWSCORE distribution and compare with human-annotated ones. Specifically, following the binary classification setup, we report the proportion of *misinformed* review points in Table 3. There is a clear trend that models more frequently decide review points as *Misinformed* than humans. Models decide 31.2% of reviews points as *Misinformed* in average, where humans decide only 17.8% as *Misinformed*. The discrepancy is the largest in claims where models decide 22.8% as *Misinformed* and humans decide only 5.9% as *Misinformed*. To understand these results more deeply, we continue our analysis on the following human-model disagreements.

Table 3: Proportion of *misinformed* review points estimated by models and humans.

	Model	Human
Claims	22.8	5.9
Arguments	30.1	17.7
Questions	40.0	26.4
Review points	31.2	17.8

**Human-model disagreements.** We analyze types of human-model disagreements in Figure 3a. Across all review points, there are two common types of disagreements, which are models misunderstand or incorrectly reason and models predict scores that are minority of human-annotated

Table 4: Effect of providing Authors Response (AR) to a model for REVIEWSCORE evaluation.

	F1		Kappa	
	w/o AR	w/ AR	w/o AR	w/ AR
WScore	0.211	<b>0.286</b>	0.178	<b>0.212</b>
ArgScore w/o Agg	0.230	<b>0.283</b>	0.271	<b>0.312</b>
ArgScore	0.329	<b>0.447</b>	0.266	<b>0.340</b>
QScore	0.507	<b>0.513</b>	<b>0.357</b>	0.324
ClaimScore	0.125	<b>0.133</b>	0.130	<b>0.225</b>
REVIEWSCORE	0.367	<b>0.449</b>	0.333	<b>0.377</b>

scores. These two types comprise nearly all disagreements on arguments and questions, and about half of disagreements on claims. For claims, we observe that 36% of disagreements are caused by either claims are value-laden or lack supporting reason. These make the factuality judgment of claims subjective, leading to a low human-model agreement. Practical limitations such as incorrect pdf parsing and not providing appendix to models cause nontrivial portion of disagreements. We also analyze the reason why `claude-sonnet-3.7` shows higher human-model agreements than `claude-sonnet-4` in Figure 3b. The advanced model shows smaller number of misunderstanding or incorrect reasoning, however, it shows higher number of overthinking<sup>10</sup> and minority human-annotated score predictions, leading to a lower human-model agreement.

**Helpfulness of authors response.** We study whether providing authors responses of reviews to a model benefits the automatic REVIEWSCORE evaluation using `claude-sonnet-4` in Table 4. Providing authors response clearly shows higher human-model agreement than not providing it. Specifically, the agreement of REVIEWSCORE increases by 22.3% in F1 score and 13.2% in Kappa score by incorporating authors response. Note that `claude-sonnet-4` with authors response achieves the highest agreement compared to all models without authors response. Furthermore, the agreement of WScore and ArgScore without aggregation (i.e., premise factuality) also increases significantly by incorporating authors response. These observations show that authors response largely benefits detecting *misinformed* reviews by providing additional cues for model judgments.

## 5 LIMITATIONS

We acknowledge three types of limitations of our work. First, there are technical and practical limitations in the automated REVIEWSCORE evaluation. Throughout Section 4, human-model agreement seems low even with authors response. However, compared to Du et al. (2024) which measures human-model agreement on *Deficient* review detection and reports the best F1 score as 0.220, our work reports the best F1 score as 0.449, indicating more than 2× agreement. To save API calling costs, when the paper is given to LLMs, we only provide text and tables of the main paper, where any figure or appendix is excluded. We also observe nontrivial amount of incorrect pdf parsing. Second, there are technical limitations in the automatic argument reconstruction. The reconstruction output is not always perfect since it depends on the capability of the base model. However, we observe a considerable output quality improvement when we upgrade the base model to `claude-sonnet-4`, indicating that the reconstruction engine would perform better as the model improves in general. Lastly, there are practical limitations in the dataset construction. Since human annotators are graduate students with varying skillfulness and the annotation requires significant amount of cognitive load, there are unavoidable noise in human annotation. Furthermore, a manual selection of papers by human annotators might introduce any kind of unintended biases. However, to collect the most reliable annotation under our limited budget, we inevitably choose this method.

## 6 CONCLUSION

We introduce REVIEWSCORE, a new evaluation of peer review quality, focusing on detecting *misinformed* review points. To automate REVIEWSCORE evaluation using LLMs and validate it, we also

<sup>10</sup>We judge a model overthinks if its justification of a score is significantly longer than another model. We also manually verify if a model overanalyzes a review point.

486 construct a trustworthy human-annotated dataset. The results show a moderate human-model agree-  
487 ment, and further comprehensive disagreement analysis supports the potential of a fully automated  
488 REVIEWSCORE evaluation with current state-of-the-art LLMs. We hope our work could contribute  
489 to mitigate the issue of low-quality reviews in AI conferences.  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539

## REFERENCES

- 540  
541  
542 AI@Meta. The llama 3 herd of models. 07 2024. URL <https://arxiv.org/abs/2407.21783>. Family report for Llama 3; used to cite Llama 3.3 70B Instruct.
- 543  
544 Anthropic. Claude 3.7 sonnet system card. Technical report, Anthropic PBC, 02 2025a.  
545 URL <https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf>. System card / technical report.
- 546  
547 Anthropic. System card: Claude opus 4 & claude sonnet 4. Technical report,  
548 Anthropic PBC, 05 2025b. URL <https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf>. System card / technical report.
- 549  
550  
551  
552 Simran Arora, Avaniika Narayan, Mayee F Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami,  
553 and Christopher Re. Ask me anything: A simple strategy for prompting language models. In  
554 *The Eleventh International Conference on Learning Representations, 2023*. URL <https://openreview.net/forum?id=bhUPJnS2g0X>.
- 555  
556 Jc Beall, Greg Restall, and Gil Sagi. Logical Consequence. In Edward N. Zalta and Uri Nodelman  
557 (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University,  
558 Summer 2024 edition, 2024.
- 559  
560 Gregor Betz and Kyle Richardson. Deepa2: A modular framework for deep argument analysis with  
561 pretrained neural text2text language models. *arXiv preprint arXiv:2110.01509*, 2021.
- 562  
563 Vadim Borisov, Kathrin Sessler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language  
564 models are realistic tabular data generators. In *The Eleventh International Conference on Learning Representations, 2023*. URL <https://openreview.net/forum?id=cEYgmQNOeI>.
- 565  
566  
567 Tracy Bowell and Gary Kemp. *Critical Thinking: A Concise Guide*. Routledge, 2014.
- 568  
569 Jordan Boyd-Graber, Naoaki Okazaki, and Anna Rogers. Acl’23 peer review policies. ACL 2023  
570 Blog, February 2023. URL <https://2023.aclweb.org/blog/review-acl23/>. Program Chairs blog post.
- 571  
572 Georg Brun and Gregor Betz. Analysing practical argumentation. In *The argumentative turn in policy analysis: Reasoning about uncertainty*, pp. 39–77. Springer, 2016.
- 573  
574 Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.
- 575  
576 Corinna Cortes and Neil D. Lawrence. Inconsistency in conference peer review: Revisiting the 2014  
577 neurips experiment. *arXiv preprint arXiv:2109.09774*, 2021. doi: 10.48550/arXiv.2109.09774.  
578 URL <https://arxiv.org/abs/2109.09774>.
- 579  
580 Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-  
581 based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations, 2023*. URL <https://openreview.net/forum?id=3lge0p5o-M->.
- 582  
583  
584 Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large  
585 language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations, 2023*. URL <https://openreview.net/forum?id=3Pf3Wg6o-A4>.
- 586  
587  
588 Leonardo De Moura and Nikolaj Bjørner. Z3: An efficient smt solver. In *International conference on Tools and Algorithms for the Construction and Analysis of Systems*, pp. 337–340. Springer, 2008.
- 589  
590  
591  
592 DeepSeek-AI, Aixin Liu, Bei Feng, et al. Deepseek-v3 technical report. 12 2024. URL <https://arxiv.org/abs/2412.19437>. Technical report; applicable to the V3-0324 checkpoint lineage.
- 593

- 594 Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas  
595 Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An  
596 image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint*  
597 *arXiv:2010.11929*, 2020.
- 598 Bradley H Dowden. Logical reasoning (dowden), 2024. URL [https://human.  
599 libretexts.org/Bookshelves/Philosophy/Logic\\_and\\_Reasoning/  
600 Logical\\_Reasoning\\_\(Dowden\)](https://human.libretexts.org/Bookshelves/Philosophy/Logic_and_Reasoning/Logical_Reasoning_(Dowden)).  
601
- 602 Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng  
603 Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, et al. Llms assist nlp researchers:  
604 Critique paper (meta-) reviewing. *arXiv preprint arXiv:2406.16253*, 2024.
- 605 Catarina Dutilh Novaes. Argument and Argumentation. In Edward N. Zalta and Uri Nodelman  
606 (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University,  
607 Fall 2022 edition, 2022.  
608
- 609 Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting  
610 for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*,  
611 2023. URL <https://openreview.net/forum?id=yflicZHC-19>.
- 612 Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and  
613 Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using  
614 textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023.  
615 URL <https://openreview.net/forum?id=NAQvF08TcyG>.  
616
- 617 Gemini Team, Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodal-  
618 ity, long context, and next generation agentic capabilities. Technical report, Google DeepMind,  
619 06 2025. URL [https://storage.googleapis.com/deepmind-media/gemini/  
620 gemini\\_v2\\_5\\_report.pdf](https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf). Technical report; covers Gemini 2.5 Pro and 2.5 Flash.
- 621 Alexander Goldberg, Ivan Stelmakh, Kyunghyun Cho, Alice Oh, Alekh Agarwal, Danielle Bel-  
622 grave, and Nihar B Shah. Peer reviews of peer reviews: A randomized controlled trial and other  
623 experiments. *PLoS one*, 20(4):e0320444, 2025.
- 624 Leo Groarke. Informal Logic. In Edward N. Zalta and Uri Nodelman (eds.), *The Stanford En-  
625 cyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2024 edition,  
626 2024.  
627
- 628 Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan  
629 Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. A survey on llm-as-a-judge.  
630 *arXiv preprint arXiv:2411.15594*, 2024. URL <https://arxiv.org/abs/2411.15594>.  
631 Recommends temperature=0 to reduce randomness and stabilize judge outputs.
- 632 Yanzhu Guo, Guokan Shang, Virgile Rennard, Michalis Vazirgiannis, and Chloé Clavel. Automatic  
633 analysis of substantiation in scientific peer reviews. *arXiv preprint arXiv:2311.11967*, 2023.  
634
- 635 Kilem Gwet. Handbook of inter-rater reliability. *Gaithersburg, MD: STATAXIS Publishing Com-  
636 pany*, pp. 223–246, 2001.
- 637 Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or.  
638 Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Con-  
639 ference on Learning Representations*, 2023. URL [https://openreview.net/forum?  
640 id=\\_CDixzkzeyb](https://openreview.net/forum?id=_CDixzkzeyb).
- 641 Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. Language model  
642 compression with weighted low-rank factorization. In *International Conference on Learning  
643 Representations*, 2022. URL <https://openreview.net/forum?id=uPv9Y3gmAI5>.  
644
- 645 Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,  
646 and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Con-  
647 ference on Learning Representations*, 2022. URL [https://openreview.net/forum?  
id=nZeVKeeFYf9](https://openreview.net/forum?id=nZeVKeeFYf9).

