

# NovBench: Evaluating Large Language Models on Academic Paper Novelty Assessment

Wenqing Wu<sup>1,2</sup>, Yi Zhao<sup>3</sup>, Yuzhuo Wang<sup>3</sup>, Siyou Li<sup>2</sup>,  
Juexi Shao<sup>2</sup>, Yunfei Long<sup>2</sup><sup>†</sup>, Chengzhi Zhang<sup>1</sup><sup>†</sup>

<sup>1</sup>School of Economics and Management, Nanjing University of Science and Technology,  
<sup>2</sup>School of Electronic Engineering and Computer Science, Queen Mary University of London,  
<sup>3</sup>School of Management, Anhui University

<sup>†</sup>Correspondence: zhangcz@njust.edu.cn, yunfei.long@qmul.ac.uk

## Abstract

Novelty is a core requirement in academic publishing and a central focus of peer review, yet the growing volume of submissions has placed increasing pressure on human reviewers. While large language models (LLMs), including those fine-tuned on peer review data, have shown promising results in generating review comments, the absence of dedicated benchmarks has limited systematic evaluation of their ability to assess research novelty. To address this gap, we introduce NovBench, the first large-scale benchmark designed to evaluate LLMs' capability to generate novelty evaluations in support of human peer review. NovBench comprises 1,684 paper-review pairs from a leading NLP conference, including novelty descriptions extracted from paper introductions and corresponding expert-written novelty evaluations. We focus on both sources because the introduction provides a standardized and explicit articulation of novelty claims, while expert-written novelty evaluations constitute one of the current gold standards of human judgment. Furthermore, we propose a four-dimensional evaluation framework (including Relevance, Correctness, Coverage, and Clarity) to assess the quality of LLM-generated novelty evaluations. Extensive experiments on both general and specialized LLMs under different prompting strategies reveal that current models exhibit limited understanding of scientific novelty, and that fine-tuned models often suffer from instruction-following deficiencies. Our findings underscore the need for targeted fine-tuning strategies that jointly improve novelty comprehension and instruction adherence.

## 1 Introduction

Novelty is a fundamental aspect of publication decisions in academic research, requiring that a paper's content or methodology makes a meaningful contribution to advancing existing knowledge, rather than simply replicating or validating established findings (Veugelers and Wang, 2019). As the primary

quality control mechanism in scientific research, assessing novelty is one of the core functions of the peer review system (Alberts et al., 2008; Cohen, 2017). However, the peer review process is currently facing pressure due to the explosive growth in academic submissions (Huang et al., 2025b) and the widening gap in the availability of qualified reviewer resources (Adam, 2025). This pressure directly manifests as a significant challenge for the robust evaluation of a paper's novelty (Zhao and Zhang, 2025). Although methods for novelty evaluation have been proposed based on bibliometric (Uzzi et al., 2013; Matsumoto et al., 2021; Shibayama et al., 2025) and deep learning (Huang et al., 2025a; Jeon et al., 2023; Liu et al., 2025a) approaches, these techniques predominantly focus on quantitative metrics (or numerical indicators). Compared to the textual evaluations provided by reviewers, these quantitative metrics inherently lack the necessary interpretability, making it difficult to effectively assist reviewers during the peer review process or to provide authors with targeted advice for manuscript improvement.

Large Language Models (LLMs) have demonstrated exceptional capabilities across a wide range of scientific tasks (Wang et al., 2024; Bao et al., 2025) and show potential in assisting academic peer review (Zhou et al., 2024; Zhuang et al., 2025). Recent researches have attempted to enhance the capabilities of LLMs in automated peer review through various avenues, such as applying more effective prompting strategies (Liang et al., 2024), fine-tuned model (Yu et al., 2024a; Zhu et al., 2025a), and multi-agent frameworks (Jin et al., 2024; D'Arcy et al., 2024). Despite the promising progress demonstrated by recent advances, several critical issues remain insufficiently explored. Although these methods achieve competitive performance on paper-level scoring tasks, the evaluation of generated review text itself has received limited attention. Existing evaluation approaches

are largely limited to metrics such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and BERTScore (Zhang\* et al., 2020), or the adoption of the LLM-as-judge (Bedemariam et al., 2025). However, these methods either rely on surface-level lexical similarity or non-transparent LLM-based judgments, and therefore fail to reliably assess the semantic adequacy and aspect-specific correctness required for evaluating free-form novelty evaluations (Kuznetsov et al., 2024). Furthermore, existing research primarily focuses on generating holistic peer review text, treating the review as a monolithic output. As a result, the ability of both general purpose and fine tuned models to perform aspect specific evaluation, particularly novelty evaluation, remains poorly characterized. Without isolating novelty as an independent evaluation target, it is difficult to determine whether a model genuinely evaluates research novelty or merely generates fluent, plausible sounding review like language. Therefore, the performance of existing general and fine-tuned models on the novelty evaluation task constitutes an important area for further investigation.

To address these issues, we present a novel evaluation framework, to be more specific: (1) We introducing a structured benchmark for novelty evaluation. This resource, which incorporates the textual evaluations of the novelty dimension of the reviewers alongside the novelty descriptions of the paper introductions, constitutes a critical resource for future research. (2) We propose a four-dimensional, interpretable, and semantics-aware framework for evaluating LLM-generated novelty review text. This metric suite surpasses the limitations of traditional lexical overlap metrics (e.g., ROUGE/BLEU) and more effectively captures the quality of "novelty" evaluation in free-form text. (3) We conduct systematic evaluation and benchmarking of current general LLMs (e.g., GPT-5 (OpenAI, 2025), Gemini-2.5-flash (Gemini Team, 2025)) and specialized LLMs on the novelty evaluation task. Based on this benchmark, we further analyze the gap between LLM-generated evaluations and human judgments, providing an in-depth understanding of these models' advantages and limitations in identifying and articulating novelty. (4) We conduct a comprehensive empirical analysis that identifies key factors for generating high-quality, highly interpretable novelty review text, and reveals important behavioral patterns of LLMs in novelty evaluation, thus directing the future development of more

reliable AI-assisted peer review.<sup>1</sup>

## 2 Related Work

Automated Scholarly Paper Review (ASPR) (Lin et al., 2023) refers to the process in which computers or intelligent machines independently evaluate the content of a scholarly paper and generate a review report automatically. Early research concerning ASPR predominantly focused on paper rating recommendation (Kang et al., 2018; Li et al., 2020; Wang et al., 2020). Furthermore, some studies (Yuan et al., 2022; Yuan and Liu, 2022) have attempted to fine-tune pre-trained models to generate paper reviews.

With the LLMs demonstrate powerful text generation capabilities (Yu et al., 2025; Chen et al., 2025), the application of LLMs to help peer review has rapidly become a significant research focus. Numerous studies have evaluated or benchmarked the performance of LLMs in generating reviews of academic papers (Zhou et al., 2024; Liang et al., 2024; Du et al., 2024). The results of these investigations indicate that while LLMs are capable of providing meaningful feedback, they often lack critical analysis, and the comments generated frequently lack the insights and specificity found in human-written reviews.

Consequently, the research community has devoted significant efforts for enhancing LLM performance in review generation (Yu et al., 2024b; Gao et al., 2024; Idahl and Ahmadi, 2025; Weng et al., 2025; Zhu et al., 2025b; Chang et al., 2025). However, recent advancements have broadly evaluated and improved the overall performance of LLMs in review generation. Despite this progress, their efficacy in assessing specific, fine-grained aspects of a paper (especially novelty) remains an area requiring further investigation.

Novelty is recognized as a key aspect for measuring the quality and contribution of academic papers. Currently, as LLM continue to demonstrate increased capability, researchers have advanced their investigation of the role of LLMs in scientific novelty, including the evaluation based on LLM of paper novelty (Huang et al., 2025a; Liu et al., 2025a; Lin et al., 2025; Liu et al., 2025b; Wu et al., 2025a,b; Tan et al., 2026) and the generation of novel research ideas (Shahid et al., 2025; Su et al., 2025; Kumar et al., 2025; da Silva et al., 2025).

---

<sup>1</sup>All resources publicly available at <https://github.com/njust-winchy/llm4novelty>.

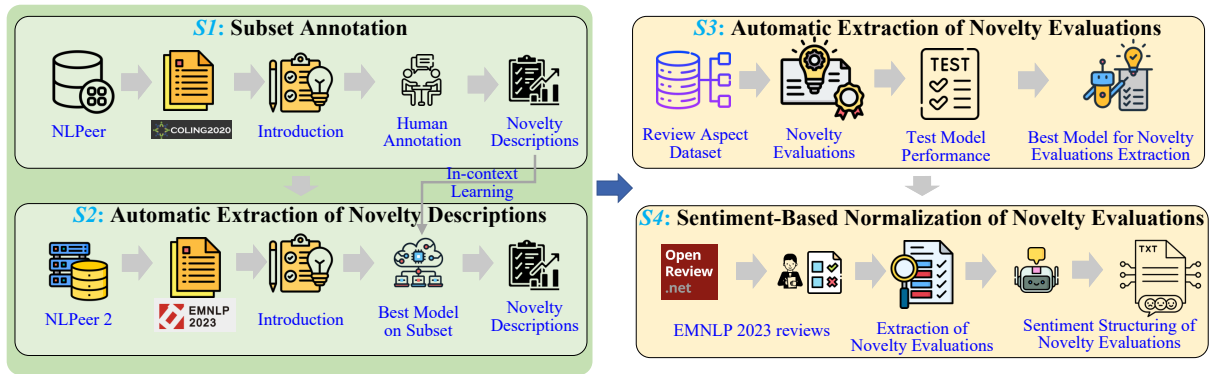


Figure 1: The pipeline for constructing NovBench, consisting of four stages.

Although these studies suggest that LLMs possess a certain capacity for novelty assessment, the evaluation of LLM performance is largely centered on quantitative output scores or relies on human evaluation. There remains a notable absence of dedicated assessment targeting the LLM generated textual novelty evaluations themselves, with current practices frequently defaulting to the "LLM as a judge" paradigm. This reliance exposes a gap in the development of robust evaluation methodologies for free form assessment text. While prior work has introduced resources for analyzing research limitations (Xu et al., 2025), there is currently no publicly available resource dedicated to novelty assessment. To address this gap, we introduce a dedicated resource for research novelty and a novel methodology for assessing novelty evaluation. This paper deliberately focuses on novelty, as it is widely regarded as a central criterion for publication decisions and one of the most conceptually challenging aspects to evaluate (Afzal et al., 2026).

### 3 NovBench

#### 3.1 Dataset Construction

We initially collected the COLING 2020 and EMNLP 2023 datasets (include parsed paper in intertextual graphs (ITG) format (Kuznetsov et al., 2022) and review report) from NLPeer (Dycke et al., 2023) and OpenReview<sup>2</sup>, followed by the execution of three automated stages: (1) extraction of novelty descriptions from the paper introductions, (2) extraction of novelty-related evaluations from the reviewer comments, and (3) structuring novelty evaluations based on sentiment polarity. We extract novelty descriptions from paper introductions because they provide the most explicit and

<sup>2</sup><https://openreview.net/>

standardized articulation of authors' claimed contributions (Lauscher et al., 2018; Li et al., 2021). Figure 1 shows the construction workflow.

#### 3.1.1 Automatic Extraction of Novelty Descriptions from Introduction

We first extracted all content under the heading "Introduction" from the parsed paper. Subsequently, we adopted a strategy of annotating a small subset (COLING 2020, 87 papers) of the data to evaluate the performance of LLMs, which was then applied to the large dataset (EMNLP 2023, 1,684 papers). Specifically, we manually annotated the novelty descriptions extracted from the introductions of the COLING 2020 papers. We then designed various prompts to evaluate the performance of mainstream LLMs on this specific task, the detailed results are presented in the Appendix A. Finally, we selected GPT-5 with in-context-learning prompt as the method for the automatic extraction of novelty descriptions for the introduction of EMNLP 2023.

#### 3.1.2 Automatic Extraction of Novelty Evaluations from Peer Review Texts

To extract novelty related evaluations from reviewer comments, it is necessary to perform aspect identification on the review texts. Recently, Lu et al. (Lu et al., 2025) introduced a resource for aspect identification in peer review text. We obtained all data corresponding to the novelty aspect of reviews from this resource. We then evaluated both prior aspect identification models and LLMs on this data, with the specific results reported in Appendix B. Based on these results, we selected GPT-4o-mini as the model for automatic extraction of novelty evaluations. Using this model, we extracted all content related to novelty evaluation from reviewer reports of EMNLP 2023.

### 3.1.3 Sentiment-Based Structuring of Novelty Evaluations

During the review process, multiple reviewers may independently praise a paper for proposing a novel approach. To facilitate the evaluation of novelty evaluations generated by the LLM, we designed a prompt that instructs GPT-4o to remove redundant evaluations and organize novelty evaluations according to sentiment polarity. The specific details are provided in Appendix C.

Dataset	#Paper	Avg. Nov Desc	Avg. Nov Eval
COLING 2020	87	6.1	-
NovBench	1,684	5.3	7.7

Table 1: Dataset statistics of Novbench and COLING 2020 (subset in Figure 1), including paper count, average novelty description (Avg. Nov Desc) sentence count, and average novelty evaluations (Avg. Nov Eval) count. '-' indicates not applicable.

Paper Novelty Description:	Structured Novelty Evaluations:
<p>To answer the question, we first define an evaluation suite of physical concepts covering visual and embodied concepts.</p> <p>We compose a Visual and Embodied Concepts evaluation benchmark VEC, with examples shown in Table 1.</p> <p>With the benchmark, we examine a wide range of LMs. We design two forms of probing tasks.</p> <p>The former (Color, Shape and Material) asks models to make a choice between two tail options given the head object.</p> <p>The latter (Size, Height, and all embodied concepts) requires LMs to judge whether the relation is valid given the head and the tail.</p> <p>Our zero (few)-shot results on the VEC benchmark show that: (i) Moderate-sized LMs such as BERT and RoBERTa exhibit a random-level understanding of both visual and embodied concepts.</p> <p>(ii) A decent visual understanding of specific concepts emerges as LMs scale up, while they still struggle to understand the embodied knowledge with performance slightly better than random guessing.<sup>2</sup></p> <p>(iii) Image-grounded caption text-only pre-training, instruction tuning, and visual supervision could provide performance gain regarding visual concepts, yet only the last one enhances the understanding of embodied knowledge of LMs.</p> <p>We further investigate the source of embodied knowledge in VLMs.</p> <p>A case study demonstrates that embodied knowledge in the VLM of CLIP is potentially rooted in the rich semantics of image representations.</p> <p>We thus devise a knowledge distillation method to transfer the learned embodied knowledge in VLMs into LMs, resulting in an average accuracy gain of 3.38.</p>	<p><b>Positive Novelty Evaluations</b></p> <ul style="list-style-type: none"> <li>- The paper introduces a novel benchmark evaluation suite (VEC) that covers a wide range of physical concepts, including both visual and embodied attributes, filling a gap in the field by providing a standardized framework to evaluate language models' understanding of physical attributes.</li> <li>- The finding that we can effectively distill embodied knowledge from VLMs to less capable LLMs is quite inspiring.</li> <li>- Probing LMs with physical concepts is a novel and interesting topic that provides insights on how to further improve and align current LMs to the real world.</li> <li>- The paper is well-written and easy to follow.</li> </ul> <p><b>Neutral Novelty Evaluations</b></p> <ul style="list-style-type: none"> <li>- The main contributions of this paper include conducting in-depth analysis with different types of LLMs and VLMs on their understanding of visual concepts and embodied concepts, providing insights on what physical concepts are well-captured and struggled by LLMs, and showing evidence that vision-language pretraining helps learn embodied knowledge.</li> </ul> <p><b>Negative Novelty Evaluations</b></p> <ul style="list-style-type: none"> <li>- Compared to GPT-3 era models, newer models like ChatGPT contain much richer physical commonsense knowledge and can perform complex reasoning about physical interactions, suggesting that the paper's findings may be less novel in the context of recent advancements.</li> </ul>

Figure 2: One Example of NovBench. The left side shows the novelty descriptions from the paper introductions, while the right side presents the structured novelty evaluations.

### 3.1.4 Dataset Statistics

Table 1 summarizes the datasets used in this study. The COLING 2020 dataset corresponds to a manually annotated subset of NovBench. This subset includes only novelty descriptions from paper introductions and is primarily used to evaluate the performance of automatic novelty description extraction models, rather than for large-scale benchmarking. Such a design supports controlled

model selection for the novelty description extraction stage. NovBench constitutes the full benchmark constructed in this work. For each paper, we automatically annotate two sources of novelty-related information: (1) novelty descriptions extracted from the paper introduction, and (2) novelty evaluations written by human reviewers. This dual-source design captures both author-stated novelty claims and independent human evaluations, enabling a systematic evaluation of LLM-generated novelty evaluations against human judgments. An example of the dataset is illustrated in Figure 2.

### 3.2 Dataset Evaluation Protocol

In this work, we define the task as follows: given the novelty description from an academic paper's introduction, the LLM is required to generate an assessment of the novelty according to specific instructions. The output format mandates a structured output based on sentiment polarity.

Academic peer review is typically conducted along multiple quality dimensions rather than a single overall score. Reviewer guidelines<sup>3</sup> for major NLP venues explicitly encourage assessments of a paper's novelty, correctness or soundness, relevance or impact, and clarity of presentation. Such multi-dimensional evaluation practices motivate our design of four dimensions for assessing LLM-generated novelty evaluations. We fix the evaluation rubric to ensure a controlled and comparable setting, allowing us to isolate the novelty evaluation capability of LLMs.

**Relevance.** This dimension is defined as the degree to which the LLM-generated evaluation accurately comprehends the novelty description presented in the paper's introduction. To quantify this, we calculate the Information Matching Score (IMS) between the model-generated evaluation and the novelty description in the introduction, utilizing a method (Wright et al., 2022) designed to measure the information alignment between two scientific sentences. Our objective is to determine if B (LLM-generated evaluation) truly understands the content of A (the source text). Therefore, we adopt the *Maximum Matching Average IMS* (AvgIMS), which measures how well each sentence in the source text is semantically covered by the most relevant sentence in the evaluation. For each sentence in A, we identify the review sentence that yields the highest IMS value, and compute the average of these

<sup>3</sup><https://aclrollingreview.org/reviewerguidelines>

maxima across all sentences:

$$\text{AvgIMS} = \frac{1}{N} \sum_{i=1}^N \max_j S_{ij} \quad (1)$$

where  $S_{ij}$  denotes the IMS between the  $i$ -th sentence in the source text and the  $j$ -th sentence in the LLM-generated evaluation. A higher AvgIMS indicates that the evaluation closely aligns with and accurately reflects the novelty description.

**Correctness.** This dimension assesses the agreement between the model-generated positive, neutral, or negative novelty evaluations and human reviewers (Zhang et al., 2022). To achieve this, we compare the resulting sentiment distribution produced by the LLM against the distribution established by the human reviewers. Following prior distribution-matching metrics, we define correctness as:

$$\text{DistAcc} = 1 - \frac{\sum |p_i - t_i|}{2} \quad (2)$$

where  $p_i$  denotes the proportion of model-generated evaluations labeled with sentiment class  $i$ , and  $t_i$  represents the corresponding human evaluations proportion. The numerator computes the  $L_1$  distance between the two distributions, and the division by 2 normalizes the maximum possible distance to 1. A higher DistAcc value indicates better alignment between model generation and human sentiment judgments.

**Coverage.** To evaluate whether the LLM-generated novelty evaluations adequately capture the key points identified by human reviewers, we define a *Coverage* dimension. Let  $G$  denote the set of review novelty evaluations, and  $M$  denote the set of LLM-generated evaluations. For each review novelty evaluation  $g \in G$ , we compute the cosine similarity with all LLM-generated evaluation  $m \in M$ , and count  $g$  as covered if the maximum similarity exceeds a threshold  $\tau$ . Formally, Coverage is defined as:

$$\text{Coverage} = \frac{1}{|G|} \sum_{g \in G} \mathbf{1} \left[ \max_{m \in M} \cos(g, m) \geq \tau \right], \quad (3)$$

where  $\cos(g, m)$  denotes the cosine similarity between the embeddings of  $g$  and  $m$ , and  $\mathbf{1}[\cdot]$  is the indicator function. In our experiments, we set  $\tau = 0.7$ , following common practice in prior work on sentence-level semantic similarity using sentence embeddings (Reimers and Gurevych, 2019).

A higher Coverage score indicates that more of the expert-identified novelty points are captured by the LLM-generated evaluations.

**Clarity.** The goal of this dimension is to determine if the generated review text is explicit and focused, ensuring the evaluation is easily understood and not overly generalized. We measure the clarity of LLM-generated novelty evaluations using a combined metric that accounts for both lexical grounding and sentence elaboration (Louis and Nenkova, 2013). Let  $K$  denote the set of keywords extracted from the introduction, and let  $T$  denote the set of LLM-generated evaluation sentences.

The first component, *Keyword Coverage* (KC), assesses whether each generated sentence contains at least one introduction keyword, indicating lexical grounding in the source text:

$$\text{KC} = \frac{1}{|T|} \sum_{t \in T} \mathbf{1} \left[ \exists k \in K \text{ such that } k \subset t \right] \quad (4)$$

In our implementation, keywords are automatically derived from the introduction novelty sentences by extracting alphanumeric and hyphenated tokens using regular expressions, followed by filtering tokens with length greater than 5 to remove function words and generic short tokens. A generated evaluation is considered keyword-covered if it contains at least one such keyword (case-insensitive), and matching is performed at the token level to avoid spurious substring matches. The second component, *Length Score* (LS), encourages sufficiently informative sentences without enforcing verbosity by computing the average sentence length (in tokens), normalized by 20 and clipped to a maximum of 1:

$$\text{LS} = \min \left( \frac{1}{20|T|} \sum_{t \in T} \text{len}(t), 1 \right) \quad (5)$$

where  $\text{len}(t)$  denotes the token length of sentence  $t$ . Then, to capture linguistic well-formedness and readability, we incorporate a fluency-based component derived from language model perplexity. Let  $PPL(t)$  denote the perplexity of sentence  $t$  computed using a pretrained causal language model (distilgpt2 (Sanh et al., 2020)). We define:

$$\text{FS} = \frac{1}{|T|} \sum_{t \in T} \frac{1}{1 + PPL(t)} \quad (6)$$

Lower perplexity corresponds to higher fluency and better grammatical quality. The inverse transformation ensures that the score lies within (0, 1).

Model	Zero Shot				Few Shot				RAG			
	Rel.	Cov.	Clarity	DistAcc	Rel.	Cov.	Clarity	DistAcc	Rel.	Cov.	Clarity	DistAcc
<b>General LLMs</b>												
DeepSeek-R1-70B	3.4885	0.2074	0.6470	0.6572	3.4452	0.2112	0.6455	0.6274	3.0376	0.1500	0.6626	0.6260
DeepSeek-R1-14B	3.4058	0.2053	0.6404	0.6436	3.3844	0.2252	0.6483	0.6606	3.0715	0.1713	0.6628	0.6417
DeepSeek-R1-8B	2.9475	0.1603	0.4949	0.5143	3.4703	0.2190	0.6160	0.6517	2.4988	0.1284	0.5459	0.5220
Qwen3-32B	3.4747	0.2065	0.6609	0.6555	3.4175	0.2242	0.6497	0.6944	3.2555	0.1692	0.6679	0.6604
Qwen3-14B	3.3892	0.1974	<b>0.6634</b>	0.6627	3.3656	0.2238	0.6487	0.6734	3.2015	0.1673	0.6681	0.6245
Qwen3-8B	3.5769	0.1996	0.6471	0.6595	3.4061	0.2240	0.6487	0.6784	3.3042	0.1732	0.6593	0.6737
GPT-4o	<b>3.6983</b>	<b>0.2332</b>	0.6595	<b>0.6979</b>	<b>3.5609</b>	<b>0.2391</b>	0.6587	<b>0.7091</b>	3.4481	0.2237	0.6668	<b>0.6965</b>
GPT-5	3.2772	0.1591	0.6209	0.4830	3.3124	0.1806	0.6164	0.5411	3.2300	0.1673	0.6666	0.6453
gpt-oss-120b	3.2586	0.1787	0.6535	0.4376	3.1897	0.1830	<b>0.6613</b>	0.5107	3.1424	0.1644	0.6656	0.6027
gpt-oss-20b	3.3158	0.1843	0.6270	0.5785	3.3298	0.2098	0.6364	0.6461	3.1372	0.1676	0.6269	0.6430
Gemini-2.5-flash	3.4711	0.2120	0.6414	0.6011	3.4726	0.2364	0.6573	0.6590	<b>3.5089</b>	<b>0.2270</b>	<b>0.6682</b>	0.5923
<b>Specialized LLMs</b>												
CycleReviewer-70B	3.4632	0.2198	0.6598	0.6326	3.3426	0.2209	0.6494	0.6522	3.0292	0.1522	0.6587	0.4892
CycleReviewer-8B	3.0712	0.1577	0.6377	0.3837	3.0833	0.1892	0.6206	0.4088	2.8853	0.1336	0.5817	0.2785
DeepReviewer-14B	2.7402	0.1173	0.6425	0.6356	2.7161	0.1140	0.6130	0.6301	2.7288	0.1193	0.6161	0.5246
DeepReviewer-7B	2.4654	0.0745	0.5992	0.6134	2.6589	0.1051	0.6160	0.6308	2.5465	0.1004	0.5697	0.3907
Llama-OpenReviewer-8B	2.1293	0.0604	0.3339	0.2216	2.3047	0.0823	0.5381	0.6205	1.2235	0.0317	0.2749	0.1031
Reviewer2	1.8377	0.0408	0.4496	0.4317	0.9993	0.0013	0.3402	0.4083	0.1556	0.0000	0.0184	0.0709
SEA-E	3.4259	<u>0.2610</u>	0.6497	0.6834	3.3356	<u>0.2483</u>	0.6395	0.6609	3.3807	<u>0.2712</u>	0.6585	0.5965
SEA-S	<u>3.6304</u>	<u>0.2576</u>	<u>0.6630</u>	<u>0.7162</u>	3.4091	0.2454	0.6519	0.7149	3.5170	<u>0.2474</u>	<u>0.6662</u>	<u>0.6740</u>
Human	2.7899	-	-	-	2.7899	-	-	-	2.7899	-	-	-

Table 2: The evaluation performance of different models under various prompting strategies. For each metric, the best-performing general model is highlighted in bold, and the best-performing specialized model is underlined. Rel. denotes Relevance, Cov. denotes Coverage, and DistAcc denotes Correctness. Human refers to the performance of human reviewers on our evaluation metrics. '-' indicates not applicable.

The final *Clarity Score* is defined as the mean of these two components:

$$Clarity = \frac{1}{3}(KC + LS + FS) \quad (7)$$

A higher score indicates that the model generates sentences that are both lexically grounded and sufficiently elaborated.

Following the ACL/EMNLP review scoring system, Relevance is scored on a scale from 1 to 5, while the other dimensions are scored from 0 to 1.

## 4 Experiments

### 4.1 Baselines Selection

Based on our proposed evaluation metrics, we assessed a total of 11 general-purpose LLMs across two categories: (1) **Closed-source LLMs**: GPT-4o (OpenAI et al., 2024), GPT-5 (OpenAI, 2025), and Gemini-2.5-flash (Gemini Team, 2025); (2) **Open-source LLMs**: DeepSeek-R1 (70B, 14B, 8B) (DeepSeek-AI et al., 2025), Qwen3 (32B, 14B, 8B) (Yang et al., 2025), and gpt-oss (120B, 20B) (OpenAI et al., 2025). Furthermore, we also evaluated eight domain specialized LLMs that were **fine-tuned on peer review dataset**: CycleReviewer-70B, CycleReviewer-8B (Weng et al., 2025), DeepReviewer-14B, DeepReviewer-7B (Zhu et al., 2025b), Llama-OpenReviewer-8B (Idahl and Ahmadi, 2025), Reviewer2 (Gao et al., 2024), SEA-E and SEA-S (Yu et al., 2024b). We

access closed-source models via their official APIs, while the open-source models were downloaded locally from HuggingFace<sup>4</sup> for inference. During testing on NovBench, we used greedy decoding with a maximum token limit of 4096 to guarantee output determinism and prevent truncation. We retained the default values for all other hyperparameters. We adopt three prompting strategies: zero-shot, few-shot, and Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). Implementation details are shown in Appendix E.