- 648 Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei  
649 Han. Large language models can self-improve, 2023. URL [https://openreview.net/  
650 forum?id=NiEtU7blzN](https://openreview.net/forum?id=NiEtU7blzN).
- 651 John Hughes. krippendorffsalpha: An r package for measuring agreement using krippendorff’s  
652 alpha coefficient. *arXiv preprint arXiv:2103.12170*, 2021.
- 653  
654 Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun KIM, Stan-  
655 ley Jungkyu Choi, and Minjoon Seo. Towards continual knowledge learning of language  
656 models. In *International Conference on Learning Representations*, 2022. URL [https://openreview.net/  
657 forum?id=vfsRB5MIo9](https://openreview.net/forum?id=vfsRB5MIo9).
- 658 Haozhe Ji, Pei Ke, Zhipeng Hu, Rongsheng Zhang, and Minlie Huang. Tailoring language genera-  
659 tion models under total variation distance. In *The Eleventh International Conference on Learning  
660 Representations*, 2023. URL <https://openreview.net/forum?id=VELL0P1Wfc>.
- 661  
662 Jacalyn Kelly, Tara Sadeghieh, and Khosrow Adeli. Peer review in scientific publications: benefits,  
663 critiques, & a survival guide. *Ejifcc*, 25(3):227, 2014.
- 664 Jaeho Kim, Yunseok Lee, and Seulki Lee. Position: The ai conference peer review crisis demands  
665 author feedback and reviewer rewards. *arXiv preprint arXiv:2505.04966*, 2025.
- 666  
667 Yoonjeon Kim, Hyunsu Kim, Junho Kim, Yunjey Choi, and Eunho Yang. Learning input-agnostic  
668 manipulation directions in styleGAN with text guidance. In *The Eleventh International Confer-  
669 ence on Learning Representations*, 2023. URL [https://openreview.net/forum?id=  
670 47B\\_ctC4pJ](https://openreview.net/forum?id=47B_ctC4pJ).
- 671 Anna Kukleva, Moritz Böhle, Bernt Schiele, Hilde Kuehne, and Christian Rupprecht. Temperature  
672 schedules for self-supervised contrastive methods on long-tail data. In *The Eleventh International  
673 Conference on Learning Representations*, 2023. URL [https://openreview.net/forum/  
674 id=ejHUr4nfHhD](https://openreview.net/forum?id=ejHUr4nfHhD).
- 675 Joe Lau and Jonathan Chan. What is an argument? Critical Thinking Web, 2025. URL [https://  
676 philosophy.hku.hk/think/arg/arg.php](https://philosophy.hku.hk/think/arg/arg.php). Module A01: Argument analysis.
- 677  
678 Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be  
679 strong differentially private learners. In *International Conference on Learning Representations*,  
680 2022. URL <https://openreview.net/forum?id=bVuP3ltATMz>.
- 681  
682 Kevin J Liang, Weituo Hao, Dinghan Shen, Yufan Zhou, Weizhu Chen, Changyou Chen, and  
683 Lawrence Carin. Mix{kd}: Towards efficient distillation of large-scale language models. In  
684 *International Conference on Learning Representations*, 2021. URL [https://openreview.  
net/forum?id=UFGEelJkLu5](https://openreview.net/forum?id=UFGEelJkLu5).
- 685  
686 Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian  
687 Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language  
688 models. *arXiv preprint arXiv:2211.09110*, 2022. URL [https://arxiv.org/abs/2211.  
09110](https://arxiv.org/abs/2211.09110). Standardizes eval settings; many HELM runs use temperature=0 for reproducibility.
- 689  
690 Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow  
691 matching for generative modeling. In *The Eleventh International Conference on Learning Repre-  
692 sentations*, 2023. URL <https://openreview.net/forum?id=PqvMRDCJT9t>.
- 693  
694 Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval:  
695 Nlg evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Confer-  
696 ence on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2511–2522. Asso-  
697 ciation for Computational Linguistics, 2023. URL [https://aclanthology.org/2023.  
emnlp-main.153/](https://aclanthology.org/2023.emnlp-main.153/). Uses low temperature / deterministic decoding for stable evaluator out-  
698 puts.
- 699  
700 Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter  
701 Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured  
mathematical reasoning. In *The Eleventh International Conference on Learning Representations*,  
2023. URL <https://openreview.net/forum?id=DHyHRBwJUTN>.

- 702 Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon.  
703 SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. URL [https://openreview.net/forum?id=aBsCjcPu\\_tE](https://openreview.net/forum?id=aBsCjcPu_tE).
- 706 Jaehyun Nam, Jihoon Tack, Kyungmin Lee, Hankook Lee, and Jinwoo Shin. STUNT: Few-shot  
707 tabular learning with self-generated tasks from unlabeled tables. In *The Eleventh International  
708 Conference on Learning Representations*, 2023. URL [https://openreview.net/forum?id=\\_xlsjehDvly](https://openreview.net/forum?id=_xlsjehDvly).
- 711 NeurIPS 2025 Program Chairs. Neurips 2025 reviewer guidelines. NeurIPS 2025 website, 2025.  
712 URL <https://neurips.cc/Conferences/2025/ReviewerGuidelines>. Guide-  
713 lines page.
- 714 Ansong Ni, Jeevana Priya Inala, Chenglong Wang, Alex Polozov, Christopher Meek, Dragomir  
715 Radev, and Jianfeng Gao. Learning math reasoning from self-sampled correct and partially-  
716 correct solutions. In *The Eleventh International Conference on Learning Representations*, 2023.  
717 URL <https://openreview.net/forum?id=4D4TSJE6-K>.
- 718 Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth  
719 64 words: Long-term forecasting with transformers. In *The Eleventh International Confer-  
720 ence on Learning Representations*, 2023. URL <https://openreview.net/forum?id=Jbdc0vTOcol>.
- 723 OpenAI. Gpt-4o system card. Technical report, OpenAI, 08 2024. URL <https://cdn.openai.com/gpt-4o-system-card.pdf>. System card / technical report.
- 726 OpenAI. Gpt-5 system card. Technical report, OpenAI, 08 2025. URL <https://cdn.openai.com/gpt-5-system-card.pdf>. System card / technical report.
- 729 Jiefu Ou, William Gantt Walden, Kate Sanders, Zhengping Jiang, Kaiser Sun, Jeffrey Cheng,  
730 William Jurayj, Miriam Wanner, Shaobo Liang, Candice Morgan, et al. Claimcheck: How  
731 grounded are llm critiques of scientific papers? *arXiv preprint arXiv:2503.21717*, 2025.
- 732 Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d  
733 diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. URL  
734 <https://openreview.net/forum?id=FjNys5c7VyY>.
- 735 Sukannya Purkayastha, Zhuang Li, Anne Lauscher, Lizhen Qu, and Iryna Gurevych. Lazyreview a  
736 dataset for uncovering lazy thinking in nlp peer reviews. *arXiv preprint arXiv:2504.11042*, 2025.
- 738 Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Alma-  
739 hairi. Progressive prompts: Continual learning for language models. In *The Eleventh Interna-  
740 tional Conference on Learning Representations*, 2023. URL [https://openreview.net/  
741 forum?id=UJTgQBc91\\_](https://openreview.net/forum?id=UJTgQBc91_).
- 742 Laura Eline Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward  
743 Grefenstette. Large language models are not zero-shot communicators, 2023. URL <https://openreview.net/forum?id=WgbcOQMNXB>.
- 745 Abdelrahman Sadallah, Tim Baumgärtner, Iryna Gurevych, and Ted Briscoe. The good, the bad  
746 and the constructive: Automatically measuring peer review’s utility for authors. *arXiv preprint  
747 arXiv:2509.04484*, 2025.
- 748 Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine  
749 Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker,  
750 Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, De-  
751 bajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen,  
752 Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen,  
753 Abheesh Sharma, Andrea Santilli, Thibault Evry, Jason Alan Fries, Ryan Teehan, Teven Le  
754 Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. Multitask prompted  
755 training enables zero-shot task generalization. In *International Conference on Learning Re-  
presentations*, 2022. URL <https://openreview.net/forum?id=9Vrb9D0WI4>.

- 756 Nihar B. Shah. Challenges, experiments, and computational solutions in  
757 peer review. *Communications of the ACM*, 65(6):76–87, June 2022.  
758 doi: 10.1145/3528086. URL [https://cacm.acm.org/research/  
759 challenges-experiments-and-computational-solutions-in-peer-review/](https://cacm.acm.org/research/challenges-experiments-and-computational-solutions-in-peer-review/).
- 760 Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei  
761 Yao, and Kurt Keutzer. How much can CLIP benefit vision-and-language tasks? In *International  
762 Conference on Learning Representations*, 2022. URL [https://openreview.net/forum?  
763 id=zf\\_Ll3HZWgy](https://openreview.net/forum?id=zf_Ll3HZWgy).
- 764 Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi,  
765 Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Lan-  
766 guage models are multilingual chain-of-thought reasoners. In *The Eleventh International Confer-  
767 ence on Learning Representations*, 2023. URL [https://openreview.net/forum?id=  
768 fR3wGCK-IXp](https://openreview.net/forum?id=fR3wGCK-IXp).
- 769 Hyungyu Shin, Jingyu Tang, Yoonjoo Lee, Nayoung Kim, Hyunseung Lim, Ji Yong Cho, Hwajung  
770 Hong, Moontae Lee, and Juho Kim. Mind the blind spots: A focus-level evaluation framework  
771 for llm reviews. *arXiv preprint arXiv:2502.17086*, 2025.
- 772 Satya Narayan Shukla and Benjamin Marlin. Multi-time attention networks for irregularly sampled  
773 time series. In *International Conference on Learning Representations*, 2021. URL [https://  
774 openreview.net/forum?id=4c0J6lwQ4\\_](https://openreview.net/forum?id=4c0J6lwQ4_).
- 775 Satya Narayan Shukla and Benjamin Marlin. Heteroscedastic temporal variational autoencoder for  
776 irregularly sampled time series. In *International Conference on Learning Representations*, 2022.  
777 URL <https://openreview.net/forum?id=Az7opqbQE-3>.
- 778 Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber,  
779 and Lijuan Wang. Prompting GPT-3 to be reliable. In *The Eleventh International Confer-  
780 ence on Learning Representations*, 2023. URL [https://openreview.net/forum?id=  
781 98p5x51L5af](https://openreview.net/forum?id=98p5x51L5af).
- 782 Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry  
783 Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-  
784 video: Text-to-video generation without text-video data. In *The Eleventh International Confer-  
785 ence on Learning Representations*, 2023. URL [https://openreview.net/forum?id=  
786 nJfylDvgz1q](https://openreview.net/forum?id=nJfylDvgz1q).
- 787 Ivan Stelmakh, Nihar B. Shah, Aarti Singh, and Hal Daumé III. A novice-reviewer experiment to  
788 address scarcity of qualified reviewers in large conferences. In *Proceedings of the AAAI Confer-  
789 ence on Artificial Intelligence*, volume 35, pp. 4785–4793, 2021. doi: 10.1609/aaai.v35i6.16610.  
790 URL <https://ojs.aaai.org/index.php/AAAI/article/view/16610>.
- 791 Taylor and Francis. Understanding the peer review process. what is peer review? a guide  
792 for authors, 2025. URL [https://authorservices.taylorandfrancis.com/  
793 publishing-your-research/peer-review](https://authorservices.taylorandfrancis.com/publishing-your-research/peer-review).
- 794 Huiqiang Wang, Jian Peng, Feihu Huang, Jince Wang, Junhui Chen, and Yifei Xiao. MICN: Multi-  
795 scale local and global context modeling for long-term series forecasting. In *The Eleventh Interna-  
796 tional Conference on Learning Representations*, 2023. URL [https://openreview.net/  
797 forum?id=zt53IDUR1U](https://openreview.net/forum?id=zt53IDUR1U).
- 798 Matthijs J Warrens. Some paradoxical results for the quadratically weighted kappa. *Psychometrika*,  
799 77(2):315–323, 2012.
- 800 Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du,  
801 Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *Interna-  
802 tional Conference on Learning Representations*, 2022. URL [https://openreview.net/  
803 forum?id=gEZrGCozdqR](https://openreview.net/forum?id=gEZrGCozdqR).
- 804 An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. 05 2025. URL [https://  
805 arxiv.org/abs/2505.09388](https://arxiv.org/abs/2505.09388). Covers the Qwen3 family including the 235B A22B series  
806 (e.g., 2507 checkpoint).