### 4.2 Overall Performance of the Baseline Model with Automatic Metrics

Table 2 reports model performance in different evaluation metrics and prompting strategies. From the results, we observe that across prompting settings, closed-source general LLMs (like GPT-4o and Gemini-2.5-Flash) have stronger performance, likely due to their larger parameter scales and undisclosed model architectures. When comparing models with comparable parameter sizes, specialized LLMs generally outperform general models, this advantage mainly depends on the choice of backbone and the fine-tuning strategy. For instance, SEA-S and SEA-E are built on Mistral (a mix-of-experts) backbone, which provides an inherent advantage for expert-level tasks such as novelty evaluation. Nevertheless, even with the same back-

<sup>4</sup><https://huggingface.co/>

bone, performance differences remain, driven by variations in fine-tuning approaches, as illustrated by CycleReviewer-8B and SEA-S. Overall performance tends to improve with increasing model size, though notable exceptions are observed for general-purpose LLMs. This may be because larger models’ stronger reasoning and generative abilities can induce over-interpretation and distributional drift under strict evaluation constraints. Additionally, from the results in the table, we observe that Human achieves relatively lower scores on the Relevance metric. This is because human reviewers typically rely on their domain knowledge and experience to make high-level judgments, rather than explicitly restating or strictly aligning their comments with the novelty descriptions in the paper introduction.

### 4.3 Human Agreement with Automatic Metrics

To validate the effectiveness of our proposed metrics in assessing the generation of novelty evaluations, we randomly selected 100 samples for human evaluation. Specifically, we established a controlled comparison wherein evaluators were tasked with judging which model (Model A or Model B) produced the higher-quality novelty evaluation. This judgment was performed by providing the evaluators with the novelty description from the paper’s introduction and the human reviewer’s evaluation. The detailed examples and evaluation guidelines (Figure 15) are provided in the Appendix D. Four human evaluators with strong expertise in Natural Language Processing, including two Ph.D. students, one Associate Professor, and one Lecturer, independently conducted the evaluation following the same guidelines. The inter-annotator agreement, measured by Fleiss’  $\kappa$ , reached 0.72, indicating substantial agreement. Our proposed automatic metrics demonstrated a high correlation with the corresponding human judgments (Spearman  $\rho$  is 0.61, with  $p < 0.001$ ). This result confirms that our metrics are capable of correctly identifying superior model-generated evaluation, consistent with human preference (Agreement = 78%).

## 5 Result Analysis

### 5.1 How does Different Prompt Strategies Affect Novelty Evaluation?

For this question, we focus exclusively on the results pertaining to prompt tuning. As evidenced by the findings in Table 2, most models achieve their

best performance in Relevance under the zero-shot prompting strategy. However, the maximum average score attained is only 3.6983, which suggests that LLMs may be unable to fully grasp the novelty described in the paper. Conversely, in the few-shot setting, model capability in Coverage and Correctness (DistAcc) demonstrates a noticeable improvement. However, this improvement is accompanied by a decrease in relevance. This trade-off suggests that when provided with human-evaluated examples, the LLM may be merely simulating human expression patterns and sentiment distribution rather than writing a genuine novelty evaluation. Furthermore, performance in clarity improves significantly in the RAG scenario. This outcome shows that the utilization of externally retrieved information helps in organizing and articulating the evaluation, resulting in output text with a clearer and more comprehensible structure. Simultaneously, the RAG approach leads to a reduction in relevance compared to both zero-Shot and few-Shot methods. This potential trade-off implies a key issue: while the retrieved information is comprehensive, the model may be misled by the retrieval results in knowledge retrieval process. Consequently, this weakens its focus on the paper’s novelty.

### 5.2 Can Specialized LLMs Improve Novelty Evaluation?

We hypothesized that models subjected to parameter fine-tuning on peer review datasets would exhibit better performance. However, the results presented in Table 2 indicate that these models only show a marginal advantage. Specifically, only the CycleReviewer-70B (large-parameter models), and the SEA series (contain data in NLP conferences), demonstrate better performance. We observe that CycleReviewer-70B and the SEA models maintain comparable scores in *Relevance* while demonstrating superior performance over the general-purpose models across the other three dimensions. This finding suggests that while learning from human data results in a more anthropomorphic output style, it does not translate to a deeper, more robust understanding of novelty evaluation for this task. Furthermore, the Reviewer2 model performed particularly poor across all metrics. An inspection of its generated output revealed a significant issue with instruction following, as illustrated in the accompanying Figure 3. We suspect that this model struggles to follow the given prompt instructions. This may be due to fine-tuning on highly

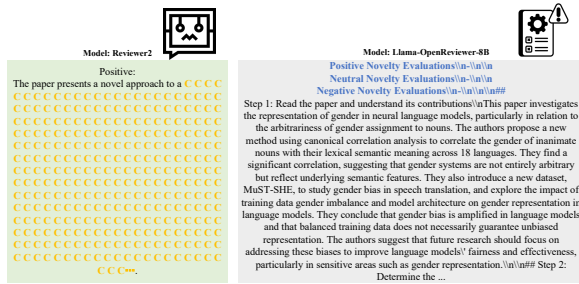


Figure 3: Examples of Instruction-Following Failures by the Specialized Model.

specific training prompts, which weakens its general instruction-following ability. We checked that other specialized models (The detail in Appendix F) exhibit similar problems, though the deficiency is most pronounced in Reviewer2. From these results, we think that models with a larger number of parameters and those designed to handle low-quality and inconsistent data are better equipped to provide strong instruction-following capabilities, rather than being fixed to a specific prompt.

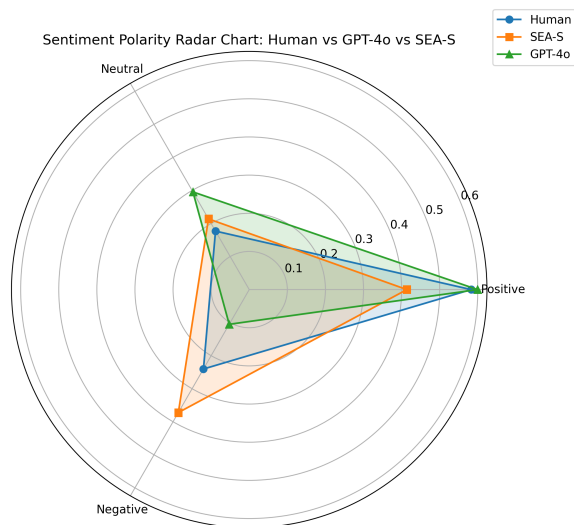


Figure 4: Comparison of Sentiment Polarity Distributions Among Human, General LLM, and Specialized LLM.

### 5.3 How Do LLM Novelty Evaluations Differ from Human Judgments?

We selected two comparatively strong models, GPT-4o as a representative general model and SEA-S as a Specialized model, and evaluated their performance across all dimensions. As shown in the Appendix G, for *Relevance*, both models produce evaluations that are highly aligned with the nov-

elty descriptions provided in the original papers. The generated positive evaluations in particular are almost entirely grounded in the explicitly stated methodologies and contributions. This suggests that both models are capable of identifying the core innovation claims of a paper. However, they also exhibit several issues, including exaggerating positive contributions, forcefully identifying negative aspects, introducing details not present in the source text, and producing overly templated and verbose assessments. For *Coverage*, the LLMs reliably capture the primary contributions, but they fall short in assessing the breadth of novelty. When a paper contains multiple innovation points, the models often fail to cover them comprehensively, potentially due to low sensitivity to different types of novelty. The models' performance on *Clarity* is strong, indicating that they are able to extract and articulate the core concepts described in the paper. Finally, we compared the sentiment distributions of model-generated evaluations against human-written evaluations, as shown in the Figure 4. GPT-4o exhibits a distribution similar to humans for positive evaluations, but produces fewer negative evaluations and more neutral ones. In contrast, SEA-S displays the opposite trend: it produces substantially more negative and fewer positive evaluations. This suggests that general-purpose models tend to accommodate user expectations by generating more favorable feedback, whereas models fine-tuned on peer review data adopt a more critical stance, sometimes excessively so, potentially leading them to overemphasize or even fabricate negative points.

### 5.4 Analysis of LLM Performance Across Novelty Evaluation Metrics

Regarding *Relevance*, although LLMs exhibit a surface-level understanding of novelty, they struggle to capture the specific and fine-grained content of novelty claims. This limitation is particularly evident under the RAG prompting setting, where performance degrades noticeably. These results indicate that retrieval augmentation or advanced prompting alone is insufficient to support genuine novelty understanding, and that specialized fine-tuning remains necessary.

For *Correctness*, better-performing specialized models achieve higher scores, suggesting that fine-tuning allows LLMs to learn human expressive and structural patterns. However, due to limited novelty understanding, these models often produce hedg-

ing evaluations with mixed sentiment, preventing optimal performance.

Across all models, *Coverage* remains sub-optimal. Even when restricted to novelty descriptions from the introduction, LLMs emphasize points that diverge from those identified by human reviewers. This highlights an important open challenge: enabling LLMs to better model how humans assess the breadth of novelty.

In contrast, LLMs perform well on *Clarity*, effectively identifying key term and major contributions in novelty descriptions, largely due to strong information extraction capabilities rather than a deeper understanding of novelty.

Finally, we observe that some models fine-tuned on peer review data exhibit severe instruction-following issues, leading to substantial performance decrease and highlighting the need for improved fine-tuning strategies.

## 5.5 Additional Analyses

To further assess potential data contamination, temporal effects, and model behavior under varying conditions, we conduct a series of additional analyses. The results show that performance is largely consistent across model generations and publication years, and remains stable under controlled input perturbations, suggesting that it is not primarily driven by memorization or temporal leakage.

We also examine performance across paper types and levels of reviewer disagreement. Models perform better on resource papers than on methodological papers, indicating that evaluation difficulty varies by contribution type. In cases of reviewer disagreement, LLM-generated evaluations show higher similarity to high-confidence reviews, suggesting that their alignment behavior is not arbitrary.

Overall, these findings demonstrate that the proposed benchmark supports systematic and fine-grained analysis of LLM behavior beyond aggregate performance. Detailed results are provided in Appendix H.

## 6 Conclusion

This paper proposes NovBench, a benchmark designed to systematically evaluate the ability of LLMs to assess academic novelty. NovBench employs four distinct dimensions to quantify evaluation quality, using a controlled and homogeneous setting to ensure reliability and isolate the novelty

assessment task. We demonstrate the performance of both general-purpose and specialized LLMs in evaluating the novelty of academic papers under varying prompting conditions. Through a comprehensive analysis of the novelty evaluations generated by different LLMs across all four dimensions, we discuss key insights intended to guide future development in this field. In future work, we plan to extend the benchmark to additional venues using the same data construction pipeline, enabling the study of cross-venue and cross-domain generalization of LLMs. Automatically deriving evaluation dimensions from reviewer guidelines is also an interesting direction for future work, and our framework can be extended to support such dynamic rubrics.

## Limitations

This study is subject to several limitations. First, our work exclusively uses the paper introduction as the text source for novelty evaluation. While the introduction contains the primary novelty claims, relying solely on this section rather than on the full paper may omit detailed content needed to fully support the evaluation.

Second, the data are sourced from COLING and EMNLP proceedings, where the readily available peer review text predominantly corresponds to accepted papers, potentially introducing selection bias. In addition, our benchmark is constructed from a limited set of NLP venues, which may restrict its generalizability to broader research domains, as conferences such as ICLR and NeurIPS adopt different review formats and scoring rubrics and cover broader interdisciplinary areas.

Third, we employ relatively simple prompt engineering strategies and do not explore more advanced prompting techniques or multi-agent architectures. Moreover, the credibility of reviewer comments remains an important concern, and we do not incorporate numerical scores (e.g., confidence scores) into the analysis. Furthermore, calibrating model confidence (Zhang et al., 2025, 2026) could further improve the reliability of generated outputs. Meanwhile, identifying and removing potentially AI-generated (Liu et al., 2023; Li et al., 2025) traces in human review comments is also an important step in ensuring data quality.

Fourth, although EMNLP places greater emphasis on methodological novelty, our analysis does not distinguish between different types of novelty.

Finally, despite the effectiveness of our proposed metrics, further research is needed to develop more robust evaluation methods.

Future work may explore more fine-grained taxonomy design, analyze hallucination patterns in novelty evaluation, and investigate multi-model aggregation approaches (e.g., ensembling or multi-agent methods) within the proposed framework. Despite these limitations, our study provides a useful reference for automated academic novelty assessment and LLM-based evaluation.

## Ethics Statement

This study is conducted in accordance with established ethical standards for research involving human-authored text. All data used in this work are openly available peer review reports released by conferences or journals, and do not contain personally identifiable information beyond what is already publicly disclosed. We do not collect new personal data, and our analysis poses no additional risk of privacy leakage or harm to authors or reviewers.

Importantly, the goal of this work is not to develop or promote automated peer review systems as a replacement for human expert reviewers. Instead, our focus is on evaluating the ability of large language models to assist in specific, well-scoped aspects of the review process—namely, the analysis and assessment of novelty—under controlled and transparent settings. We view such tools as potential supporting instruments that may help reduce reviewer workload or provide complementary perspectives, rather than substitutes for human judgment, expertise, or accountability.

We acknowledge the broader ethical concerns surrounding the use of LLMs in peer review, including risks of over-reliance, bias amplification, and misuse. Our work is intended to contribute empirical evidence that informs these discussions, rather than to advocate for the deployment of LLMs as autonomous reviewers.

## Acknowledgments

This work is supported by the Major Project of the National Social Science Fund of China (Grant No.25&ZD298). This research utilised Queen Mary’s Apocrita HPC facility, supported by QMUL Research-IT (King et al., 2017).