810 Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan  
811 Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International  
812 Conference on Learning Representations*, 2023. URL [https://openreview.net/forum?  
813 id=WE\\_vluYUL-X](https://openreview.net/forum?id=WE_vluYUL-X).

814 Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network  
815 for irregularly sampled multivariate time series. In *International Conference on Learning Repre-  
816 sentations*, 2022. URL <https://openreview.net/forum?id=Kwm8I7dU-15>.

817 Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency  
818 for multivariate time series forecasting. In *The Eleventh International Conference on Learning  
819 Representations*, 2023. URL <https://openreview.net/forum?id=vSVLM2j9eie>.

820 Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in  
821 large language models. In *The Eleventh International Conference on Learning Representations*,  
822 2023. URL <https://openreview.net/forum?id=5NTt8GFjUHkr>.

823 Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schu-  
824 urmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. Least-to-most prompting  
825 enables complex reasoning in large language models. In *The Eleventh International Confer-  
826 ence on Learning Representations*, 2023a. URL [https://openreview.net/forum?id=  
827 WZH7099tgfM](https://openreview.net/forum?id=WZH7099tgfM).

828 Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and  
829 Jimmy Ba. Large language models are human-level prompt engineers. In *The Eleventh Interna-  
830 tional Conference on Learning Representations*, 2023b. URL [https://openreview.net/  
831 forum?id=92gvk82DE-](https://openreview.net/forum?id=92gvk82DE-).

832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863

## A FUTURE IMPACT

We expect automatic REVIEWSCORE evaluation could greatly benefit different roles of the current peer reviewing system. For authors, by providing reconstruction of argumentative review points, it helps to understand or clarify the reviewer’s intention and to formulate a rebuttal. For reviewers, by providing REVIEWSCORE of their review points, it allows them to verify the review quality by themselves and helps reviewers to better understand the paper. For metareviewers, by providing REVIEWSCORE of each reviewer, it could assist their final decisions. To summarize, it could serve as an automated system for managing the review quality.

## B REVIEW QUALITY CRITERIA DISCUSSION

First, we recruit a group of three graduate students studying AI and NLP and let them independently analyze reviews of ten common manually selected submitted papers from ICLR 2021–2023<sup>11</sup>. Specifically, we guided the group to: decompose a review into several independent review points, evaluate quality of the review points by a 5-point scale with their own criteria, and justify their scores that ends with a meta-sentence that does not involve paper’s context which is used for our further analysis. During human evaluation, we minimize our effort to provide detailed guidelines to facilitate bias-free human analysis.

Then, we categorize these justifications in order to find common features of low-quality review points. There are five common types: (1) questions that can be already addressed by the paper, (2) comments that reflect a misunderstanding of the paper, (3) out-of-scope remarks, (4) observations pointing out minor details, and (5) unclear points. However, for the last three types, we observe that majority of review points are agreed or argued by a single human annotator. In other words, given a review point, only a single human argues it is out-of-scope, whereas other two humans argue it is within scope and could be a potential drawback of the submitted paper. Following this pattern, a single human argues a review point is addressing minor details but others argue it is a major point, and a single human argues a review point is unclear whereas others do not agree with this. In contrast, the first two types of review points (i.e., answerable questions and misunderstood comments) are mostly agreed by two or more human annotators, indicating that these two types have more objective and trustworthy criteria for detecting low-quality reviews. Based on these observations, we focus on evaluating review points based on the first two criteria.

Lastly, we share some meaningful insights from this group discussion that strongly motivates our work. Based on the analysis of score justifications, given a review point, some (sub)sentences are high-quality or factually correct but others could be misinformed or factually incorrect. This later motivates our premise-level factuality evaluation (Section 3.1). Furthermore, during the group discussion, we observe that human annotators sometimes struggle which parts in a review point they should weigh more to evaluate the review point. This later motivates our aggregation methods which include logical conjunction and weighted average by untrivialness (Section 3.1). We also discover that human evaluations are sometimes incorrect. We leave this as a limitation of our work. However, to alleviate this issue, we recruit three human annotators for every instance and further ensure the annotation quality by providing them careful guidelines and actively communicating with them to build a global consensus on the evaluation criteria (Section 3.3).

---

<sup>11</sup>Since those papers include papers written by authors of our work, we do not share the full list of target papers to keep anonymity. We will uncover the full list after the paper is published.

## C AUTOMATIC ARGUMENT RECONSTRUCTION DETAILS AND RESULTS

### C.1 IMPLEMENTATION DETAILS

We elaborate components of automatic argument reconstruction engine (Section 3.2, Figure 2) in the following. We also refer the corresponding prompt used in each step if it exists.

1. Given an argumentative review point, an LLM extracts a verbatim conjecture and its verbatim reason statements (Figure 23).
2. Given a verbatim conclusion and reasons of the argumentative review point and a corresponding paper parsed from a pdf file, an LLM reconstructs the argument into a premise-conclusion structure. At the same inference, the model also translates (or formalizes) NL premises and conclusion into corresponding FOL formulas and generate keys which assign NL meaning to variables and predicates. To facilitate the model to generate a *valid* formalization, the model generates a deductive proof using formalized premises and conclusion at the end (Figure 24).
3. Given FOL premises and conclusion with keys and a deductive proof, an LLM extracts necessary FOL premises for the deductive proof, write a python program using Z3 (De Moura & Bjørner, 2008) that automatically checks the *validity* of the necessary FOL premises and conclusion, extracts a final FOL conclusion that is used in the python program, and judges whether the proof is circular (i.e., whether the final FOL conclusion is included in one of necessary FOL premises.) (Figure 25, Figure 26).
4. If the proof is circular, then a NL feedback indicating circularity of the proof is sent to Step 2, and the model re-generates an argument reconstruction. Otherwise, we run the python program that checks the *validity* of the reconstruction. However, if the program returns an error, the model takes this error message and re-generate the python program that fixes the error (Figure 27). If the reconstruction is *invalid*, then a NL feedback indicating *invalidity* of the reconstruction is sent to Step 2, and the model re-generates an argument reconstruction. Otherwise (i.e., if the reconstruction is *valid*), we proceed to the next step.
5. To check *faithfulness* of the reconstruction, the model first translates FOL premises and conclusion (one of the outputs of the model in Step 2) with keys (one of the outputs of the model in Step 1) into NL premises and conclusion (Figure 28). This process is called *logical streamlining* in logic and critical thinking (Bowell & Kemp, 2014; Brun & Betz, 2016; Betz & Richardson, 2021).
6. Lastly, given an original argumentative review point (or an argument) and the *streamlined* NL premises and conclusion, the model judges whether the reconstruction is *faithful* with justifications (Figure 29). If the reconstruction is *unfaithful*, then a NL feedback including the model’s justifications is sent to Step 2, and the model re-generates an argument reconstruction accordingly. Otherwise, since the reconstruction is *valid* yet *faithful*, the feedback loop is finished and the *streamlined* NL premises and conclusion become a final argument reconstruction.

## C.2 QUANTITATIVE RESULTS

To verify effectiveness of the feedback loop, we report the argument reconstruction performance with and without feedback in Table 5. We mainly measure average *validity* and *faithfulness* of the reconstructed arguments using a SAT solver and human annotators, respectively. We provide score rubric for evaluating *faithfulness* in Figure 4. We also include the pass rate, which indicates whether the (last) reconstructed argument fulfill the *validity* and *faithfulness* criteria in the feedback loop. Furthermore, we report the average number of loop iterations to check if the feedback loops are actively used.

In Table 5, we verify that the proposed method (i.e., w/ feedback) achieves a perfect *validity* and nearly perfect *faithfulness* and pass rate, whereas the performance of the direct reconstruction (i.e., w/o feedback) largely lags behind that. We observe that the average number of loop iterations is 3.09, indicating active usage of feedback signals, but the number varies a lot depending on the clarity of logical structures of the arguments.

Table 5: Performance of Automatic Argument Reconstruction.

	w/o feedback	w/ feedback
Validity	0.895	<b>1.00</b>
Faithfulness	2.91	<b>4.47</b>
Pass Rate	0.369	<b>0.935</b>
# of loops	N/A	$3.09 \pm 2.73$

### Faithfulness Score Rubric for Argument Reconstruction

- 5: Definitely faithful, no change is essential
- 4: Faithful, but one or two minor details in premises need to change in order to fully express the original context
- 3: Faithful, but more than two minor details in premises need to change
- 2: Not faithful, one or two major error exist
- 1: Definitely not faithful, more than two major errors exist

Figure 4: Score rubric for evaluating *faithfulness* of argument reconstruction by human annotators.

## C.3 QUALITATIVE RESULTS

We present three representative examples of model-generated argument reconstruction and compare outputs with and without the feedback loop. We observe the direct reconstruction (i.e., w/o feedback) often generates either *invalid* or *unfaithful* reconstructed arguments, however, the proposed method (i.e., w/ feedback) generates *valid* yet *faithful* reconstructed arguments. We provide detailed analysis in Figure 5–7.

1026  
1027  
1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079

Input: Argument (Review Point) & Paper
<p><i>Paper:</i> Least-to-Most Prompting Enables Complex Reasoning in Large Language Models (Zhou et al., 2023a)</p> <hr/> <p><i>Argument (Review Point):</i> The paper lacks insights on key questions. Does the model have ability to decompose questions very different from demo examples into correct sub-problems? Even the model solves each sub-problem correctly, what makes the model compose the final solution correctly?</p>
Output: Reconstructed Argument
<p><i>Direct Reconstruction (w/o feedback):</i> P1: The paper does not demonstrate whether the model has the ability to decompose questions very different from demo examples into correct sub-problems. P2: The paper does not explain what makes the model compose the final solution correctly even when it solves each sub-problem correctly. P3: If a paper fails to provide insights on key questions about a proposed method, then the paper lacks insights on key questions. P4: Questions about decomposition ability and solution composition are key questions for the least-to-most prompting method. ∴ C: The paper lacks insights on key questions.</p> <hr/> <p><i>Proposed Method (w/ feedback):</i> P1: The paper does not demonstrate whether the model has the ability to decompose questions very different from demo examples into correct sub-problems. P2: If the paper does not demonstrate whether the model has the ability to decompose questions very different from demo examples into correct sub-problems, then the paper fails to provide insights on the key question about decomposition ability. P3: If the paper fails to provide insights on the key question about decomposition ability or fails to provide insights on the key question about solution composition, then the paper fails to provide insights on key questions. P4: If the paper fails to provide insights on key questions, then the paper lacks insights on key questions. ∴ C: The paper lacks insights on key questions.</p>
Analysis
<ul style="list-style-type: none"> <li>- For the direct reconstruction, P1, P2, P4 semantically implies an intermediate conclusion that “The paper fails to provide insights on key questions about the proposed method.”, but this connecting premise is not explicitly reconstructed. Therefore, this reconstruction is <i>invalid</i> [✗]. However, if we consider that this connecting premise is implicitly presumed, then the reconstruction <i>faithfully</i> [✓] represents the original argument.</li> <li>- For the proposed method, the reconstruction is logically <i>valid</i> [✓] yet <i>faithfully</i> [✓] represents the original argument.</li> </ul>

Figure 5: Example #1 of automatic argument reconstruction.