## References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. [Phi-4 technical report](#). *Preprint*, arXiv:2412.08905.
- David Adam. 2025. The peer-review crisis: how to fix an overloaded system. *Nature*, 644(8075):24–27.
- Osama Mohammed Afzal, Preslav Nakov, Tom Hope, and Iryna Gurevych. 2026. [Beyond “not novel enough”: Enriching scholarly critique with LLM-assisted feedback](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2648–2671, Rabat, Morocco. Association for Computational Linguistics.
- Bruce Alberts, Brooks Hanson, and Katrina L. Kellner. 2008. [Reviewing peer review](#). *Science*, 321(5885):15–15.
- Tong Bao, Mir Tafseer Nayeem, Davood Rafiei, and Chengzhi Zhang. 2025. [SurveyGen: Quality-aware scientific survey generation with large language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2712–2736, Suzhou, China. Association for Computational Linguistics.
- Rewina Bedemariam, Natalie Perez, Sreyoshi Bhaduri, Satya Kapoor, Alex Gil, Elizabeth Conjar, Ikkei Itoku, David Theil, Aman Chadha, and Naumaan Nayyar. 2025. [Potential and perils of large language models as judges of unstructured textual data](#). *Preprint*, arXiv:2501.08167.
- Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Hayden Kwok-Hay So, Zhijiang Guo, Liya Zhu, and Ngai Wong. 2025. [TreeReview: A dynamic tree of questions framework for deep and efficient LLM-based scientific peer review](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15662–15693, Suzhou, China. Association for Computational Linguistics.
- Junyi Chen, Shihao Bai, Zaijun Wang, Siyu Wu, Chuheng Du, Hailong Yang, Ruihao Gong, Shengzhong Liu, Fan Wu, and Guihai Chen. 2025. [Pre<sup>3</sup>: Enabling deterministic pushdown automata for faster structured LLM generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11253–11267, Vienna, Austria. Association for Computational Linguistics.
- Barak A Cohen. 2017. [Point of view: How should novelty be valued in science?](#) *eLife*, 6:e28699.

- Italo Luis da Silva, Hanqi Yan, Lin Gui, and Yulan He. 2025. [GraphMind: Interactive novelty assessment system for accelerating scientific discovery](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 286–294, Suzhou, China. Association for Computational Linguistics.
- Mike D’Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. [Marg: Multi-agent review generation for scientific papers](#). *Preprint*, arXiv:2401.04259.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, and 21 others. 2024. [LLMs assist NLP researchers: Critique paper \(meta-\)reviewing](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099, Miami, Florida, USA. Association for Computational Linguistics.
- Nils Dycke, Iliia Kuznetsov, and Iryna Gurevych. 2023. [NLPeer: A unified resource for the computational study of peer review](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychol. Bull.*, 76(5):378–382.
- Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. 2024. [Reviewer2: Optimizing review generation through prompt generation](#). *Preprint*, arXiv:2402.10886.
- Google Gemini Team. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Shengzhi Huang, Yong Huang, Yinpeng Liu, Zhuoran Luo, and Wei Lu. 2025a. [Are large language models qualified reviewers in originality evaluation?](#) *Information Processing & Management*, 62(3):103973.
- Xuanjing Huang, Shihan Dou, and Zhangyue Yin. 2025b. The dual-edged sword: artificial intelligence’s evolving role in academic peer review. *Sci. China Inf. Sci.*, 68(11).
- Maximilian Idahl and Zahra Ahmadi. 2025. [OpenReviewer: A specialized large language model for generating critical scientific paper reviews](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 550–562, Albuquerque, New Mexico. Association for Computational Linguistics.
- Daeseong Jeon, Junyoun Lee, Joon Mo Ahn, and Changyong Lee. 2023. [Measuring the novelty of scientific publications: A fasttext and local outlier factor approach](#). *Journal of Informetrics*, 17(4):101450.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. [AgentReview: Exploring peer review dynamics with LLM agents](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1208–1226, Miami, Florida, USA. Association for Computational Linguistics.
- Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. [A dataset of peer reviews \(PeerRead\): Collection, insights and NLP applications](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.
- Thomas King, Simon Butcher, and Lukasz Zalewski. 2017. [Apocrita - High Performance Computing Cluster for Queen Mary University of London](#).
- Sandeep Kumar, Tirthankar Ghosal, Vinayak Goyal, and Asif Ekbal. 2025. [Can large language models unlock novel scientific research ideas?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33551–33575, Suzhou, China. Association for Computational Linguistics.
- Iliia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K. Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, Sheng Lu, Mausam, Margot Mieskes, Aur  lie N  v  ol, Danish Pruthi, Lizhen Qu,

- Roy Schwartz, Noah A. Smith, Thamar Solorio, and 5 others. 2024. [What can natural language processing do for peer review?](#) *Preprint*, arXiv:2405.06563.
- Iliia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. 2022. [Revise and resubmit: An intertextual model of text-based collaboration in peer review.](#) *Computational Linguistics*, 48(4):949–986.
- Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. [An argument-annotated corpus of scientific publications.](#) In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, Brussels, Belgium. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA. Curran Associates Inc.
- Jiyi Li, Ayaka Sato, Kazuya Shimura, and Fumiyo Fukumoto. 2020. [Multi-task peer-review score prediction.](#) In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 121–126, Online. Association for Computational Linguistics.
- Xiangci Li, Gully Burns, and Nanyun Peng. 2021. [Scientific discourse tagging for evidence extraction.](#) In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2550–2562, Online. Association for Computational Linguistics.
- Yuanfan Li, Zhaohan Zhang, Chengzhengxu Li, Chao Shen, and Xiaoming Liu. 2025. [Iron sharpens iron: Defending against attacks in machine-generated text detection with adversarial training.](#) *Preprint*, arXiv:2502.12734.
- Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou. 2024. [Can large language models provide useful feedback on research papers? a large-scale empirical analysis.](#) *NEJM AI*, 1(8):AIoa2400196.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries.](#) In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Ethan Lin, Zhiyuan Peng, and Yi Fang. 2025. [Evaluating and enhancing large language models for novelty assessment in scholarly publications.](#) In *Proceedings of the 1st Workshop on AI and Scientific Discovery: Directions and Opportunities*, pages 46–57, Albuquerque, New Mexico, USA. Association for Computational Linguistics.
- Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023. [Automated scholarly paper review: Concepts, technologies, and challenges.](#) *Information Fusion*, 98:101830.
- Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. 2023. Coco: Coherence-enhanced machine-generated text detection under low resource with contrastive learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16167–16188.
- Yan Liu, Zonglin Yang, Soujanya Poria, Thanh-Son Nguyen, and Erik Cambria. 2025a. [Harnessing large language models for scientific novelty detection.](#) *Preprint*, arXiv:2505.24615.
- Yan Liu, Zonglin Yang, Soujanya Poria, Thanh-Son Nguyen, and Erik Cambria. 2025b. [Harnessing large language models for scientific novelty detection.](#) *Preprint*, arXiv:2505.24615.
- Annie Louis and Ani Nenkova. 2013. [Automatically assessing machine summary content without a gold standard.](#) *Computational Linguistics*, 39(2):267–300.
- Sheng Lu, Iliia Kuznetsov, and Iryna Gurevych. 2025. [Identifying aspects in peer reviews.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6145–6167, Suzhou, China. Association for Computational Linguistics.
- Kuniko Matsumoto, Sotaro Shibayama, Byeongwoo Kang, and Masatsura Igami. 2021. Introducing a novelty indicator for scientific research: validating the knowledge-based combinatorial approach. *Scientometrics*, 126(8):6891–6915.
- OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastian Bubeck, and 108 others. 2025. [gpt-oss-120b & gpt-oss-20b model card.](#) *Preprint*, arXiv:2508.10925.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. [Gpt-4o system card.](#) *Preprint*, arXiv:2410.21276.
- OpenAI. 2025. [Gpt-5 system card.](#) <https://cdn.openai.com/gpt-5-system-card.pdf>.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation.](#) In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. [Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter](#). *Preprint*, arXiv:1910.01108.
- Simra Shahid, Marissa Radensky, Raymond Fok, Pao Siangliulue, Daniel S Weld, and Tom Hope. 2025. [Literature-grounded novelty assessment of scientific ideas](#). In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 96–113, Vienna, Austria. Association for Computational Linguistics.
- Sotaro Shibayama, Zhao Wu, Deyun Yin, and Kazuki Yokota. 2025. [State of the art of novelty indicators](#). Technical report, SSRN Electronic Journal. Available at SSRN: <https://ssrn.com/abstract=5379973>.
- Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, Philip Torr, Bowen Zhou, and Nanqing Dong. 2025. [Many heads are better than one: Improved scientific idea generation by a LLM-based multi-agent system](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28201–28240, Vienna, Austria. Association for Computational Linguistics.
- Hongming Tan, Shaoxiong Zhan, Fengwei Jia, Hai-Tao Zheng, and Wai Kin (Victor) Chan. 2026. [A hierarchical framework for measuring scientific paper innovation via large language models](#). *Information Sciences*, 728:122787.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruiti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. [Atypical combinations and scientific impact](#). *Science*, 342(6157):468–472.
- Reinhilde Veugelers and Jian Wang. 2019. [Scientific novelty and technological impact](#). *Research Policy*, 48(6):1362–1372.
- Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. [ReviewRobot: Explainable paper review generation based on knowledge synthesis](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 384–397, Dublin, Ireland. Association for Computational Linguistics.
- Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. [Autosurvey: Large language models can automatically write surveys](#). In *The Thirtieth Annual Conference on Neural Information Processing Systems*.
- Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. [Cyclereviewer: Improving automated research via automated review](#). In *The Thirteenth International Conference on Learning Representations*.
- Dustin Wright, Jiayin Pei, David Jurgens, and Isabelle Augenstein. 2022. [Modeling information change in science communication with semantically matched paraphrases](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1783–1807, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Wenqing Wu, Chengzhi Zhang, Tong Bao, and Yi Zhao. 2025a. [Sc4anm: Identifying optimal section combinations for automated novelty prediction in academic papers](#). *Expert Systems with Applications*, 273:126778.
- Wenqing Wu, Chengzhi Zhang, and Yi Zhao. 2025b. [Automated novelty evaluation of academic paper: A collaborative approach integrating human and large language model knowledge](#). *Journal of the Association for Information Science and Technology*, 76(11):1452–1469.
- Zhijian Xu, Yilun Zhao, Manasi Patwardhan, Lovekesh Vig, and Arman Cohan. 2025. [Can LLMs identify critical limitations within scientific research? a systematic evaluation on AI research papers](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20652–20706, Vienna, Austria. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Heng Yu, Junfeng Kang, Rui Li, Qi Liu, Liyang He, Zhenya Huang, Shuanghong Shen, and Junyu Lu.

2025. **CA-GAR: Context-aware alignment of LLM generation for document retrieval**. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5836–5849, Vienna, Austria. Association for Computational Linguistics.
- Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, Renjing Cui, Chengcheng Han, Qiushi Sun, Zhiyong Wu, Yunshi Lan, and Xiang Li. 2024a. **Automated peer reviewing in paper SEA: Standardization, evaluation, and analysis**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10164–10184, Miami, Florida, USA. Association for Computational Linguistics.
- Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, Renjing Cui, Chengcheng Han, Qiushi Sun, Zhiyong Wu, Yunshi Lan, and Xiang Li. 2024b. **Automated peer reviewing in paper SEA: Standardization, evaluation, and analysis**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10164–10184, Miami, Florida, USA. Association for Computational Linguistics.
- Weizhe Yuan and Pengfei Liu. 2022. **Kid-review: Knowledge-guided scientific review generation with oracle pre-training**. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11639–11647.
- Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. **Can we automate scientific reviewing?** *Journal of Artificial Intelligence Research*, 75:171–212.
- Hongxin Zhang, Yanzhe Zhang, Ruiyi Zhang, and Diyi Yang. 2022. **Robustness of demonstration-based learning under limited data scenario**. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1769–1782, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tianyi Zhang\*, Varsha Kishore\*, Felix Wu\*, Kilian Q. Weinberger, and Yoav Artzi. 2020. **Bertscore: Evaluating text generation with bert**. In *International Conference on Learning Representations*.
- Zhaohan Zhang, Chengzhengxu Li, Xiaoming Liu, Chao Shen, Ziquan Liu, and Ioannis Patras. 2026. Confidence should be calibrated more than one turn deep. *arXiv preprint arXiv:2604.05397*.
- Zhaohan Zhang, Ziquan Liu, and Ioannis Patras. 2025. Grace: A generative approach to better confidence elicitation in large language models. *arXiv preprint arXiv:2509.09438*.
- Yi Zhao and Chengzhi Zhang. 2025. **A review on the novelty measurements of academic papers**. *Scientometrics*, 130(2):727–753.
- Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. **Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks**. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, Torino, Italia. ELRA and ICCL.
- Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025a. **DeepReview: Improving LLM-based paper review with human-like deep thinking process**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29330–29355, Vienna, Austria. Association for Computational Linguistics.
- Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025b. **DeepReview: Improving LLM-based paper review with human-like deep thinking process**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29330–29355, Vienna, Austria. Association for Computational Linguistics.
- Zhenzhen Zhuang, Jiandong Chen, Hongfeng Xu, Yuwen Jiang, and Jialiang Lin. 2025. **Large language models for automated scholarly paper review: A survey**. *Information Fusion*, 124:103332.

## A Supplement of Automatic Extraction of Novelty Descriptions

The accurate extraction of novelty descriptions from the paper introduction constitutes a critical step. We began by manually annotating the novelty descriptions within the introductions of the COLING 2020 papers sourced from NLPeer (Dycke et al., 2023), covering 87 papers, 2,300 total sentences, of which 533 were classified as novelty description sentences. The manual annotation process was executed by the two experienced journal and conference reviewers, we ask them to determine whether a given sentence constitutes a description of the paper’s novelty, with reference to the surrounding context, achieving a Cohen’s  $\kappa$  inter-rater agreement of 0.831. We framed the automatic novelty description extraction as a binary classification task, where the model is required to identify whether a given sentence constitutes a novelty description. Specifically, we designed various prompting strategies (zero-shot see Figure 5, few-shot see Figure 6, step-by-step see Figure 7, and in-context learning prompt see Figure 8) to benchmark the performance of various LLMs on this task. The results are presented in the Figure 9. As shown in the Figure 9, the context prompt strategy yielded the best performance across all models, with GPT-5 achieving the highest Accuracy (0.89) and Macro F1 score (0.84). Consequently, we selected the context-prompted GPT-5 as the model for the automatic extraction of novelty descriptions.

You are a senior academic reviewer familiar with how research papers describe their novelty.

Given a sentence from the Introduction section of a paper, determine whether it describes the novelty of this paper. Novelty typically includes introducing new methods, models, tasks, datasets, perspectives, or achieving new improvements or combinations that have not been explored before.

Only consider whether the sentence refers to this paper’s novelty—not the novelty of prior work.

Answer with:

Yes — if the sentence describes the novelty of this paper

No — otherwise

Sentence:  
{sentence}

Figure 5: The Zero-Shot Prompt for Novelty Description Extraction.

## B Supplement of Automatic Extraction of Novelty Evaluations

Similarly, the accurate extraction of novelty evaluations from the peer review text is equally crucial. We first obtained all novelty-related evaluation instances (totaling 493 comments) from the public resource shared by Lu et al. (Lu et al., 2025), a dataset related to peer review aspect identification. We then randomly selected 500 instances of non-novelty evaluations, framing the task as a binary classification task for model testing. Specifically, given a review sentence extracted from the peer review text, the model is required to judge whether it is a novelty evaluation. We benchmarked the performance of the deep learning models provided by Yuan et al. (Yuan et al., 2022) against several LLMs, which executed the task under zero-shot (see Figure 11) and RAG (see Figure 12) prompt. The specific results are presented in the accompanying Figure 10. The results in the Figure 10 indicate that GPT-4o-mini and GPT-5 achieved the best performance under the zero-shot prompting strategy, registering the highest combined Accuracy (0.93) and Macro F1 score (0.93). In consideration of cost-effectiveness, we designated the zero-shot prompted GPT-4o-mini model as the model for extracting novelty evaluations.

## C Supplement of Sentiment-Based Normalization of Novelty Evaluations

To ensure fair comparison between human-written and LLM-generated evaluations, we use a prompt that instructs GPT-4o to (1) deduplicate semantically similar comments, (2) consolidate them into concise statements, and (3) categorize them by sentiment polarity. The exact prompt used in our ex-

periments is reproduced below, and illustrated in Figure 13. This prompt ensures that all novelty-related feedback is standardized into a consistent and non-redundant set of evaluative statements, enabling more reliable automatic evaluation of novelty descriptions generated by LLMs.

## D Supplement of Agreement Evaluation

This appendix provides the detailed instructions, examples (see Figure 14), and guidelines (see Figure 15) used for the human evaluation of model-generated novelty assessments. We employ four human evaluators with strong expertise in Natural Language Processing (NLP), including two Ph.D. students, one Associate Professor, and one Lecturer. Each evaluator independently assesses, for each sample, which of the two models (Model A or Model B) produces a higher-quality novelty evaluation. The primary objective of this human evaluation is to validate the effectiveness of the proposed automatic evaluation metrics. Inter-annotator agreement is measured using Fleiss’  $\kappa$  (Fleiss, 1971), yielding a score of 0.72, which indicates substantial agreement. To compare human judgments with automatic metrics, we compute both the Spearman rank correlation coefficient ( $\rho$ ) and an agreement score that measures whether the metric selects the same preferred model as the aggregated human judgment. Formally, let  $H_i^{(j)}$  denote the preference of the  $j$ -th annotator on sample  $i$ , where  $j = 1, \dots, N$ . The aggregated human preference  $H_i$  is obtained via majority voting across annotators. Samples without a strict majority are excluded from the agreement computation. The agreement between the automatic metric and human judgments is defined as:

$$\text{Agreement} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \mathbf{1}(H_i = M_i), \quad (8)$$

where  $\mathcal{D}$  denotes the set of samples with valid aggregated labels,  $M_i$  is the prediction of the automatic metric, and  $\mathbf{1}(\cdot)$  is the indicator function.

## E Experiment Implementation Details

During testing on NovBench, we evaluated various general and specialized LLMs using three distinct prompt tuning strategies: zero-shot (see Figure 16), few-shot (see Figure 17), and RAG (see Figure 18). For the zero-shot setting, the model is provided only with the extracted novelty descriptions. For the few-shot setting, the

You are a senior academic reviewer familiar with how research papers describe their novelty.  
Your task is to determine whether a given sentence from the Introduction section describes the novelty or contribution of this paper.

Definition of Novelty:  
Novelty includes introducing new methods, models, tasks, datasets, perspectives, or achieving new improvements or combinations that have not been explored before.  
Do NOT mark sentences as novel if they only discuss prior work’s novelty or general background.

Answer strictly with:  
- Yes — if the sentence describes this paper’s novelty.  
- No — otherwise.

Examples:

# Simple clear cases  
Sentence: "In this paper, we propose a novel transformer-based architecture that integrates syntactic information into language modeling."  
Answer: Yes

Sentence: "Previous studies have explored the use of graph neural networks for text classification."  
Answer: No  
.....

# Borderline cases  
Sentence: "We show that applying existing attention mechanisms to longer documents leads to performance improvements."  
Answer: No

Sentence: "This paper is the first to conduct a comprehensive comparison of previously proposed algorithms for neural parsing."  
Answer: Yes  
.....

Now classify the following sentence:

Sentence:  
{sentence}

Figure 6: The Few-Shot Prompt for Novelty Description Extraction.

model receives the extracted novelty descriptions along with two analogous examples selected from our dataset as demonstrations. For the RAG setting, the model is provided with the extracted novelty descriptions and additional retrieved context, where the retrieval corpus consists of titles and abstracts of ACL, EMNLP, and NAACL papers published between 2019 and 2022, sourced from the ACL Anthology. Specifically, we utilized `acl-anthology-helper` (<https://github.com/tangg555/acl-anthology-helper>) to acquire and store the ACL Anthology papers in a local database. We then filtered this repository to include only the titles and abstracts from the specified ACL, EMNLP, and NAACL proceedings (2019–2022). Retrieval was executed using the abstract of each paper in NovBench as the query, ultimately yielding the 5 most relevant titles and abstracts per paper to serve as the RAG content.

Here, we provide additional details on the eight fine-tuned LLMs. The CycleReviewer (8B’s backbone is Mistral-Nemo-12B<sup>5</sup>, 70B’s backbone is Qwen2.5-Instruct-72B (Qwen et al., 2025)) se-

<sup>5</sup><https://mistral.ai/news/mistral-nemo>

ries models are primarily fine-tuned on peer review data from ICLR 2024, covering the fields of machine learning and artificial intelligence. The DeepReviewer (backbone is Phi-4 (Abdin et al., 2024)) series models are mainly fine-tuned on peer review data from ICLR 2024 and ICLR 2025, also spanning machine learning and artificial intelligence. Llama-OpenReview-8B (backbone is Llama-3.1-8B-Instruct (Grattafiori et al., 2024)) is fine-tuned on peer review data from ICLR and NeurIPS (post-2022), covering machine learning and artificial intelligence. Reviewer2 (backbone is Llama-2-7B-Chat (Touvron et al., 2023)) is primarily fine-tuned on peer review data from NLPeer (CoNLL-16, ACL-17, COLING-20, ARR-22), ICLR 2017–2023, and NeurIPS 2016–2022, covering machine learning, natural language processing, computational linguistics, and artificial intelligence, with approximately 7B parameters. SEA-E and SEA-S are mainly fine-tuned on peer review data from NLPeer (CoNLL-16, ACL-17, COLING-20, ARR-22), NeurIPS 2016–2023, and ICLR 2017–2024, covering machine learning, natural language processing, computational linguistics,

<p># Step 1 and Step 2 You are a careful academic language analyst.</p> <p>Given the sentence below, perform two tasks:</p> <ol style="list-style-type: none"> <li>Briefly summarize what the sentence is mainly saying.</li> <li>Indicate whether it refers to this paper's own work, or to prior work or general background.</li> </ol> <p>Sentence: <i>{sentence}</i></p> <p>Answer format: Main idea: ... Refers to: [This paper / Previous work / General background]</p>	<p>#Step 3 Given the following context:</p> <p>Main idea: <i>{main_idea}</i> Refers to: <i>{refers_to}</i></p> <p>Does this sentence describe the novelty or original contribution of this paper?</p> <p>Novelty may include introducing a new method, dataset, model, task, theory, or a new combination/improvement. Do not rely only on keywords like “new” or “novel” — base your judgment on meaning.</p> <p>Answer with: Yes or No.</p>
--	---

Figure 7: The Step by Step Prompt for Novelty Description Extraction.

and artificial intelligence, and both backbone is Mistral-7B-Instruct-v0.2 (Jiang et al., 2023).

LLM inference was executed utilizing A100 80GB GPUs and H100 80GB GPUs. Specifically, models sized 8B, 14B, 20B, and 32B, along with CycleReviewer-8B, DeepReviewer-7B, Llama-OpenReviewer-8B, Reviewer2, SEA-E, and SEA-S, were run on a single A100 80GB GPU. Models at the 70B parameter scale and DeepReviewer-14B required inference to be distributed across two A100 80GB GPUs. Finally, the gpt-oss-120B model was allocated across two H100 80GB GPUs. It is important to note that we employed the Fast Mode configuration for all inferences involving CycleReviewer and DeepReviewer. The total inference time per model, contingent upon its parameter size, ranged from 5 to 70 hours. For Closed-source models, the inference process was implemented through official API integration..

## F Supplemental Analysis of Instruction-Following Deficiencies in Specialized Review Generation Models

Beyond the particularly severe instruction-following deficiencies reported in Section 5.2, we observed that other models subjected to parameter fine-tuning on peer review feedback exhibit similar, and arguably unacceptable, operational failures. These specific issues are visually documented in the accompanying Figure 19. As depicted, CycleReviewer-8b suffers from the propensity to generate repetitive evaluations, whereas DeepReviewer exhibits the problem of producing null or empty evaluations.

Model	Rel.	Cov.	Clarity	DistAcc
GPT-3.5 Zero	3.556	0.228	0.663	0.676
GPT-3.5 Few	3.505	0.246	0.660	0.731
GPT-3.5 RAG	3.462	0.237	0.667	0.679
GPT-4o Zero	3.698	0.233	0.660	0.698
GPT-4o Few	3.561	0.240	0.659	0.709
GPT-4o RAG	3.448	0.224	0.667	0.697
Gemini-2.5-flash Zero	3.471	0.212	0.641	0.601
Gemini-2.5-flash Few	3.473	0.236	0.657	0.659
Gemini-2.5-flash RAG	3.509	0.227	0.668	0.592

Table 3: Results of GPT-3.5, GPT-4o and Gemini-2.5-flash on EMNLP 2023.

## G Case Studies Comparing Human and LLM-Generated Novelty Evaluations

We selected five case studies for the analysis presented in Section 5.3. Each case study comprises the novelty description extracted from the paper introduction, the corresponding novelty evaluation provided by the human reviewer, and the novelty evaluations generated by GPT-4o and SEA-S. These examples are specifically illustrated in Figures 20, 21, 22, 23, and 24.

## H Additional Analyses

### H.1 Memorization and Temporal Analysis

To assess potential data contamination and temporal leakage, we conduct a series of complementary experiments. First, we evaluate an earlier model (GPT-3.5), released prior to EMNLP 2023, under the same prompting settings as other models. The results (Table 3) show that GPT-3.5 performs competitively among general LLMs, indicating that performance is not primarily driven by access to more recent training data. Second, we perform cross-year evaluation by comparing model performance

You are a senior academic reviewer familiar with how research papers describe their novelty.

You will be given a **target sentence** from the Introduction section of a research paper, along with its **context** (several surrounding sentences).

Your task is to determine whether the **target sentence** describes the novelty or contribution of **this paper**.

- Novelty typically includes introducing new methods, models, tasks, datasets, perspectives, or achieving new improvements or combinations that have not been explored before.
- Only consider whether the target sentence refers to **this paper’s novelty**—not the novelty of prior work.
- Use the provided context to assist your judgment, but base your decision specifically on the target sentence.

Answer with:

- Yes — if the target sentence describes the novelty of this paper.
- No — otherwise.

Context:

*{context}*

Target sentence:

*{sentence}*

Figure 8: The Context Prompt for Novelty Description Extraction. We set the context window size to 2, meaning we utilized the two preceding sentences and the two succeeding sentences as contextual information. Boundary conditions were handled such that the first sentence included only succeeding context (post-text), and the last sentence included only preceding context (pre-text).

Model	Rel.	Cov.	Clarity	DistAcc
GPT-3.5 Zero	3.611	0.098	0.664	0.578
GPT-3.5 Few	3.580	0.152	0.661	0.623
GPT-4o Zero	3.762	0.111	0.659	0.583
GPT-4o Few	3.580	0.118	0.658	0.555

Table 4: Results of GPT-3.5 and GPT-4o on COLING 2020.

on COLING 2020 and EMNLP 2023 datasets (Table 4). The results show no substantial performance differences across publication years.

Third, we test for verbatim memorization by prompting models to continue review sentences. In all cases, models respond that they are not certain about the continuation, suggesting the absence of exact recall. Finally, we conduct input perturbation experiments by modifying novelty descriptions through paraphrasing (“change”) and partial deletion (“del”). As shown in Table 5, model performance remains largely stable across all evaluation dimensions.

Overall, these results consistently suggest that

Setting	Rel.	Cov.	Clarity	DistAcc
GPT-4o Few	3.569	0.221	0.659	0.699
GPT-4o Few (change)	3.480	0.238	0.661	0.675
GPT-4o Few (del)	3.466	0.201	0.658	0.671
GPT-4o RAG	3.460	0.243	0.665	0.684
GPT-4o RAG (change)	3.430	0.196	0.667	0.681
GPT-4o RAG (del)	3.392	0.168	0.668	0.687
GPT-4o Zero	3.702	0.222	0.659	0.680
GPT-4o Zero (change)	3.657	0.197	0.657	0.677
GPT-4o Zero (del)	3.614	0.184	0.658	0.637

Table 5: Results of perturbation experiments on GPT-4o.

model behavior is not explained by memorization or temporal leakage, but reflects the intrinsic difficulty of novelty evaluation.