1080  
1081  
1082  
1083  
1084  
1085  
1086  
1087  
1088  
1089  
1090  
1091  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099  
1100  
1101  
1102  
1103  
1104  
1105  
1106  
1107  
1108  
1109  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1130  
1131  
1132  
1133

Input: Argument (Review Point) & Paper
<p><i>Paper:</i> ReAct: Synergizing Reasoning and Acting in Language Models (Yao et al., 2023)</p> <p>-----</p> <p><i>Argument (Review Point):</i> ReAct alone works fairly well for some tasks, but self-consistent Chain of Thought often outperforms ReAct alone, and leads to fairly dramatic increases in performance. The best models use a smart combination of both CoT and ReAct, using heuristics that likely maximize performance on the dataset(s) (see ‘ReAct → CoT-SC’ strategy), which makes the performance of ReAct by itself slightly less impressive (though I still think this is a useful contribution for the community)</p>
Output: Reconstructed Argument
<p><i>Direct Reconstruction (w/o feedback):</i> P1: The best models use a smart combination of both CoT and ReAct. P2: Models use heuristics that likely maximize performance on the dataset(s) (see ‘ReAct → CoT-SC’ strategy). P3: If models use heuristics that likely maximize performance on the dataset(s), then the performance of ReAct by itself is less impressive. P4: If the best models use a combination of both CoT and ReAct rather than ReAct alone, then the performance of ReAct by itself is less impressive. P5: <b>The performance of ReAct by itself being less impressive still means it is a useful contribution for the community.</b> ∴ C: <b>ReAct alone works fairly well for some tasks, but self-consistent Chain of Thought often outperforms ReAct alone, and leads to fairly dramatic increases in performance.</b></p> <p>-----</p> <p><i>Proposed Method (w/ feedback):</i> P1: ReAct alone works fairly well for some tasks. P2: Self-consistent Chain of Thought often outperforms ReAct alone. P3: The best models use a smart combination of both CoT and ReAct. P4: Models use heuristics like the ‘ReAct → CoT-SC’ strategy that maximize performance. P5: If Self-consistent Chain of Thought often outperforms ReAct alone, and the best models use a smart combination of both CoT and ReAct, and models use heuristics like the ‘ReAct → CoT-SC’ strategy that maximize performance, then the performance of ReAct by itself is slightly less impressive. P6: If ReAct alone works fairly well for some tasks, and the performance of ReAct by itself is slightly less impressive, and a technique can be a useful contribution even if its performance is less impressive, then ReAct is still a useful contribution for the community. P7: A technique can be a useful contribution even if its performance is less impressive. ∴ C: The performance of ReAct by itself is slightly less impressive, and ReAct is still a useful contribution for the community.</p>
Analysis
<p>- For the direct reconstruction, both pairs of premises, P1 &amp; P4, and P2 &amp; P3, deductively imply an intermediate conclusion that “The performance of ReAct by itself is less impressive.”. However, there is no logical connection between this intermediate conclusion and the final conclusion C, which indicates the reconstruction is <i>invalid</i> [X]. Furthermore, the final conclusion C does not correctly represent the original argument, which means the reconstruction is <i>unfaithful</i> [X].</p> <p>- For the proposed method, the premises deductively imply the final conclusion (i.e., <i>valid</i> [✓]), and the reconstruction correctly yet completely represents the original argument (i.e., <i>faithful</i> [✓]), including the correct final conclusion.</p>

Figure 6: Example #2 of automatic argument reconstruction.

1134  
1135  
1136  
1137  
1138  
1139  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187

Input: Argument (Review Point) & Paper
<p><i>Paper:</i> Automatic Chain of Thought Prompting in Large Language Models (Zhang et al., 2023)</p> <p>-----</p> <p><i>Argument (Review Point):</i> Finally, if clustering was the way to go, the authors could have tried to cluster samples into <math>n</math>-clusters, see where the test sample is fitting, and could have tried to use just one prompt instead of <math>k</math> different ones at each time. This might have reduced the compute cost and would be interesting to see the performance vs resource trade-off.</p>
Output: Reconstructed Argument
<p><i>Direct Reconstruction (w/o feedback):</i></p> <p>P1: The authors could have tried to cluster samples into <math>n</math>-clusters and see where the test sample is fitting.</p> <p>P2: The authors could have tried to use just one prompt instead of <math>k</math> different ones at each time.</p> <p>P3: Using one prompt instead of <math>k</math> different ones would reduce the compute cost.</p> <p>P4: If a method reduces compute cost, it would be interesting to see the performance vs resource trade-off.</p> <p>P5: If the authors could have done something that would be interesting to see, then they should have done it.</p> <p>∴ C: The authors should have tried to cluster samples into <math>n</math>-clusters, see where the test sample is fitting, and use just one prompt instead of <math>k</math> different ones at each time.</p> <p>-----</p> <p><i>Proposed Method (w/ feedback):</i></p> <p>P1: An alternative approach would be to cluster samples into <math>n</math>-clusters, see where the test sample is fitting, and use just one prompt instead of <math>k</math> different ones.</p> <p>P2: If an alternative approach would use just one prompt instead of <math>k</math> different ones, then the alternative approach would reduce compute cost.</p> <p>P3: If the alternative approach would reduce compute cost, then examining the performance vs. resource trade-off of the alternative approach would be interesting.</p> <p>P4: If an alternative approach would use just one prompt instead of <math>k</math> different ones and examining the performance vs. resource trade-off of the alternative approach would be interesting, then the alternative approach would have been interesting to evaluate for its performance vs. resource trade-off.</p> <p>∴ C: The alternative approach would have been interesting to evaluate for its performance vs. resource trade-off.</p>
Analysis
<p>- For the direct reconstruction, P3 and P4 imply an intermediate conclusion that “Using one prompt instead of <math>k</math> different ones would be interesting to see the performance vs resource trade-off.”. Let us call this IC1. IC1 and P5 imply another intermediate conclusion that “The authors should have tried using one prompt instead of <math>k</math> different ones.”. Let us call this IC2. Lastly, P1, P2, and IC2 imply the final conclusion C, indicating the reconstruction is <i>valid</i> [✓]. However, the final conclusion is stronger than what the original argument states. The original argument suggests an alternative approach, but the reconstructed final conclusion obligates authors to try it, meaning the reconstruction is <i>unfaithful</i> [✗].</p> <p>- For the proposed method, the reconstruction is <i>valid</i> [✓] and <i>faithful</i> [✓], including the correct degree of strength of the final conclusion.</p>

Figure 7: Example #3 of automatic argument reconstruction.

## D DATASET DETAILS

### D.1 TERMS OF USE AND LICENSE

All papers and corresponding reviews used in our work are crawled from OpenReview<sup>12</sup>, and our work is consistent with the Openreview terms of use<sup>13</sup>. Following these terms, we will release our work with a CC-BY 4.0 license.

### D.2 PILOT DATASET

Before we construct the main dataset in Section 3.3, to ensure a trustworthiness and reliability of human annotation, we conduct a pilot study. We recruit three graduate students studying AI as human annotators, and let them choose total five papers submitted to ICLR 2021–2025<sup>14</sup>. Specifically, each human annotates reviews of three papers in OpenReview, where only one paper is authored by themselves and the other two non-authored papers are assigned in common. To ensure trustworthiness of non-authors’ annotation, three humans annotate reviews and then we take a median value as a final human label.

### D.3 DATASET STATISTICS

We present the REVIEWSCORE dataset statistics including the pilot and main subsets in Table 6. We include number of instances and percentage of *misinformed* labels.

Table 6: REVIEWSCORE dataset statistics.

	Pilot	Main	Total
ICLR Years	2021–2025	2021–2023	-
# Papers	5	40	45
# Review(er)s	19	155	174
<i>Number of instances</i>			
Review Points	84	573	657
Questions	22	121	143
Claims	5	87	92
Arguments	57	365	422
Premises	227	1,521	1,748

### D.4 FULL LIST OF PAPERS

We provide a full list of papers used in the main subset of REVIEWSCORE dataset in Table 7. As explained in Section 3.3, eight papers are selected by each human annotator group with a common research interest.

## E ANALYSIS OF HUMAN ANNOTATORS

### E.1 HUMAN ANNOTATOR EXPERTISE

We report human annotator’s expertise on REVIEWSCORE evaluation in Table 8. Specifically, for each annotator, we indicate a number of publications in AI/ML (including arXived works) and an averaged paper relevance score across eight assigned papers. We also indicate averaged numbers and scores for each group. The results show that an average number of publication is 3.93 and average paper relevance score is 4.06 / 5, indicating highly-experienced and relevant experts conduct a human annotation process. However, we observe that there are inter-group gaps in human expertise. Specifically, Group 2 and 4 shows significantly higher number of publications and paper relevance

<sup>12</sup><https://openreview.net>

<sup>13</sup><https://openreview.net/legal/terms>

<sup>14</sup>Unlike the main dataset, we include ICLR 2024–2025 since authors can access to their own submissions.

Table 7: Full list of ICLR submitted papers used in the REVIEWSCORE dataset.

Group	Topic	Paper Title
1	Image Generation	Lipman et al. (2023), Hertz et al. (2023), Couairon et al. (2023), Singer et al. (2023), Meng et al. (2022), Gal et al. (2023), Poole et al. (2023), Kim et al. (2023)
2	Time Series ML	Shukla & Marlin (2021), Zhang et al. (2022), Borisov et al. (2023), Nie et al. (2023), Nam et al. (2023), Zhang & Yan (2023), Wang et al. (2023), Shukla & Marlin (2022)
3	LLM Reasoning / Compression	Hsu et al. (2022), Ji et al. (2023), Lu et al. (2023), Ni et al. (2023), Fu et al. (2023), Liang et al. (2021), Ruis et al. (2023), Arora et al. (2023)
4	LLM / VLM	Hu et al. (2022), Kukleva et al. (2023), Wei et al. (2022), Sanh et al. (2022), Shen et al. (2022), Zhou et al. (2023b), Si et al. (2023), Shi et al. (2023)
5	LLM Prompting	Zhang et al. (2023), Yao et al. (2023), Zhou et al. (2023a), Razdaibiedina et al. (2023), Li et al. (2022), Jang et al. (2022), Creswell et al. (2023), Huang et al. (2023)

than other groups. Detailed score rubric for paper relevance is described in Figure 8, and we note that there is no instance where paper relevance of any human annotator is less than 3 (i.e., Moderate relevance).

Table 8: Human annotator’s expertise on REVIEWSCORE evaluation. A **bold** indicates the highest number/score across different groups, and an underline indicates the second highest.