## H.2 Analysis by Paper Type

To investigate whether model performance varies across paper types, we classify papers into coarse-grained categories (methodological and resource papers) using GPT-4o based on titles and abstracts. The benchmark results are then grouped accordingly. Specifically, we report a subset of repre-

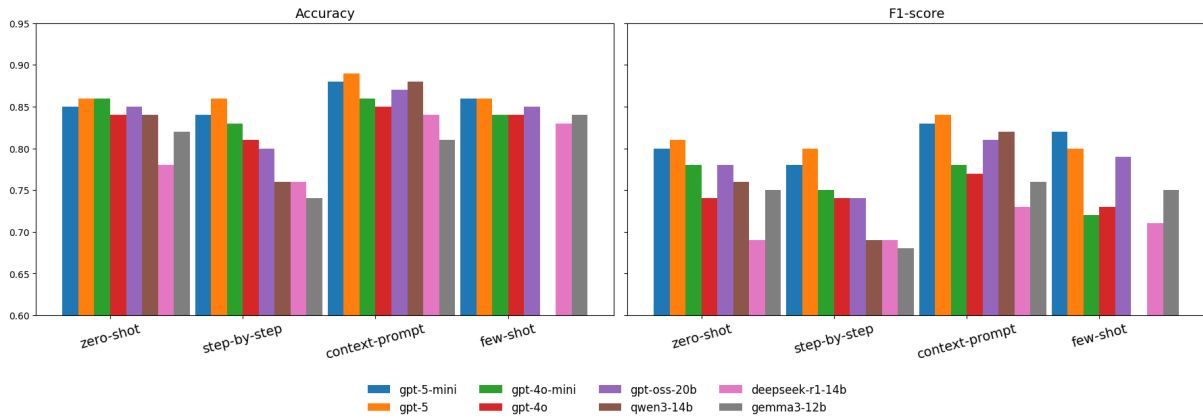


Figure 9: The performance of various LLMs on novelty description extraction under different prompt.

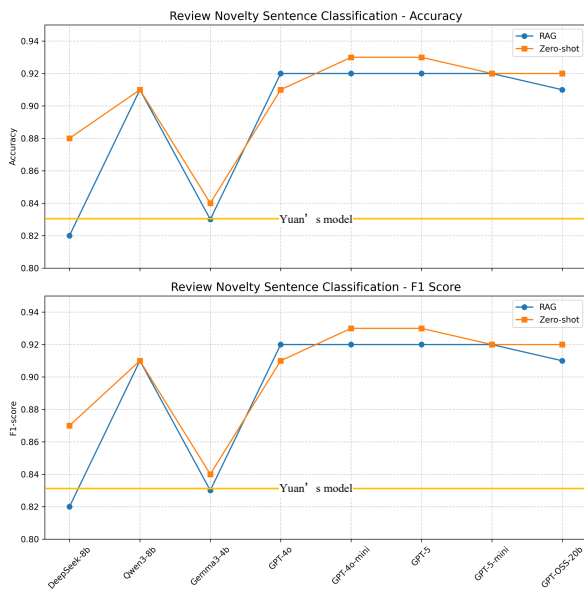


Figure 10: The performance of various LLMs on novelty evaluation extraction under different prompt.

sentative models selected from the main results, including several top-performing models, which sufficiently reflect the overall trends.

Model	Rel.	Cov.	Clarity	DistAcc
SEA-E	3.4337	0.2974	0.6505	0.6712
SEA-S	3.6402	<b>0.3025</b>	<b>0.6656</b>	0.7048
GPT-4o	<b>3.7266</b>	0.2668	0.6582	<b>0.7094</b>
Gemini-2.5-flash	3.4900	0.2608	0.6429	0.6187

Table 7: Results on resource papers.

Tables 6 and 7 report the results for methodological and resource papers, respectively. Models consistently achieve better performance on resource papers than on methodological papers. This is likely because resource papers (e.g., benchmarks) present more explicit and concrete contributions, whereas

You are analyzing whether the following peer review sentence evaluates the **\*\*novelty\*\*** of a paper.

Sentence:  
*{sentence}*

Answer with only 'Yes' or 'No': Does this sentence evaluate novelty?"

Figure 11: The Zero-Shot Prompt for Novelty Evaluations Extraction.

You are analyzing whether the following peer review sentence evaluates the **\*\*novelty\*\*** of a paper.

Sentence:  
*{sentence}*

Here are examples of novelty-related evaluation sentences:  
*{retrieved\_sentence}*

Answer with only 'Yes' or 'No': Does this sentence evaluate novelty?"

Figure 12: The RAG Prompt for Novelty Evaluations Extraction. The retrieved sentences were obtained by calculating the similarity between the query and the sentences related to novelty contained within the ReviewAdvisor (Yuan et al., 2022).

methodological papers often require more nuanced reasoning to assess novelty.

These findings indicate that paper characteristics affect evaluation difficulty, while model rankings remain broadly consistent across categories.

### H.3 Alignment under Reviewer Disagreement

We analyze model behavior under reviewer disagreement by examining whether LLM-generated evaluations align differently with reviewers of varying confidence levels. Specifically, we report a subset of representative models selected from the main results, including several top-performing models, which sufficiently reflect the overall trends.

You are a review analysis assistant. I am providing a set of reviewer comments regarding the novelty of a paper.

Your task:

1. Deduplicate and consolidate comments that are semantically identical or very similar into a single, concise statement.
2. Categorize each consolidated comment into one of the following classes:
  - Positive Novelty Evaluations
  - Neutral Novelty Evaluations
  - Negative Novelty Evaluations
3. Use the exact output format:

Positive Novelty Evaluations

- [Consolidated positive comment 1]
- [Consolidated positive comment 2]

Neutral Novelty Evaluations

- [Consolidated neutral comment 1]

Negative Novelty Evaluations

- [Consolidated negative comment 1]
- [Consolidated negative comment 2]

Here are the evaluations:

*{reviews}*

Figure 13: The Prompt for Structuring Novelty Evaluations based Sentiment.

Model	Rel.	Cov.	Clarity	DistAcc
SEA-E	3.4234	0.2507	0.6495	0.6869
SEA-S	3.6270	0.2445	0.6622	0.7194
GPT-4o	3.6899	0.2233	0.6599	0.6947
Gemini-2.5-flash	3.4647	0.1976	0.6409	0.5962

Table 6: Results on methodological papers.

We select samples with substantial disagreement (confidence gap  $\geq 3$ ) and divide reviews into high-confidence and low-confidence groups. We then compute the semantic similarity between LLM-generated evaluations and each group.

As shown in Table 8, models consistently exhibit higher similarity to high-confidence reviews. This suggests that LLM-generated evaluations tend to align more closely with reviewers who express stronger certainty, rather than behaving arbitrarily under disagreement.

Model	Mode	High	Low	Diff
SEA-S	Zero	0.6761	0.6488	0.0274
SEA-S	Few	0.6579	0.6351	0.0228
SEA-S	RAG	0.6860	0.6474	0.0387
GPT-4o	Zero	0.6632	0.6322	0.0310
GPT-4o	Few	0.6615	0.6404	0.0211
GPT-4o	RAG	0.6964	0.6535	0.0428
SEA-E	Zero	0.6597	0.6456	0.0141
SEA-E	RAG	0.6803	0.6506	0.0297
GPT-5	RAG	0.6537	0.6108	0.0429
Gemini-2.5-flash	RAG	0.6755	0.6338	0.0417

Table 8: Similarity of LLM-generated evaluations to high- and low-confidence reviews under disagreement.

<p><b>review_text</b></p> <ul style="list-style-type: none"> <li>- This paper conducts in-depth research on the relationship between the weight space and loss landscape of fine-tuned language models, providing new ideas and methods for the study of deep learning.</li> <li>- Additionally, they proposed a new method to visualize the similarities between fine-tuned models and discussed the limitations of fine-tuning techniques in deep learning.</li> <li>- The paper offers a unique and intriguing perspective by investigating the relationships between different models in the weight space and loss landscape.</li> </ul> <ul style="list-style-type: none"> <li>- While most previous research has concentrated on understanding individual models trained on specific datasets, this study delves into the connections between various models trained or tested on distinct datasets.</li> </ul> <ul style="list-style-type: none"> <li>- Besides, this paper explores the limitations of fine-tuned models, providing new ideas and methods for the study of deep learning.</li> </ul>	<p>Positive Novelty Evaluations</p> <p>Neutral Novelty Evaluations</p> <p>Negative Novelty Evaluations</p>
<p><b>introduction</b></p> <p>explores the relationship between the weights of different finetuned models and the capabilities they exhibit.</p> <p>We analyze the weight space , where each model is represented by a weight vector <math>\theta \in \mathbb{R}^n</math> .</p> <p>We find that distance characterizes models' knowledge and similarity.</p> <p>Particularly, after finetuning a pretrained model on similar datasets, the resulting models are close to each other in weight space (§2.3). Throughout the paper, we consider 3 granularities (§3.1), showing that (i) models finetuned on the same data are closer to each other than to other models; (ii) models finetuned on the same task also cluster together; and (iii) models finetuned on general language tasks are not spread arbitrarily around the pretrained model, but fall in a constrained region in space.</p> <p>We find that different finetuning runs on the same data tend to converge on similar points in weight space rather than dispersed points.</p>	
<p><b>original_text_1</b></p> <p>This observation highlights a potential limitation or area for further research in the field of fine-tuning large language models. It suggests that simply fine-tuning on similar datasets might not be sufficient to achieve significant improvements, warranting further investigation into more sophisticated fine-tuning strategies or model architectures.</p> <p>The finding that models remain close in weight space after fine-tuning on similar datasets could be a valuable insight for practitioners, indicating that the pre-trained knowledge is robust and transferable, but also suggesting that more targeted approaches might be necessary for specific downstream tasks.</p> <p>This observation could serve as a starting point for developing methods to better leverage pre-trained models for fine-tuning, potentially leading to more efficient and effective transfer learning techniques.</p>	<p>Positive:</p> <p>Neutral:</p> <p>Negative:</p>
<p><b>original_text_2</b></p> <p>The paper introduces a novel approach to analyzing the relationship between model weights and their capabilities, which could provide new insights into model performance and convergence behavior during fine-tuning.</p> <p>The analysis of weight space as a means to characterize models' knowledge and similarity is innovative, potentially offering a new perspective on model comparisons and evaluations.</p> <p>The granularity of analysis across different datasets and tasks is a significant contribution, highlighting the nuanced ways in which models converge in weight space.</p> <p>The paper's approach to analyzing model weights and their implications for model capabilities is not entirely new, as similar studies have been conducted in the field. However, the specific focus on weight space and its implications for model behavior is a refinement of existing knowledge.</p> <p>The paper's findings on the clustering of models in weight space based on their fine-tuning tasks and datasets are expected given the current understanding of model convergence. While the paper provides a detailed analysis, it does not fundamentally challenge existing theories.</p> <p>The paper does not introduce any groundbreaking methodologies or theories that significantly advance the field beyond current understanding. The analysis primarily reaffirms existing knowledge about model convergence in weight space during fine-tuning.</p> <p>The paper's focus on weight space as a means to evaluate model capabilities is not particularly novel, as similar approaches have been explored in previous studies. The paper does not offer a significant departure from existing methodologies or provide new insights that substantially advance the field.</p> <p>The paper's contribution to the field is limited by its reliance on existing models and datasets, which may not sufficiently challenge or expand current understandings of model behavior during fine-tuning.</p>	<p>Positive:</p> <p>Neutral:</p> <p>Negative:</p>

Figure 14: An Example for Human Evaluation.

## Human Evaluation Guideline

You will be given two model-generated novelty evaluations (e.g., Model A and Model B) for the same novelty description. Notable, review\_text, introduction, original\_text\_1 and original\_text\_2 are human reviewer evaluation, novelty description, model A's output and model B's output.

Use the following four dimensions **as reference only**.

Your final task is simply to decide **which model produced the better overall evaluation**.

### Reference Dimensions

#### 1. Relevance

Determine whether the model's evaluation genuinely addresses the **novelty** of the paper rather than drifting into unrelated topics (e.g., dataset details, experimental results, writing quality, or background information).

#### 2. Correctness

Determine whether the model's positive/neutral/negative stance aligns with the expert reviewer's actual novelty evaluation.

#### 3. Coverage

Determine whether the model captures **multiple novelty points** mentioned by the expert, rather than focusing on only one point or missing important aspects.

#### 4. Clarity

Determine whether the evaluation is **clear, specific, and meaningful**, rather than vague, generic, or lacking detail.

### Final Output (Choose One)

Please choose **only one** of the following:

- **Model A is better**
- **Model B is better**
- **Both are comparable**

Figure 15: Guideline of Human Evaluation.

You are a peer-review expert specializing in novelty evaluation of academic papers.  
I will provide sentences from the introduction of a paper that describe its novelty.  
Your task is to evaluate the novelty of the paper and present your evaluation strictly in the following format:

Positive Novelty Evaluations

-

Neutral Novelty Evaluations

-

Negative Novelty Evaluations

-

Rules:

- If the work is highly innovative, add an item under "Positive".
- If the work shows some improvement but is only moderately new, add it under "Neutral".
- If the work lacks substantial novelty, add it under "Negative".
- Each category can contain multiple evaluations or remain empty (with only a dash).
- Keep evaluations concise and academic in tone.

Input:

|"|"|"

{sentences}

|"|"|"

Figure 16: The zero shot prompt for different LLMs on NovBench.

You are a peer-review expert specializing in novelty evaluation of academic papers.  
I will provide sentences from the introduction of a paper that describe its novelty.  
Your task is to evaluate the novelty of the paper and present your evaluation strictly in the following format:

Positive Novelty Evaluations  
-

Neutral Novelty Evaluations  
-

Negative Novelty Evaluations  
-

Rules:

- If the work is highly innovative, add an item under "Positive".
- If the work shows some improvement but is only moderately new, add it under "Neutral".
- If the work lacks substantial novelty, add it under "Negative".
- Each category can contain multiple evaluations or remain empty (with only a dash).
- Keep evaluations concise and academic in tone.

### Example 1  
Input:  
*{input\_1}*  
Output:  
*{output\_1}*

### Example 2  
Input:  
*{input\_2}*  
Output:  
*{output\_2}*

### Now it's your turn:  
Input:  
*{sentences}*

Figure 17: The few shot prompt for different LLMs on NovBench.

You are a peer-review expert specializing in novelty evaluation of academic papers.  
I will provide:

1. Sentences from the introduction of a paper that describe its novelty.
2. Titles and abstracts of 5 related papers retrieved from a local literature database.

Your task is to evaluate the novelty of the given paper by comparing it against the retrieved references.  
Output your evaluation strictly in the following format:

Positive Novelty Evaluations  
-

Neutral Novelty Evaluations  
-

Negative Novelty Evaluations  
-

Rules:

- If the contribution introduces a fundamentally new idea not present in the retrieved references, write it under "Positive".
- If the contribution is similar to but not exactly the same as retrieved references, write it under "Neutral".
- If the claimed contribution overlaps strongly with retrieved references and lacks originality, write it under "Negative".
- Each section can be empty if not applicable (just keep the dash).
- Keep evaluations concise, academic in tone, and grounded in the retrieved references.

[Input Sentences]  
*{sentences}*

[Retrieved References]

1. Title: *{retrieved\_title\_1}*  
Abstract: *{retrieved\_abstract\_1}*
2. Title: *{retrieved\_title\_2}*  
Abstract: *{retrieved\_abstract\_2}*
3. Title: *{retrieved\_title\_3}*  
Abstract: *{retrieved\_abstract\_3}*
4. Title: *{retrieved\_title\_4}*  
Abstract: *{retrieved\_abstract\_4}*
5. Title: *{retrieved\_title\_5}*  
Abstract: *{retrieved\_abstract\_5}*

Figure 18: The RAG prompt for different LLMs on NovBench.

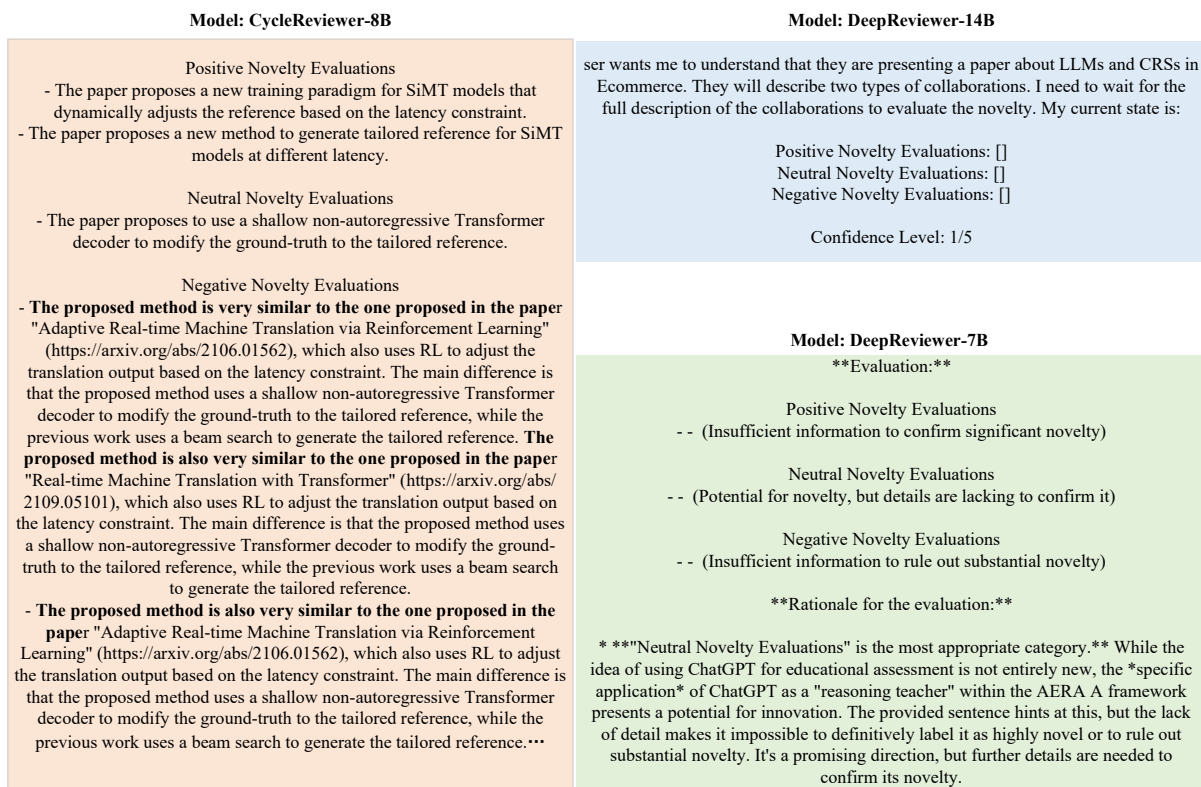


Figure 19: Examples of Instruction-Following Failures by other Specialized Models.

**Paper Novelty Description:**

"Given an utterance  $x_1$  from pre-training corpus, we generate a pseudo intent name  $y_{pseudo_1}$  using labels from the intent role labeling (IRL) tagger.",  
 "Our PIE model is then optimized by pulling the gold utterance  $x_{gold_1}$ , gold intent  $y_1$ , and pseudo intent  $y_{pseudo_1}$  close to the input utterance  $x_1$  in the embedding space.",  
 "In this paper, we propose a novel pre-training method for zero and few-shot IC tasks (Figure 1).",  
 "Once we obtain the IRL predictions, we convert them to the pseudo intent names of query utterances and use them to pre-train the encoder in a contrastive learning fashion.",  
 "This intent-aware contrastive learning aims to not only align utterances with their pseudo intent names in the semantic embedding space, but also to encourage the encoder to pay attention to the intent-relevant spans that are important for distinguishing intents.",  
 "To the best of our knowledge, this work is the first to extract key information from utterances and use it as pseudo labels for pre-training intent-aware text encoders."

**SEA-S output:**

## Positive Novelty Evaluations

- The paper introduces a novel pre-training method for zero- and few-shot intent classification tasks, which is a significant advancement in the field of natural language processing.
- The use of intent-aware contrastive learning to align utterances with their pseudo intent names in the semantic embedding space is a novel approach that could potentially enhance the performance of text encoders in distinguishing intents.
- The paper is the first to extract key information from utterances and use it as pseudo labels for pre-training intent-aware text encoders, which represents a significant innovation in the application of pseudo labels in NLP tasks.

## Neutral Novelty Evaluations

- The paper describes a method of optimizing a pre-training model by aligning utterances with their pseudo intent names in the embedding space, which is a known approach with some modifications.
- The use of intent role labeling (IRL) tagger to generate pseudo intent names is a known technique, but the paper's application of this method in a novel way for pre-training could be seen as an improvement rather than a complete novelty.

## Negative Novelty Evaluations

- The paper does not introduce any fundamentally new techniques or approaches to the field of natural language processing, as it primarily builds upon existing methods such as intent role labeling and contrastive learning.
- The use of pseudo labels for pre-training is not a new concept, and the paper does not provide a significant advancement in the application of this technique beyond what has already been explored in previous works.
- The paper's claims of novelty are somewhat overstated, as the core ideas presented are not significantly different from existing methods in the field.

**GPT-4o output:**

## Positive Novelty Evaluations

- The paper proposes a novel pre-training method specifically designed for zero- and few-shot intent classification (IC) tasks, which represents an innovative direction in pre-training techniques for this domain.
- To the best of the authors' knowledge, the work is the first to extract key information from utterances and use it as pseudo labels for pre-training intent-aware text encoders, indicating significant novelty.

## Neutral Novelty Evaluations

- The use of intent-aware contrastive learning to align utterances with pseudo intent names while encouraging the encoder to focus on intent-relevant spans introduces a moderately new application of contrastive learning in IC tasks.
- The approach of using predictions from an intent role labeling (IRL) tagger to generate pseudo intent names for pre-training contributes some improvement to pre-trained encoder performance, but this may not represent a fundamentally new concept in natural language understanding.

## Negative Novelty Evaluations

-

**Gold output:**

## Positive Novelty Evaluations

- The paper presents a novel PIE method that enhances an encoder's ability to represent intent.
- The proposed method demonstrates significant improvement in the representational ability of LLMs used as encoders in task-oriented dialogue tasks.
- This paper introduces a new intent-aware contrastive learning objective that improves intent classification performance compared to baselines.
- The paper introduces an innovative algorithm for generating pseudo intent names from utterances across multiple dialogue datasets.
- The idea of assigning roles to words using IRL for creating pseudo intents to pre-train the encoder is innovative and interesting.
- The authors introduce a unique pre-training approach for text encoders, leveraging contrastive learning combined with intent pseudo-labels.
- The paper proposes a novel and inspiring method for the intent classification task, achieving state-of-the-art performance.

## Neutral Novelty Evaluations

- None

## Negative Novelty Evaluations

- The proposed method seems to be very similar to already published works.

Figure 20: Case Outputs of SEA-S and GPT-4o Compared with Novelty Descriptions from the Paper Introduction and Human Reviewer Evaluations.

**Paper Novelty Description:**

"To answer the question, we first define an evaluation suite of physical concepts covering visual and embodied concepts.",  
 "We compose a Visual and Embodied Concepts evaluation benchmark VEC , with examples shown in Table 1.",  
 "With the benchmark, we examine a wide range of LMs.",  
 "We design two forms of probing tasks.",  
 "The former (Color, Shape and Material) asks models to make a choice between two tail options given the head object.",  
 "The latter (Size, Height, and all embodied concepts) requires LMs to judge whether the relation is valid given the head and the tail.",  
 "Our zero (few)-shot results on the VEC benchmark show that: (i) Moderate-sized LMs such as BERT and RoBERTa exhibit a random-level understanding of both visual and embodied concepts.",  
 "(ii) A decent visual understanding of specific concepts emerges as LMs scale up, while they still struggle to understand the embodied knowledge with performance slightly better than random guessing.",  
 "(iii) Image-grounded caption text-only pre-training, instruction tuning, and visual supervision could provide performance gain regarding visual concepts, yet only the last one enhances the understanding of embodied knowledge of LMs.",  
 "We further investigate the source of embodied knowledge in VLMs.",  
 "A case study demonstrates that embodied knowledge in the VLM of CLIP is potentially rooted in the rich semantics of image representations.",  
 "We thus devise a knowledge distillation method to transfer the learned embodied knowledge in VLMs into LMs, resulting in an",  
 "average accuracy gain of 3.38%."

**SEA-S output:**

## Positive Novelty Evaluations

- The paper introduces a novel evaluation suite, VEC, specifically designed for assessing visual and embodied concepts in large language models (LMs), which is a significant contribution to the field.
- The paper presents innovative probing tasks that differentiate between visual and embodied concepts, providing a nuanced evaluation of LMs' understanding of these areas.
- The use of a knowledge distillation method to transfer embodied knowledge from vision-language models (VLMs) to LMs represents a novel approach to enhancing LMs' understanding of embodied concepts.

## Neutral Novelty Evaluations

- The paper's zero-shot results on the VEC benchmark show that while LMs exhibit some improvement in understanding visual concepts as they scale up, their understanding of embodied knowledge remains limited, which is a known issue in the field.
- The paper's investigation into the source of embodied knowledge in VLMs is a common area of research, but the specific focus on CLIP's rich semantics in image representations adds a slight novelty to the discussion.

## Negative Novelty Evaluations

- The paper does not introduce any fundamentally new theoretical frameworks or methodologies for understanding or enhancing LMs' understanding of visual and embodied concepts.
- The results presented, such as the average accuracy gain of 3.38%, are modest and do not significantly advance the state of the art in this area.
- The paper's focus on specific LMs like BERT and RoBERTa does not provide a comprehensive evaluation across a broader range of models, which could limit the generalizability of the findings.

**GPT-4o output:**

## Positive Novelty Evaluations

- The definition of a comprehensive evaluation suite (Visual and Embodied Concepts evaluation benchmark, VEC) is a significant contribution and valuable asset for assessing a wide range of language models and their understanding.
- The introduction of probing tasks specifically addressing visual (Color, Shape, Material) and embodied (Size, Height, Relations) concepts to test language models suggests a highly innovative approach to evaluating knowledge.
- The use of knowledge distillation to transfer embodied knowledge from Vision-Language Models (VLMs) to other Language Models (LMs), resulting in measurable accuracy gains, is a novel methodological advancement.

## Neutral Novelty Evaluations

- The identification of scaling effects in large language models demonstrating emerging understanding of visual concepts, while embodied knowledge remains limited, offers valuable insights but builds upon established findings.
- The performance improvement achieved through techniques like pretraining and instruction tuning is insightful but reflects incremental advancements rather than groundbreaking novelties.

## Negative Novelty Evaluations

- Examining existing Vision-Language Models' (e.g., CLIP) source of embodied knowledge, while useful, does not introduce substantially new methods or paradigms.

**Gold output:**

## Positive Novelty Evaluations

- The paper introduces a novel benchmark evaluation suite (VEC) that covers a wide range of physical concepts, including both visual and embodied attributes, filling a gap in the field by providing a standardized framework to evaluate language models' understanding of physical attributes.
- The finding that we can effectively distill embodied knowledge from VLMs to less capable LLMs is quite inspiring.
- Probing LMs with physical concepts is a novel and interesting topic that provides insights on how to further improve and align current LMs to the real world.
- The paper is well-written and easy to follow.

## Neutral Novelty Evaluations

- The main contributions of this paper include conducting in-depth analysis with different types of LLMs and VLMs on their understanding of visual concepts and embodied concepts, providing insights on what physical concepts are well-captured and struggled by LLMs, and showing evidence that vision-language pretraining helps learn embodied knowledge.

## Negative Novelty Evaluations

- Compared to GPT-3 era models, newer models like ChatGPT contain much richer physical commonsense knowledge and can perform complex reasoning about physical interactions, suggesting that the paper's findings may be less novel in the context of recent advancements.

Figure 21: Case Outputs of SEA-S and GPT-4o Compared with Novelty Descriptions from the Paper Introduction and Human Reviewer Evaluations.