Group	Annotator ID	# AI/ML Pub	Paper Relevance	Avg # Pub	Avg Paper Relevance
1	anno_11	2	3.50	3.00	4.00
	anno_12	3	4.13		
	anno_13	4	4.38		
2	anno_21	6	4.50	<u>5.33</u>	<b>4.38</b>
	anno_22	3	4.38		
	anno_23	7	4.25		
3	anno_31	0	4.00	1.33	3.96
	anno_32	1	4.00		
	anno_33	3	3.88		
4	anno_41	1	4.00	<b>6.67</b>	<u>4.04</u>
	anno_42	6	3.88		
	anno_43	13	4.25		
5	anno_51	3	3.13	3.33	3.92
	anno_52	1	4.13		
	anno_53	6	4.50		
Total	-	-	-	3.93	4.06

**Paper Relevance Score Rubric**

5: Direct expertise – Works in the exact subtopic and can judge nuanced claims, methods, and datasets.  
 4: Strong relevance – Adjacent/overlapping subtopic with regular use of the paper’s methods or domain; can evaluate technical choices with minimal ramp-up.  
 3: Moderate relevance – Same broad area (e.g., NLP ↔ NLP; CV ↔ CV) but different subtopic or methods; will understand contributions but may miss edge-case nuances.  
 2: Low relevance – Only tangential connection (e.g., general ML experience while the paper is domain-specific) and limited familiarity with core methods or domain.  
 1: No clear relevance – Outside the field; would require substantial background reading to assess claims/methodology.

Figure 8: Score rubric for evaluating paper relevance of human annotators.

## E.2 INTER-ANNOTATOR AGREEMENT

To ensure trustworthiness of human annotation, we report inter-annotator agreement in Krippendorff’s Alpha (Hughes, 2021) on REVIEWSCORE evaluation in Table 9. Overall, REVIEWSCORE shows 0.301 Krippendorff’s Alpha, indicating a moderate inter-annotator agreement. Specifically, ClaimScore shows the highest agreement, QScore follows subsequently, and ArgScore shows the lowest agreement.

By comparing inter-group agreements, Group 2 and 4 show significantly higher REVIEWSCORE agreement than other groups. In contrast, Group 1 and 5 include negative inter-annotator agreement, which is a critical concern for reliable human annotation. By a manual disagreement analysis, we confirm that most disagreements come from human annotators with low paper relevance (i.e., anno\_11 and anno\_51 in Table 8). To summarize, we observe that the inter-annotator agreement aligns with the annotator expertise (Table 8), and we (unsurprisingly) confirm that human annotations could be more reliable if their research interests become more relevant to the assigned papers.

Table 9: Inter-annotator agreement (Krippendorff’s Alpha) on REVIEWSCORE evaluation. A **bold** indicates the highest agreement across different groups, and an underline indicates the second highest.

Group	ClaimScore	ArgScore	QScore	REVIEWSCORE
1	<b>0.399</b>	-0.023	0.245	0.125
2	0.328	<u>0.329</u>	0.350	<b>0.428</b>
3	<b>0.400</b>	0.146	<u>0.436</u>	0.301
4	0.375	<b>0.361</b>	<b>0.465</b>	<u>0.385</u>
5	0.203	-0.023	-0.078	-0.041
Median	0.375	0.146	0.350	0.301

## F ADDITIONAL RESULTS OF AUTOMATIC REVIEWSCORE EVALUATION

### F.1 QUANTITATIVE RESULTS

We additionally report human-model agreement on REVIEWSCORE evaluation using different evaluation metrics in Table 10–13. Specifically, we report agreement on ClaimScore evaluation in Table 10, agreement on ArgScore evaluation in Table 11, agreement on QScore evaluation in Table 12, and agreement on REVIEWSCORE evaluation in Table 13. For the binary classification setup, we use Precision, Recall, and F1 Score, and for the 5-point scale setup, we use Pearson rank correlation, Gwet’s AC2 (Gwet, 2001), and Quadratic Weighted Kappa (Warrens, 2012).

Table 10: Human-model agreement on ClaimScore evaluation.

Model	Binary			5-point Scale		
	Precision	Recall	F1	Pearson	AC2	Kappa
<i>Proprietary models</i>						
claude-sonnet-3.7	<b>0.091</b>	<b>0.667</b>	<b>0.160</b>	<b>0.192</b>	0.035	<b>0.137</b>
claude-sonnet-4	<b>0.091</b>	0.200	0.125	0.172	0.048	0.130
gpt-4o	0.000	0.000	0.000	0.130	<b>0.064</b>	0.119
gpt-5	0.083	0.250	0.125	-0.014	-0.057	-0.004
gemini-2.5-flash	0.059	0.200	0.091	0.146	0.054	0.106
<i>Open-sourced models</i>						
qwen3-235b-a22b	<b>0.136</b>	<b>0.750</b>	<b>0.231</b>	0.135	0.045	0.094
llama-3.3	0.091	0.250	0.133	0.144	<b>0.122</b>	0.153
deepseek-v3	0.000	0.000	0.000	<b>0.195</b>	0.086	<b>0.165</b>

1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359  
1360  
1361  
1362  
1363  
1364  
1365  
1366  
1367  
1368  
1369  
1370  
1371  
1372  
1373  
1374  
1375  
1376  
1377  
1378  
1379  
1380  
1381  
1382  
1383  
1384  
1385  
1386  
1387  
1388  
1389  
1390  
1391  
1392  
1393  
1394  
1395  
1396  
1397  
1398  
1399  
1400  
1401  
1402  
1403

Table 11: Human-model agreement on ArgScore evaluation.

Model	Binary			5-point Scale		
	Precision	Recall	F1	Pearson	AC2	Kappa
<i>Proprietary models</i>						
claude-sonnet-3.7	0.345	<b>0.679</b>	<b>0.458</b>	<b>0.425</b>	<b>0.369</b>	<b>0.380</b>
claude-sonnet-4	0.250	0.482	0.329	0.258	0.239	0.266
gpt-4o	<b>0.483</b>	0.250	0.329	0.280	0.221	0.239
gpt-5	0.353	0.536	0.426	0.368	0.329	0.334
gemini-2.5-flash	0.333	<b>0.679</b>	0.447	0.419	0.358	0.366
<i>Open-sourced models</i>						
qwen3-235b-a22b	<b>0.381</b>	<b>0.429</b>	<b>0.403</b>	<b>0.281</b>	0.012	0.149
llama-3.3	0.344	0.196	0.250	-0.014	0.124	0.107
deepseek-v3	0.267	0.286	0.276	0.238	<b>0.188</b>	<b>0.196</b>

Table 12: Human-model agreement on QScore evaluation.

Model	Binary			5-point Scale		
	Precision	Recall	F1	Pearson	AC2	Kappa
<i>Proprietary models</i>						
claude-sonnet-3.7	<b>0.571</b>	0.516	<b>0.542</b>	<b>0.394</b>	<b>0.380</b>	<b>0.387</b>
claude-sonnet-4	0.462	0.563	0.507	0.387	0.335	0.357
gpt-4o	0.447	0.677	0.538	<b>0.395</b>	0.298	0.340
gpt-5	0.500	0.594	<b>0.543</b>	0.320	0.285	0.300
gemini-2.5-flash	0.386	<b>0.759</b>	0.512	0.283	0.154	0.225
<i>Open-sourced models</i>						
qwen3-235b-a22b	0.394	0.813	0.531	0.325	0.117	0.234
llama-3.3	0.409	<b>0.844</b>	0.551	<b>0.375</b>	<b>0.196</b>	<b>0.284</b>
deepseek-v3	<b>0.469</b>	0.821	<b>0.597</b>	0.325	0.101	0.228

Table 13: Human-model agreement on REVIEWSCORE evaluation.

Model	Binary			5-point Scale		
	Precision	Recall	F1	Pearson	AC2	Kappa
<i>Proprietary models</i>						
claude-sonnet-3.7	0.350	0.622	<b>0.448</b>	0.380	0.339	0.355
claude-sonnet-4	0.291	0.495	0.367	0.360	0.308	0.333
gpt-4o	<b>0.417</b>	0.393	0.405	<b>0.389</b>	<b>0.359</b>	<b>0.362</b>
gpt-5	0.370	0.543	0.441	0.327	0.300	0.310
gemini-2.5-flash	0.324	<b>0.678</b>	0.439	0.369	0.296	0.318
<i>Open-sourced models</i>						
qwen3-235b-a22b	0.351	<b>0.576</b>	<b>0.436</b>	0.336	0.161	0.243
llama-3.3	<b>0.358</b>	0.424	0.388	0.004	<b>0.322</b>	<b>0.316</b>
deepseek-v3	0.328	0.459	0.382	<b>0.361</b>	0.278	0.310

1404 F.2 QUALITATIVE RESULTS  
1405

1406 We present a thorough qualitative human-model disagreement analysis on REVIEWSCORE eval-  
1407 uation in Figure 9–14. For each review point type, we demonstrate two representative examples  
1408 which include model-based scores and corresponding justifications using `claude-sonnet-3.7`  
1409 and `claude-sonnet-4`.

1410	1411	1412	1413	1414	1415	1416	1417
<b>Input: Claim (Review Point) &amp; Paper</b>							
<i>Paper:</i> Towards Continual Knowledge Learning of Language Models (Jang et al., 2022)							
-----							
<i>Claim (Review Point):</i> The experimental findings in this paper are somewhat trivial.							
<b>Human-annotated Label (5-point Scale)</b>							
3, 4, 4 (Median: 4)							
<b>Output: ClaimScore Evaluation</b>							
<i>Model:</i> <code>claude-sonnet-3.7</code>							
<i>Score:</i> <b>2</b>							
<i>Human-model Disagreement Type:</i> <b>Value judgment</b>							
<i>Justification:</i> The proposition states that the experimental findings in this paper are somewhat trivial. To evaluate this claim, I need to assess the novelty, significance, and depth of the experimental findings presented in the paper. Key experimental findings in the paper include: .....							
These findings are not trivial because:							
- They establish a new problem formulation and benchmark for an important real-world challenge							
- They provide novel insights about how different methods perform in this setting							
- They identify key factors affecting knowledge retention and acquisition							
- The findings have practical implications for developing ever-changing language models							
-----							
<i>Model:</i> <code>claude-sonnet-4</code>							
<i>Score:</i> <b>2</b>							
<i>Human-model Disagreement Type:</i> <b>Value judgment</b>							
<i>Justification:</i> To evaluate whether the experimental findings are “somewhat trivial”, I need to assess the novelty, significance, and non-obviousness of the results presented in the paper. **Novel contributions and non-trivial findings:** .....							
**Overall assessment:** While some high-level expectations (like catastrophic forgetting occurring) are predictable, the specific findings about which methods work best, the quantitative trade-offs, the multi-phase degradation effects, and the detailed analysis of what causes forgetting provide substantial non-trivial insights. The systematic benchmarking and the discovery that traditional CL methods don’t transfer well to this domain are significant contributions.							
The experimental work is comprehensive, the findings have practical implications for maintaining up-to-date language models, and several results challenge conventional wisdom from traditional continual learning.							