<p><b>Paper Novelty Description:</b></p> <p>"In this work, we address some shortcomings of one of the first loss formulations proposed for learning deep neural networks under LLP setup by (Ardehaly and Culotta, 2017), termed as DLLP method.",</p> <p>"In Section 3, we highlight certain properties of DLLP objective that can be highly undesirable for training deep networks.",</p> <p>"Motivated by this, we propose a novel objective function that is a parametrization of the Total Variation Distance (TVD), which itself is a lower bound to the KL via the Pinsker's inequality .",</p> <p>"Our formulation enjoys more functional flexibility because of the introduced parameter while retaining the outlier robustness property of the TVD.",</p> <p>"We also discuss some theoretical results for the proposed novel formulation.",</p> <p>"Lastly, we combine our formulation with an auxiliary self-supervised objective that greatly aids in representation learning during the fine-tuning stage of the large scale NLP models experimented with.",</p> <p>"Experimentally, we first demonstrate that the proposed formulation is indeed better and align with the theoretical motivation provided.",</p> <p>"In the main results, we demonstrate that our formulation achieves better results compared to the baselines in almost 87% of the 20 extensive configurations across 4 widely used models and 5 datasets.",</p> <p>"We observe up to 40% improvement in weighted precision metric on the BERT model.",</p> <p>"To summarize, we have the following contributions: (i) A novel loss formulation that addresses the shortcomings of the previous work with supporting theoretical and empirical results.",</p> <p>"(ii) One of the preliminary works discussing the application of LLP to natural language tasks.",</p> <p>"(iii) Strong empirical results demonstrating that our method outperforms the baselines in most of the configurations."</p>
<p><b>SEA-S output:</b></p> <p>Positive Novelty Evaluations</p> <ul style="list-style-type: none"> <li>- The paper introduces a novel loss formulation that addresses the shortcomings of previous methods, specifically the DLLP method, by proposing a new objective function based on the Total Variation Distance (TVD), which is a lower bound to the KL via Pinsker's inequality.</li> <li>- The proposed formulation offers more functional flexibility through the introduction of a new parameter while maintaining the outlier robustness property of the TVD.</li> <li>- The paper presents theoretical results to support the novel formulation, which is a significant contribution to the field.</li> <li>- The combination of the proposed formulation with an auxiliary self-supervised objective enhances representation learning during the fine-tuning stage, which is a novel approach.</li> <li>- The empirical results demonstrate that the proposed method outperforms baselines in most configurations, showing up to 40% improvement in weighted precision metric on the BERT model, which is a substantial improvement.</li> </ul> <p>Neutral Novelty Evaluations</p> <ul style="list-style-type: none"> <li>- The paper discusses the application of the LLP to natural language tasks, which is a preliminary but relevant area of research.</li> </ul> <p>Negative Novelty Evaluations</p> <ul style="list-style-type: none"> <li>- The paper does not provide a clear differentiation from existing methods in terms of the fundamental approach to the problem, as it still relies on the LLP setup, which has been previously explored extensively.</li> <li>- The novelty of the proposed formulation might be limited as it is a modification of existing methods rather than a fundamentally new approach.</li> <li>- The paper does not sufficiently discuss the limitations or potential drawbacks of the proposed method, which could undermine the perceived novelty and impact of the work.</li> </ul>
<p><b>GPT-4o output:</b></p> <p>### Positive Novelty Evaluations</p> <ul style="list-style-type: none"> <li>- The introduction of a novel loss formulation based on a parametrization of the Total Variation Distance (TVD), which retains outlier robustness while offering functional flexibility.</li> <li>- Theoretical contributions supporting the proposed novel loss formulation.</li> <li>- Empirical demonstration of the novel formulation's effectiveness, with strong results across diverse configurations, models, and datasets, including up to a 40% improvement in weighted precision on the BERT model.</li> <li>- Highlighting and addressing undesirable properties of the DLLP objective, making strides in the LLP domain.</li> <li>- Preliminary work in applying the LLP paradigm to natural language processing tasks.</li> </ul> <p>### Neutral Novelty Evaluations</p> <ul style="list-style-type: none"> <li>- Combination of the proposed formulation with an auxiliary self-supervised objective for enhanced representation learning during the fine-tuning stage of large-scale NLP models.</li> </ul> <p>### Negative Novelty Evaluations</p> <ul style="list-style-type: none"> <li>- None.</li> </ul>
<p><b>Gold output:</b></p> <p>Positive Novelty Evaluations</p> <ul style="list-style-type: none"> <li>- The paper proposes a new loss function for training deep learning classifiers in the Learning under Label Proportions (LLP) setup, supported by a theoretical background and good experimental results.</li> <li>- The application of LLM to natural language tasks is new and of some interest, and the authors claim it to be one of the preliminary works discussing the application of LLP to NLP tasks.</li> <li>- This paper introduces a robust objective function for the learning from label proportions (LLP) setting, applying the total variation distance as a tight lower bound of the KL divergence.</li> </ul> <p>Neutral Novelty Evaluations</p> <ul style="list-style-type: none"> <li>- The paper is relatively clearly written and includes a simple ablation study.</li> </ul> <p>Negative Novelty Evaluations</p> <ul style="list-style-type: none"> <li>- None.</li> </ul>

Figure 22: Case Outputs of SEA-S and GPT-4o Compared with Novelty Descriptions from the Paper Introduction and Human Reviewer Evaluations.

**Paper Novelty Description:**

"Drawing on these insights, we propose to integrate word-level image information into the UNMT process to mitigate the issue of lexical confusion.",  
 "It's worth noting that we don't use sentence-level images like some multi-modal UNMT approaches (Su et al.).",  
 "Instead, we focused on word-level images since they are more effective and flexible in addressing the challenge of lexical confusion and are easier to collect.",  
 "Our approach involves augmenting the encoded images to their corresponding words, modifying the corresponding position and language encoding information, and utilizing a visible matrix to isolate the impact of images on other words.",  
 "We conduct experiments utilizing over 300,000 self-collected images, and further analysis demonstrates that we have effectively mitigated lexical confusion to a certain extent.",  
 "Furthermore, we achieve new state-of-the-art results for the UNMT task in certain language directions on the Multi30k dataset."

**SEA-S output:****Positive Novelty Evaluations**

- The paper introduces a novel approach by integrating word-level image information into the UNMT process, which is a departure from traditional sentence-level image integration methods.
- The use of word-level images is highlighted as more effective and flexible in addressing lexical confusion, which is a significant improvement over existing methods.
- The paper demonstrates a substantial collection of over 300,000 self-collected images, which is a significant dataset for this type of research.
- The proposed method achieves new state-of-the-art results for the UNMT task in certain language directions on the Multi30k dataset, indicating a significant advancement in the field.

**Neutral Novelty Evaluations**

- The paper's focus on word-level images is noted as a different approach from some multi-modal UNMT approaches, but the novelty of this approach may not be entirely new as similar ideas have been explored in the literature.
- The paper's method of augmenting encoded images to their corresponding words and utilizing a visible matrix to isolate the impact of images on other words is a modification of existing techniques, rather than a completely new approach.

**Negative Novelty Evaluations**

- The paper does not provide a clear differentiation from existing methods in terms of its theoretical contributions, as it primarily focuses on the practical application of integrating word-level images into UNMT without significant theoretical advancements.
- The paper's novelty is limited by its reliance on existing methodologies and datasets, with no significant innovation in the underlying technology or data sources used.
- The paper's claims of mitigating lexical confusion are not substantiated by a rigorous comparative analysis with other state-of-the-art methods, which undermines the strength of its novelty claims.

**GPT-4o output:****\*\*Positive Novelty Evaluations\*\***

- The integration of word-level image information into the UNMT process to address lexical confusion represents a distinct and innovative approach not commonly explored in prior multi-modal UNMT research.
- The proposal to use word-level images, rather than sentence-level images, introduces increased flexibility and practicality in data collection, representing an improvement over existing methodologies.
- The development of a unique mechanism involving augmented encoded images, adjusted position/language encoding information, and a visible matrix demonstrates a novel methodological contribution to the UNMT framework.

**\*\*Neutral Novelty Evaluations\*\***

- The demonstration of new state-of-the-art results for specific language directions on the Multi30k dataset suggests some incremental methodological advancements in UNMT.

**\*\*Negative Novelty Evaluations\*\***

- The mitigation of lexical confusion "to a certain extent" as reported may reflect an optimization rather than a fundamentally new breakthrough in addressing this issue.

**Gold output:****Positive Novelty Evaluations**

- This paper introduces a novel multi-modal unsupervised neural machine translation method based on self-collected word-level images, addressing lexical confusion and providing a new perspective on leveraging visual information for NLP.
- The proposed method offers a valuable dataset that can facilitate multimodal NLP research and development, potentially leading to new tasks or challenges that combine vision and language.

**Neutral Novelty Evaluations**

- The authors leverage search engines to extract images for training with denoising autoencoder loss and backtranslation based loss.

**Negative Novelty Evaluations**

- None.

Figure 23: Case Outputs of SEA-S and GPT-4o Compared with Novelty Descriptions from the Paper Introduction and Human Reviewer Evaluations.

**Paper Novelty Description:**

"In this paper, we propose CHAP, a novel translation-based AMR parser distinguished by three innovations.",  
 "Firstly, we introduce new target forms of AMR parsing.",  
 "1.c-e, we use multiple layers to capture different semantics, such that each layer is simple and concise.",  
 "Particularly, the base layer, which encapsulates all meanings except for coreferences (or reentrancies), is a tree-structured representation, enabling more convenient structure modeling than the graph structure of AMR.",  
 "Meanwhile, coreferences are presented through pointers, circumventing several shortcomings associated with the variable-based coreference representation (See Sec.",  
 "Secondly, we propose Causal Hierarchical Attention (CHA), the core mechanism of our incremental structure modeling, inspired by Transformer Grammars (Sartran et al., 2022).",  
 "CHA describes a procedure of continuously composing child nodes to their parent nodes and encoding new nodes with all uncomposed nodes, as illustrated in Fig.",  
 "(c)-(d) are our proposed representations.",  
 "(d) is for \u0302 double(c) (Fig.",  
 "like the causal attention in translation-based models, which allows a token to interact with all its preceding tokens, CHA incorporates a strong inductive bias of recursion, composition, and graph topology.",  
 "Thirdly, deriving from transition-based AMR parsers (Zhou et al., 2021a,b), we introduce a pointer encoder for encoding histories and a pointer net for predicting coreferences, which is proven to be an effective solution for generalizing to a variable-size output space (Vinyals et al., 2015; See et al., 2017).",  
 "We propose various alternative modeling options of CHA and strategies for integrating CHA with existing pretrained seq2seq models and investigate them via extensive experiments.",  
 "Ultimately, our model CHAP achieves superior performance on two in-distribution and three out-of-distribution benchmarks."

**SEA-S output:**

## Positive Novelty Evaluations

- The paper introduces a novel translation-based AMR parser, CHAP, which employs multiple layers to capture different semantics, each layer being simple and concise, and a tree-structured base layer for more convenient structure modeling compared to traditional graph structures.
- The introduction of Causal Hierarchical Attention (CHA) as a core mechanism for incremental structure modeling is a significant innovation, inspired by Transformer Grammars and incorporating strong inductive biases of recursion, composition, and graph topology.
- The paper proposes a novel pointer encoder for encoding histories and a pointer net for predicting coreferences, which is an effective solution for generalizing to a variable-size output space, enhancing the model's adaptability and scalability.
- The paper explores various alternative modeling options of CHA and integrates it with existing pretrained seq2seq models, demonstrating a comprehensive approach to model adaptation and integration.

## Neutral Novelty Evaluations

- The paper's novelty in using multiple layers for semantic capture and a tree-structured base layer for AMR parsing is somewhat similar to existing approaches in the field, albeit with some improvements in simplicity and structure modeling.

## Negative Novelty Evaluations

- The paper does not provide a clear differentiation from existing transition-based AMR parsers, particularly in the use of pointer encoders and nets for coreference prediction, which have been previously explored in similar models.
- The novelty of the proposed CHA mechanism, while inspired by Transformer Grammars, does not significantly deviate from existing causal attention mechanisms in translation-based models, lacking substantial novelty in its core functionality.
- The paper's claims of superior performance on benchmarks are not substantiated with a detailed comparison against existing state-of-the-art models, which could undermine the perceived novelty and impact of the proposed methods.

**GPT-4o output:**

## Positive Novelty Evaluations

- The introduction of new target forms of AMR parsing offers significant potential for advancing parsing methodologies.
- The use of a tree-structured representation for the base layer, accompanied by pointers for coreferences, simplifies AMR structure modeling and addresses shortcomings of traditional variable-based coreference approaches.
- The proposal of Causal Hierarchical Attention (CHA) inspired by Transformer Grammars introduces a unique mechanism incorporating inductive biases of recursion, composition, and graph topology.
- The introduction of a pointer encoder and pointer net for modeling coreferences presents an effective solution to generalizing variable-size output spaces.
- The integration of CHA with pretrained seq2seq models and the extensive investigation of alternative modeling options demonstrate advancement in combining incremental modeling techniques with established frameworks.

## Neutral Novelty Evaluations

- Some innovations, such as CHA procedures for child-to-parent node composition and encoding, although beneficial, are moderately new and extend existing concepts in attention-based modeling.

## Negative Novelty Evaluations

- The emphasis on extensive experimental validation and benchmark performance, while important, does not inherently indicate substantial methodological novelty.

**Gold output:**

## Positive Novelty Evaluations

- The authors contribute new AMR representations/linearizations involving pointers to dismiss variables and encode parent-child relations, along with an extension of the decoding process using hierarchical attention.
- This paper introduces a new method for AMR parsing called CHAP, which adds graph structure to the decoder architecture while maintaining compatibility with pretrained transformer decoders.
- The new decoder architecture features significant modifications, including a hierarchical attention mechanism based on Transformer Grammars.
- Overall, this paper proposes a novel architecture for AMR parsing that outperforms recent baselines, except for a concurrent project called LeakDistill.
- The method itself seems to be an innovative improvement over prior work.
- This paper presents a novel AMR parser.

## Neutral Novelty Evaluations

- No other applications beyond AMR parsing are discussed.

## Negative Novelty Evaluations

- [No negative comments found.]

Figure 24: Case Outputs of SEA-S and GPT-4o Compared with Novelty Descriptions from the Paper Introduction and Human Reviewer Evaluations.