1455 Figure 9: Example #1 of ClaimScore evaluation.  
1456  
1457

1458  
1459  
1460  
1461  
1462  
1463  
1464  
1465  
1466  
1467  
1468  
1469  
1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479  
1480  
1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511

<b>Input: Claim (Review Point) &amp; Paper</b>
<p><i>Paper:</i> Temperature Schedules for self-supervised contrastive methods on long-tail data (Kukleva et al., 2023)</p> <p>-----</p> <p><i>Claim (Review Point):</i> The improvement is quite limited compared with other long-tailed learning methods.</p>
<b>Human-annotated Label (5-point Scale)</b>
2, 2, 3 (Median: 2)
<b>Output: ClaimScore Evaluation</b>
<p><i>Model:</i> claude-sonnet-3.7</p> <p><i>Score:</i> <b>3</b></p> <p><i>Justification:</i> The proposition states that “The improvement is quite limited compared with other long-tailed learning methods”. To evaluate this claim, I need to examine how the paper compares its proposed temperature scheduling (TS) approach with other long-tailed learning methods. Looking at the paper:</p> <p>...</p> <p>Given that the paper shows consistent improvements over the baselines and SDCLR, but doesn’t compare with a broader range of long-tailed learning methods, and the improvements are modest (1-3%), the proposition that “the improvement is quite limited compared with other long-tailed learning methods” is partially supported but not fully substantiated.</p> <p>-----</p> <p><i>Model:</i> claude-sonnet-4</p> <p><i>Score:</i> <b>4</b></p> <p><i>Human-model Disagreement Type:</i> <b>Value judgment</b></p> <p><i>Justification:</i> To evaluate this proposition, I need to examine the experimental results and compare the improvements achieved by the proposed temperature scheduling (TS) method against other long-tailed learning methods. Looking at the experimental results:</p> <ol style="list-style-type: none"> <li><b>Comparison with SDCLR:</b> In Table 3, the paper compares SimCLR+TS with SDCLR (a specialized method for long-tailed contrastive learning). The improvements are modest.</li> <li><b>Absolute improvements:</b> The improvements over baseline methods are generally in the range of 1–4% across different metrics and datasets.</li> <li><b>Limited baseline comparisons:</b> The paper primarily compares against standard contrastive learning methods (MoCo, SimCLR) and only one specialized long-tailed method (SDCLR). It doesn’t compare against other established long-tailed learning approaches from supervised learning or other self-supervised long-tailed methods.</li> <li><b>Scope of evaluation:</b> The evaluation is limited to relatively small datasets (CIFAR variants and ImageNet-100) and doesn’t demonstrate performance on larger-scale long-tailed datasets that are commonly used in long-tailed learning literature.</li> <li><b>Method simplicity vs. improvement:</b> While the method is simple (just scheduling temperature), the improvements are incremental rather than substantial breakthroughs.</li> </ol> <p>The paper acknowledges this limitation implicitly by positioning the method as “simple yet effective” and emphasizing the zero additional computational cost, rather than claiming state-of-the-art performance improvements.</p>

Figure 10: Example #2 of ClaimScore evaluation.

1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565

Input: Premise of Argument (Review Point) & Paper
<p><i>Paper:</i> Automatic Chain of Thought Prompting in Large Language Models (Zhang et al., 2023)</p> <p>-----</p> <p><i>Argument (Review Point):</i> Firstly, when mentioning that the method automatically finds prompts using chain of thoughts (Auto.CoT) approach, this does not imply finding the best samples from the dataset and using the same method as before. In other words, the proposed method does not find the best chain of thought prompts as claimed but rather believes that “let’s do this step by step” is the best zero-shot-CoT prompting strategy and the method merely uses that with different prompt samples. The title and the abstract are highly misleading in that way.</p> <p>-----</p> <p><i>Premise:</i> The method uses “let’s do this step by step” with different prompt samples.</p>
Human-annotated Label (5-point Scale)
5, 5, 5 (Median: 5)
Output: ClaimScore Evaluation
<p><i>Model:</i> claude-sonnet-3.7</p> <p><i>Score:</i> 5</p> <p><i>Justification:</i> The premise states that the Auto-CoT method uses “let’s do this step by step” with different prompt samples. I need to check if this is accurate based on the paper. Looking at the paper, in the “DEMONSTRATION SAMPLING” section, the author describes how they construct demonstrations:</p> <p>“A prompted input is formulated as: [Q: q(i)j. A: [P]], where [P] is a single prompt ‘Let’s think step to step’. This formed input is fed into an LLM using Zero-Shot-CoT (Kojima et al., 2022) to output the reasoning chain consisting of the rationale r(i)j and the extracted answer a(i)j.”</p> <p>...</p> <p>So the premise is accurate - the Auto-CoT method does use “let’s think step by step” (or a very similar variant “Let’s think step to step”) with different prompt samples selected through their clustering and sampling approach.</p> <p>-----</p> <p><i>Model:</i> claude-sonnet-4</p> <p><i>Score:</i> 2</p> <p><i>Human-model Disagreement Type:</i> <b>Overthinking</b></p> <p><i>Justification:</i> To judge the factuality of this premise, I need to examine what the Auto-CoT method actually does according to the paper. From the paper, I can see that:</p> <ol style="list-style-type: none"> <li>1. Auto-CoT consists of two main stages: (i) question clustering and (ii) demonstration sampling</li> <li>2. In the demonstration sampling stage, the method uses Zero-Shot-CoT to generate reasoning chains for selected questions</li> <li>3. The paper explicitly states that Zero-Shot-CoT uses the prompt “Let’s think step by step” (not “let’s do this step by step”)</li> </ol> <p>...</p> <p>So the premise is mostly accurate in describing the approach (using Zero-Shot-CoT with different samples), <b>but contains a factual error in the exact wording of the prompt. The paper consistently uses “Let’s think step by step” rather than “let’s do this step by step.”</b></p>

Figure 11: Example #1 of ArgScore evaluation.

1566  
1567  
1568  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619

Input: Premise of Argument (Review Point) & Paper
<p><i>Paper:</i> Automatic Chain of Thought Prompting in Large Language Models (Zhang et al., 2023)</p> <p>-----</p> <p><i>Argument (Review Point):</i> Secondly, this is very obvious to anyone that the prompts must be representative of the test samples (as much as possible). This work takes the dataset, clusters them, finds the most representative sample from each cluster, and uses that as prompts. There is nothing extra-ordinary about this approach and is merely a small improvement over the randomly sampling strategy. This can also be seen from a marginal improvement in the results over a random sampling strategy.</p> <p>-----</p> <p><i>Premise:</i> The approach is obvious (i.e., prompts must be representative of test samples).</p>
Human-annotated Label (5-point Scale)
4, 4, 5 (Median: 4)
Output: ClaimScore Evaluation
<p><i>Model:</i> claude-sonnet-3.7</p> <p><i>Score:</i> <b>2</b></p> <p><i>Human-model Disagreement Type:</i> <b>Misunderstanding / incorrect reasoning</b></p> <p><i>Justification:</i> I'll examine whether the paper itself acknowledges or implies that prompts should be representative of test samples as an obvious approach:</p> <p style="text-align: center;">...</p> <p>The paper's main contribution is showing that diversity in demonstrations is important, not just representativeness. The paper states: "diversity-based clustering may mitigate misleading by similarity" and "diversity of demonstration questions is the key."</p> <p>This suggests that the premise oversimplifies the paper's findings. The paper doesn't claim that prompts must simply be representative – it argues for a balance of representativeness and diversity.</p> <p>-----</p> <p><i>Model:</i> claude-sonnet-4</p> <p><i>Score:</i> <b>4</b></p> <p><i>Justification:</i> To judge the factuality, I need to determine whether it's true that "prompts must be representative of test samples" is an obvious principle. This is a general statement about machine learning methodology and prompt engineering principles that would be known to AI practitioners before 2023. The idea that training/demonstration examples should be representative of test data is a fundamental principle in machine learning that predates this specific paper. This principle is well-established in the field and doesn't require knowledge of this specific paper to evaluate its truth.</p> <p>The paper does implement this principle through clustering and selecting representative samples, but the general principle itself is established knowledge in the field.</p>

Figure 12: Example #2 of ArgScore evaluation.

1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1630  
1631  
1632  
1633  
1634  
1635  
1636  
1637  
1638  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1670  
1671  
1672  
1673

Input: Question (Review Point) & Paper
<p><i>Paper:</i> ReAct: Synergizing Reasoning and Acting in Language Models (Yao et al., 2023)</p> <p>-----</p> <p><i>Question (Review Point):</i> Is there a way to add self-consistency to ReAct, such that it samples multiple reasoning traces and actions and uses the majority action, for example? SC seems extremely effective for Chain of Thought tasks, and it seems like something similar could work here.</p>
Human-annotated Label (5-point Scale)
2, 2, 5 (Median: 2)
Output: ClaimScore Evaluation
<p><i>Model:</i> claude-sonnet-3.7</p> <p><i>Score:</i> 4</p> <p><i>Human-model Disagreement Type:</i> <b>Minority score prediction</b></p> <p><i>Justification:</i> The question asks about adding self-consistency to ReAct, specifically whether it's possible to sample multiple reasoning traces and actions and use the majority action, similar to how self-consistency works for Chain of Thought tasks. Let me check if the paper discusses this concept or something similar:</p> <p style="text-align: center;">...</p> <p>However, the paper does not explicitly discuss applying the self-consistency technique directly to ReAct itself (i.e., sampling multiple ReAct trajectories and taking the majority action). It only discusses combining ReAct with CoT-SC as separate methods, not applying the self-consistency sampling technique to ReAct's own trajectories.</p> <p>-----</p> <p><i>Model:</i> claude-sonnet-4</p> <p><i>Score:</i> 4</p> <p><i>Human-model Disagreement Type:</i> <b>Minority score prediction</b></p> <p><i>Justification:</i> The question asks about adding self-consistency to ReAct, similar to how it's used in Chain of Thought (CoT) tasks. Let me examine what the paper says about self-consistency and whether it discusses or implements this for ReAct. Looking through the paper, I can see that:</p> <p style="text-align: center;">...</p> <p>However, the question specifically asks about applying self-consistency directly to ReAct itself – sampling multiple ReAct reasoning traces and actions and using majority voting. While the paper discusses CoT-SC and hybrid approaches that combine ReAct with CoT-SC, it doesn't appear to directly implement or discuss a "ReAct-SC" approach where multiple ReAct trajectories are sampled and majority voting is applied to the ReAct outputs themselves.</p>

Figure 13: Example #1 of QScore evaluation.

1674  
1675  
1676  
1677  
1678  
1679  
1680  
1681  
1682  
1683  
1684  
1685  
1686  
1687  
1688  
1689  
1690  
1691  
1692  
1693  
1694  
1695  
1696  
1697  
1698  
1699  
1700  
1701  
1702  
1703  
1704  
1705  
1706  
1707  
1708  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719  
1720  
1721  
1722  
1723  
1724  
1725  
1726  
1727

Input: Question (Review Point) & Paper
<p><i>Paper:</i> Large Language Models Can Self-improve (Huang et al., 2023)</p> <p>-----</p> <p><i>Question (Review Point):</i> Can you check if baseline self-consistency with more examples can match LMSI or not?</p>
Human-annotated Label (5-point Scale)
4, 5, 5 (Median: 5)
Output: ClaimScore Evaluation
<p><i>Model:</i> claude-sonnet-3.7</p> <p><i>Score:</i> <b>2</b></p> <p><i>Human-model Disagreement Type:</i> <b>Misunderstanding / incorrect reasoning</b></p> <p><i>Justification:</i> The paper includes a hyperparameter study in Figure 4(b) that shows the effect of varying the number of sampled reasoning paths for self-consistency, both with and without LMSI. <b>This directly addresses whether increasing the number of examples in baseline self-consistency can match LMSI performance.</b></p> <p>-----</p> <p><i>Model:</i> claude-sonnet-4</p> <p><i>Score:</i> <b>4</b></p> <p><i>Justification:</i> Looking at the experimental setup and results:</p> <ul style="list-style-type: none"> <li>- The paper uses m=32 reasoning paths for generating training data</li> <li>- Figure 4(b) shows accuracy vs number of sampled reasoning paths for both w/o LMSI and LMSI</li> <li>- The experiments seem to use consistent numbers of paths for comparison</li> </ul> <p>The paper doesn't systematically explore whether dramatically increasing the number of paths in baseline self-consistency (e.g., to 100+ paths) could match the LMSI performance. While Figure 4(b) shows diminishing returns after 15–20 paths, it doesn't explore very high numbers of paths that might be computationally equivalent to the LMSI training process.</p>

Figure 14: Example #2 of QScore evaluation.

## G MODEL DETAILS

We measure reliability of automatic REVIEWSCORE evaluation with eight current state-of-the-art LLMs, including five proprietary models and three open-sourced models. For proprietary models, we use Anthropic’s `claude-3-7-sonnet-20250219` (Anthropic, 2025a) and `claude-sonnet-4-20250514` (Anthropic, 2025b), OpenAI’s `gpt-4o-2024-08-06` (OpenAI, 2024) and `gpt-5-2025-08-07` (OpenAI, 2025), and Google’s `gemini-2.5-flash` (Gemini Team, Google, 2025). Since the REVIEWSCORE evaluation does not require longer chain of thought, we exclude reasoning models. Following this rule, we use Anthropic’s claude models without thinking modes and OpenAI’s gpt-5 model with minimal reasoning effort. For open-sourced models, we use Alibaba’s `qwen3-235b-a22b-2507` (Yang et al., 2025) with 235B model parameters, DeepSeek’s `deepseek-v3-0324` (DeepSeek-AI et al., 2024) with 671B model parameters and Meta’s `llama-3.3-70b-instruct` (AI@Meta, 2024) with 70B model parameters. We call the models by their APIs.

## H PROMPTS

We list every prompt used for automatic REVIEWSCORE evaluation (Section 4), automatic review point type classification (used in preprocessing stage of the dataset construction in Section 3.3), and automatic argument reconstruction (Section 3.2). Figure 15 – Figure 20 indicate prompts for automatic REVIEWSCORE evaluation and related score rubrics. Figure 21 and Figure 22 indicate prompts for automatic review point type classification, where the type is one of claim, argument, and question. Figure 23 – Figure 29 indicate prompts for automatic argument reconstruction. In the prompts, capital letters with double square brackets (i.e., `[[XYZ]]`) would be replaced by the corresponding material, and words with a curly bracket (i.e., `{abc}`) would be replaced by the corresponding score rubric (Figure 15 – Figure 17) or a python code snippet in Figure 26.

Unanswerability Score Rubric
1: Definitely Answerable by the paper
2: Probably Answerable by the paper
3: No Verdict
4: Probably Unanswerable by the paper
5: Definitely Unanswerable by the paper

Figure 15: Score rubric for evaluating *unanswerability*.

Factuality Score Rubric
1: Definitely False
2: Probably False
3: No Verdict
4: Probably True
5: Definitely True

Figure 16: Score rubric for evaluating *factuality*.

Untrivialness Score Rubric
0: Definitely Trivial
1: Probably Trivial
2: Definitely Not Trivial

Figure 17: Score rubric for evaluating *untrivialness*.

1782  
1783  
1784  
1785  
1786  
1787  
1788  
1789  
1790  
1791  
1792  
1793  
1794  
1795  
1796  
1797  
1798  
1799  
1800  
1801  
1802  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1830  
1831  
1832  
1833  
1834  
1835

<b>Prompt for evaluating QScore (<i>unanswerability of questions</i>)</b>	
# Paper	[[PAPER]]
# Question	[[QUESTION]]
You are given a paper submitted to an AI conference and a question regarding the paper.	
Judge if the question is answerable or not by the paper. You should scale 1-5 to indicate unanswerability as follows.	
{Unanswerability Score Rubric}	
If you score the question 1 or 2, then indicate which knowledge source you have grounded to (i.e., indicate corresponding section(s) and verbatim sentence(s)) and answer the question.	
Your output should be the following.	
### Reasoning	[think step-by-step]
### Unanswerability	[1 or 2 or 3 or 4 or 5]
### Source	[corresponding section(s) and verbatim sentence(s) if the score is 1 or 2, otherwise None]
### Answer	[answer the question if the score is 1 or 2, otherwise None]

Figure 18: Prompt used for evaluating QScore in a 5-point scale.

<b>Prompt for evaluating ClaimScore or WScore (<i>factuality of weaknesses</i>)</b>	
# Paper	[[PAPER]]
# Weakness	[[WEAKNESS]]
You are given a paper submitted to an AI conference and a weakness regarding the paper.	
Judge if the weakness is true or not based on the paper. You should scale 1-5 to indicate factuality as follows.	
{Factuality Score Rubric}	
Your output should be the following.	
### Reasoning	[think step-by-step]
### Factuality	[1 or 2 or 3 or 4 or 5]

Figure 19: Prompt used for evaluating ClaimScore or WScore in a 5-point scale.

1836 **Prompt for evaluating ArgScore (factuality of premises)**

1837 # Paper

1838 [[PAPER]]

1839 # Weakness

1840 [[WEAKNESS]]

1841

1842 # Premise

1843 [[PREMISE]]

1844

1845 You are given a paper submitted to an AI conference, a weakness of the paper, and one of premises of

1846 the weakness. Your task is to judge the factuality and untrivialness of the given premise.

1847

1848 First, judge the factuality of the premise. To do that, choose an appropriate knowledge source from:

1849 1. given paper

1850 2. annotator knowledge before the year [[YEAR]] (more precisely, before the paper is publicized)

1851 3. other paper(s),

1852 and then judge the factuality of the premise based on the knowledge source. You should scale 1-5 to

1853 indicate factuality as follows.

1854 {Factuality Score Rubric}

1855

1856 Here are guidelines you should follow:

1857 - Main purpose is to distinguish \*given\_paper\* and \*annotator\_knowledge\*.

1858 - Select \*other\_papers\* only if the premise refers a specific paper.

1859 - Note that you should separate judging the factuality of the premise from understanding the semantics

1860 of the premise. It does not matter whether \*given\_paper\* is needed or not to understand the semantics

1861 of the premise. The knowledge source is \*given\_paper\* only if \*given\_paper\* is needed to judge the

1862 factuality of the premise, otherwise, the knowledge source is \*annotator\_knowledge\*. For example,

1863 although you need the paper's context in order to understand what the premise means, if you do not

1864 need the paper's knowledge to judge if the premise is true or not (e.g., logical assessment), then you

1865 should choose \*annotator\_knowledge\* as a knowledge source and judge the factuality accordingly.

1866 - For premises that are conditionals (If A then B), you should presume that the antecedent (A) is al-

1867 ways true even if the antecedent does not align with the paper's knowledge. (Because the antecedent is

1868 always true, the knowledge source should only be determined while judging the factuality of the conse-

1869 quent.) Then, choose an appropriate knowledge source to judge if the consequent (B) is true or not and

1870 judge the factuality accordingly.

1871

1872 Next, decide whether the premise is trivially true or not based on the common knowledge of CS/AI-

1873 majoring undergrad students before the year [[YEAR]] (more precisely, before the paper is publicized).

1874 For premises that are conditionals (If A then B), you must assume that the antecedent (A) is true and

1875 judge if the consequent (B) is trivially true or not. You should scale the score to 0-2 as follows.

1876 {Untrivialness Score Rubric}

1877

1878 Here are guidelines you should follow:

1879 - If the knowledge source is \*given\_paper\* or \*other\_papers\*, then untrivialness should always be 2

1880 unless the premise factuality could also be determined by \*annotator\_knowledge\*.

1881 - If the knowledge source is \*annotator\_knowledge\*, then untrivialness could be 0-2.

1882

1883 Your output should be formatted as below.

1884

1885 ### Reasoning

1886 [think step-by-step]

1887

1888 ### Source

1889 [given\_paper or annotator\_knowledge or other\_papers]

1890

1891 ### Factuality

1892 [1 or 2 or 3 or 4 or 5]

1893

1894 ### Untrivialness

1895 [0 or 1 or 2]

Figure 20: Prompt used for evaluating ArgScore in a 5-point scale.

1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897  
1898  
1899  
1900  
1901  
1902  
1903  
1904  
1905  
1906  
1907  
1908  
1909  
1910  
1911  
1912  
1913  
1914  
1915  
1916  
1917  
1918  
1919  
1920  
1921  
1922  
1923  
1924  
1925  
1926  
1927  
1928  
1929  
1930  
1931  
1932  
1933  
1934  
1935  
1936  
1937  
1938  
1939  
1940  
1941  
1942  
1943

<b>Prompt for classifying review point types (<code>is_argument</code>)</b>
<pre># Review Point [[REVIEW_POINT]]  You are given an AI conference review point. Is this an argument or not? Your response should follow the format below.  ### Reasoning [think step-by-step]  ### Response [Yes or No]</pre>

Figure 21: Prompt used for deciding if a review point is an argument or not (`is_argument`).

<b>Prompt for classifying review point types (<code>is_question</code>)</b>
<pre># Review Point [[REVIEW_POINT]]  You are given an AI conference review point. Decide if this is a question or a simple claim. Your response should follow the format below.  ### Reasoning [think step-by-step]  ### Response [Question or Claim]</pre>

Figure 22: Prompt used for deciding if a review point is a question or a claim (`is_question`).

<b>Prompt for Argument Reconstruction (<code>extract_verbatim_conclusion_reason</code>)</b>
<pre>### Review Point [[REVIEW_POINT]]  Given an AI conference review point, consider it as an argument, and then returns its verbatim conjecture in the source text and verbatim reason statements of that conjecture in the source text. The output format should be as following.  ### Conjecture [main conjecture in the review point]  ### Supporting Reasons [list of supporting reasons for the conjecture]</pre>

Figure 23: Prompt used for extracting verbatim conjecture and reason statements in an argument (`extract_verbatim_conclusion_reason`).

1944  
1945  
1946  
1947  
1948  
1949  
1950  
1951  
1952  
1953  
1954  
1955  
1956  
1957  
1958  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1970  
1971  
1972  
1973  
1974  
1975  
1976  
1977  
1978  
1979  
1980  
1981  
1982  
1983  
1984  
1985  
1986  
1987  
1988  
1989  
1990  
1991  
1992  
1993  
1994  
1995  
1996  
1997

<b>Prompt for Argument Reconstruction (<code>argument_reconstruction</code>)</b>	
# Paper	[[PAPER]]
# Review Point	
## Conclusion	[[CONCLUSION]]
## Explicit reasons	[[REASONS]]
<p>You are given a paper submitted to an AI conference and a review point by a peer reviewer. A review point consists of a conclusion and its explicit reasons. Reconstruct an argument (i.e., a review) with premise-conclusion structure where premises deductively imply the conclusion. The reconstructed argument should be deductively valid, using formal logical patterns like modus ponens (e.g., Premise1: A, Premise2: If A then B, Conclusion: B). Add implicit premises and intermediate conclusions if needed.</p> <p>Your output should be composed of two parts, argument reconstruction and its formalization. In the first part, list premises, intermediate conclusions, and conclusion, and indicate their logical connection (i.e., which propositions deductively implies which). In the second part, first define variables and/or predicates, then formalize premises, intermediate conclusions, and a conclusion, and then generate a deductive proof. The output format should be as following.</p>	
# Argument Reconstruction	
## Premises	[list of explicit and implicit premises]
## Intermediate Conclusions	[list of intermediate conclusions (if intermediate conclusions are not needed, then write "None".)]
## Conclusion	[a conclusion]
## Logical Connections	[list of logical connections]
# Formalized Argument	
## Defined Variables/Predicates	[definition of each variable and/or predicate]
## Formalized Premises	[formalization of premises using definition]
## Formalized Intermediate Conclusions	[formalization of intermediate conclusions using definition (if intermediate conclusions are not needed, then write "None".)]
## Formalized Conclusion	[formalization of conclusion using definition]
## Deductive Proof	[deductive proof using formalized premises]

Figure 24: Prompt used for reconstructing an argument (`argument_reconstruction`).

1998  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051

<b>Prompt for Argument Reconstruction (streamlining)</b>	
1998	
1999	
2000	
2001	<b>Prompt for Argument Reconstruction (streamlining)</b>
2002	## Defined Variables/Predicates
2003	[[DEFINITION]]
2004	
2005	## Formalized Premises
2006	[[PREMISES]]
2007	
2008	## Formalized Conclusion
2009	[[CONCLUSION]]
2010	
2011	## Deductive Proof
2012	[[PROOF]]
2013	First, determine necessary formalized premises for the given deductive proof. This includes:
2014	1. Add any missing formalized premises that are necessary to prove conclusion but cannot be derived
2015	from the formalized premises.
2016	2. Remove any unnecessary formalized premises that are not necessary to prove conclusion but
2017	present in the formalized premises.
2018	You should format these premises into a python dictionary where keys and values are python strings.
2019	
2020	Second, write a python program using z3 that inputs the necessary formalized premises and formalized
2021	conclusion and outputs:
2022	1. Their validity, formatted as a python string of either “valid” or “invalid”.
2023	2. A smallest subset of necessary formalized premises to prove the formalized conclusion, formatted
2024	as a python list of keys of the python dictionary of the necessary formalized premises.
2025	You should therefore print two things (a python string and a python list) separately. Please use the be-
2026	low python code snippet.
2027	{Code snippet for checking validity}
2028	
2029	Third, return the final formalized conclusion that is used in the python program in step 2.
2030	
2031	Lastly, judge whether the formal proof using the necessary formalized premises (in step 1) and the final
2032	formalized conclusion (in step 3) is circular or not. If there is a single necessary formalized premise
2033	that is the same as the final formalized conclusion, then return N/A.
2034	
2035	Your response should be as following.
2036	
2037	### Necessary Formalized Premises
2038	``` python
2039	{
2040	“[Symbol of a premise #1]”: “[Formalization of a premise #1]”,
2041	“[Symbol of a premise #2]”: “[Formalization of a premise #2]”,
2042	...
2043	}
2044	```
2045	
2046	### Python Program
2047	``` python
2048	[a python program]
2049	```
2050	
2051	### Final Formalized Conclusion
2052	[Formalized conclusion in the python program]
2053	
2054	### Proof Circularity
2055	[Yes or No or N/A]

Figure 25: Prompt used for streamlining formalized reconstruction (streamlining).

2052  
2053  
2054  
2055  
2056  
2057  
2058  
2059  
2060  
2061  
2062  
2063  
2064  
2065  
2066  
2067  
2068  
2069  
2070  
2071  
2072  
2073  
2074  
2075  
2076  
2077  
2078  
2079  
2080  
2081  
2082  
2083  
2084  
2085  
2086  
2087  
2088  
2089  
2090  
2091  
2092  
2093  
2094  
2095  
2096  
2097  
2098  
2099  
2100  
2101  
2102  
2103  
2104  
2105

<b>Code snippet for checking validity</b>	
<pre> 2055 ``` python 2056 from z3 import * 2057 import itertools 2058 2059 ##### 2060 ### Write down your code here ### 2061 ##### 2062 2063 # Check validity of the argument 2064 def check_validity(premises_dict, conclusion): 2065     s = Solver() 2066     s.add(list(premises_dict.values())) 2067     s.add(Not(conclusion)) 2068     if s.check() == unsat: 2069         return "valid" 2070     else: 2071         return "invalid" 2072 2073 # Find minimal set of premises 2074 def find_minimal_premises(premises_dict, conclusion): 2075     for subset_size in range(1, len(premises_dict) + 1): 2076         for subset in itertools.combinations(premises_dict.keys(), subset_size): 2077             subset_premises = [premise_dict[key] for key in subset] 2078             if check_validity(subset_premises, conclusion) == "valid": 2079                 return list(subset) 2080     return list(premises_dict.keys()) 2081 2082 validity = check_validity(premises, conclusion) 2083 print(validity) 2084 minimal_premises = find_minimal_premises(premises, conclusion) 2085 print(minimal_premises) 2086 ``` </pre>	

Figure 26: Python code snippet for evaluating validity of reconstruction.

<b>Prompt for Argument Reconstruction (program_debugging)</b>	
<pre> 2091 When I execute the python program, I got the following error: 2092 [[ERROR]] 2093 2094 Fix the error and generate a revised python program. Your response should be as following. 2095 2096 ### Reasoning 2097 [explain why and how to fix the program] 2098 2099 ### Revised Python Program 2100 ``` python 2101 [a python program] 2102 ``` </pre>	

Figure 27: Prompt used for debugging python programs that evaluate validity of reconstruction (program\_debugging).

2106  
2107  
2108  
2109  
2110  
2111  
2112  
2113  
2114  
2115  
2116  
2117  
2118  
2119  
2120  
2121  
2122  
2123  
2124  
2125  
2126  
2127  
2128  
2129  
2130  
2131  
2132  
2133  
2134  
2135  
2136  
2137  
2138  
2139  
2140  
2141  
2142  
2143  
2144  
2145  
2146  
2147  
2148  
2149  
2150  
2151  
2152  
2153  
2154  
2155  
2156  
2157  
2158  
2159

<b>Prompt for Argument Reconstruction (deformalization)</b>	
## Defined Variables/Predicates	[[DEFINITION]]
## Formalized Premises	[[PREMISES]]
## Formalized Conclusion	[[CONCLUSION]]
Given definitions of variables and/or predicates, generate natural language (NL) descriptions of formalized premises and conclusion. Your response should be as following.	
### NL Premises	[list of premises in natural language]
### NL Conclusion	[conclusion in natural language]

Figure 28: Prompt used for translating FOL formulas with keys (i.e., defined variables/predicates) to NL reconstructed arguments (deformalization).

<b>Prompt for Argument Reconstruction (check_faithfulness)</b>	
# Argument	[[ARGUMENT]]
# Argument Reconstruction	
## Premises	[[PREMISES]]
## Conclusion	[[CONCLUSION]]
For an argument, its reconstruction as a premise-conclusion structure is given. Your task is to judge whether the construction is faithful or not. You should judge the faithfulness according to the following two criteria:	
- <b>Accuracy &amp; Charity.</b> The reconstruction should keep the author’s intended meaning while eliminating irrelevancies—i.e., obey the principle of charity and prefer the strongest sensible reading of ambiguous passages.	
- <b>Completeness.</b> All explicit premises, the main conclusion and any indispensable implicit premises must be included.	
The output format should be as following.	
# Reasoning	[Explain step-by-step]
# Faithfulness	[Yes or No]

Figure 29: Prompt used for evaluating faithfulness of reconstruction (check\_faithfulness).

## I ADDITIONAL COMPARISON WITH PRIOR WORKS

### I.1 DATASET STATISTICS

We compare dataset statistics of REVIEWSCORE with other datasets in Table 14. “-” indicates the information is not explicitly addressed in the paper. We indicate the statistics of human-annotated reviews. In terms of number of papers, our work has a similar size with Ou et al. (2025) and has a smaller size than Shin et al. (2025) and Du et al. (2024). However, in terms of number of review points, which is the primary measure of review data size, our work has a larger data size than Ou et al. (2025), Shin et al. (2025), and Purkayastha et al. (2025) since they filter out reviews that does not fit to their problem setting (We use all review points as targets for evaluation). Furthermore, while other works focus on evaluating at a level of review points, we also evaluate at a level of premises of review points, which are more fine-grained. As a result, our human experts annotate total 1,748 premises of review points. To guarantee annotation reliability, each instance is labeled by three human experts. This design choice intentionally prioritizes quality over scale, resulting in a smaller dataset under limited human annotation capacity. Furthermore, other works mostly evaluate reviews of NLP papers, whereas our work includes other domains such as CV and Time Series.

	# Papers	# RP	# Prem.	Ann./Inst.	Domains	Venues
Ou et al. (2025)	41	168	N/A	–	NLP	NeurIPS 2023–2024
Shin et al. (2025)	68	327	N/A	2	–	ICLR 2021–2024
Purkayastha et al. (2025)	–	500	N/A	2	NLP	ARR 2022
Du et al. (2024)	100	N/A	N/A	2	NLP	ICLR/NeurIPS 2020–2023
REVIEWSCORE (Ours)	40	<b>657</b>	<b>1,748</b>	3	CV, NLP, Time Series	ICLR 2021–2023

Table 14: Comparison of dataset statistics. RP = Review Points. Prem. = Premises.

### I.2 REVIEW EVALUATION METHODS

We compare review evaluation methods of REVIEWSCORE with other works in Table 15. To be specific, we compare evaluation targets and units, and review quality criteria. These four criteria are adopted from Goldberg et al. (2025) and other review evaluation works. While most of previous works focus on evaluating weaknesses, we additionally focus on evaluating questions. As a result, we are the only work that consider not only factuality of weaknesses but also unanswerability of questions. Most of the works including REVIEWSCORE evaluate review points, and we additionally evaluate premises of review points, which we define factuality at a level of premises. In contrast, Du et al. (2024) evaluates factuality of each sentence. However, sentence-level factuality is less reliable than premise-level factuality since human reviews could be repetitive, hide implicit premises, contain rhetoric, and sometimes change a stance in a single review. Premise-level factuality is robust to all of these effects. Guo et al. (2023), Sadallah et al. (2025), and Ou et al. (2025) consider groundedness of reviews to the paper, but do not consider factuality and leave it as a future work. Our work and Du et al. (2024) both considers groundedness and factuality. However, since we are focusing of detecting misinformation in reviews, we do not evaluate constructiveness of reviews. We also think that evaluating constructiveness is more subjective and value-laden than other criteria, which requires more careful human annotations.

	Targets	Units	Groundedness	Factuality	(Un)Answerability	Constructiveness
Guo et al. (2023)	S, W	R	✓	×	×	×
Sadallah et al. (2025)	W	RP	✓	×	×	✓
Ou et al. (2025)	W	RP	✓	×	×	×
Shin et al. (2025)	S, W	RP	×	×	×	×
Purkayastha et al. (2025)	W	RP	×	×	×	×
Du et al. (2024)	W	RS	✓	✓	×	✓
REVIEWSCORE (Ours)	W, Q	RP & Prem.	✓	✓	✓	×

Table 15: Comparison of review evaluation methods. S = Strengths. W = Weaknesses. Q = Questions. R = Reviews, RP = Review Points. RS = Review Sentences. Prem. = Premises.