Evaluating and Enhancing Large Language Models for Novelty Assessment in Scholarly Publications

Ethan Lin *
Santa Clara University
Santa Clara, CA, USA
enlin@scu.edu

Zhiyuan Peng *
Santa Clara University
Santa Clara, CA, USA
zpeng@scu.edu

Yi Fang
Santa Clara University
Santa Clara, CA, USA
yfang@scu.edu

Abstract

Recent studies have evaluated creativity, where novelty is an important aspect, of large language models (LLMs) primarily from a semantic perspective, using benchmarks from cognitive science. However, assessing the novelty in scholarly publications, a critical facet of evaluating LLMs as scientific discovery assistants, remains underexplored, despite its potential to accelerate research cycles and prioritize high-impact contributions in scientific workflows. We introduce SchNovel [1], a benchmark to evaluate LLMs' ability to assess novelty in scholarly papers, a task central to streamlining the discovery pipeline. SchNovel consists of 15000 pairs of papers across six fields sampled from the arXiv dataset with publication dates spanning 2 to 10 years apart. In each pair, the more recently published paper is assumed to be more novel. Additionally, we propose RAG-Novelty, a retrieval-augmented method that mirrors human peer review by grounding novelty assessment in the retrieved context. Extensive experiments provide insights into the capabilities of different LLMs to assess novelty and demonstrate that RAG-Novelty outperforms recent baseline models, highlighting LLMs' promise as tools for automating novelty detection in scientific workflows.

1 Introduction

AI-driven scientific discovery systems, such as autonomous lab platforms like Coscientist (Boiko et al., 2023), promise to accelerate research by synthesizing insights from vast literature. A critical bottleneck, however, lies in identifying which papers introduce truly novel concepts, a capability essential for prioritizing experiments, avoiding redundant work, and guiding discovery pipelines. While large language models (LLMs) are increasingly deployed to analyze scientific texts, their ability to detect scholarly novelty, particularly in evolving research contexts, remains unproven. This gap persists despite LLMs' remarkable proficiency in tasks requiring creativity, traditionally defined as producing ideas that are both novel and effective (Runco and Jaeger, 2012). LLMs now solve open-domain problems, write code, and even generate research ideas rivaling human experts' novelty (Si et al., 2024). Yet their capacity to systematically assess novelty in scholarly publications, where contributions build incrementally on prior work, remains underexplored.

Recent studies evaluating the generative creativity of LLMs have yielded inconsistent conclusions. Orwig et al. (2024) concluded that GPT-4 (OpenAI, 2023) generates stories that are comparable to those written by humans in terms of creativity. Similarly, Pépin et al. (2024) found that LLMs can even surpass humans in specific creative tasks, such as divergent association and creative writing. However, Anderson et al. (2024) argued that AI-based creativity support tools (CSTs) like ChatGPT are not yet well-suited to fostering truly original ideas, as they can lead to the homogenization of human creativity. Chakrabarty et al. (2024) observed that LLM-generated stories pass the Torrance Test for Creative Writing (TTCW) tests 3 to 10 times less frequently than those written by professionals. Additionally, Chakrabarty et al. (2023) pointed out that LLMs often rely on cliches, produce text lacking nuance, and frequently resort to overly moralistic and predictable endings in stories. These discrepancies can be attributed to using different evaluation benchmarks and metrics, highlighting the lack of widely accepted standards for accessing LLM creativity in domain-specific contexts like scientific discovery.

The evaluation benchmarks used in current studies are primarily derived from cognitive science, such as the Torrance Tests of Creative Thinking (TTCT) (Lissitz and Willhoft, 1985), Alternative

*These authors contributed equally to this work.

[1]The SchNovel dataset and RAG-Novelty code are available at: https://github.com/ethannlin/SchNovel

Use Task (AUT) (Guilford, 1964), and the Runco Creativity Assessment Battery (rCAB) (Runco, 2011). These benchmarks focus on assessing semantic creativity by tasks like generating responses to pictures or listing as many uses as possible for a common object. Corresponding metrics include fluency, flexibility, originality, and elaboration. However, these metrics primarily assess semantic novelty, which does not fully capture the kind of novelty emphasized in scholarly research. Novelty in scholarly work is especially critical, as each paper undergoes rigorous peer review, particularly in high-prestige venues. Novel papers typically build upon existing research while introducing new ideas, methods, or insights, making novelty assessment heavily dependent on current and past trends in research.

While LLMs have shown great capability in generating text and mimicking human reasoning, their ability to assess novelty in scholarly publications remains largely unexamined. To address this gap, we present a scholarly novelty benchmark (SchNovel) to evaluate LLMs' capability of assessing novelty in scholarly papers. Specifically, we leverage the arXiv dataset to create a collection of 15,000 paper pairs. In each pair, we assume that the more recently published paper is more novel. Papers are selected across six categories, with publication dates spaced by gaps ranging from 2 to 10 years between the paired papers. We evaluate various LLMs on their ability to assess novelty and report their accuracy.

To further improve novelty assessment, we propose RAG-Novelty, a retrieval-augmented generation method. This method assumes that more novel papers will retrieve more recently published works, enhancing the novelty prediction. Our extensive experiments demonstrate that RAG-Novelty outperforms recent baseline models in assessing novelty in scholarly papers. Our key contributions include:

- We release the first benchmark, SchNovel, specifically designed to evaluate LLMs' capability in assessing novelty within scholarly publications.
- We conduct comprehensive experiments to explore how variations in categories, starting years, and year gaps affect LLMs' ability to assess paper novelty.
- We propose a novel method, RAG-Novelty, to enhance LLMs' performance in assessing paper novelty.

2 Related Work

2.1 Existing Benchmarks

TTCT (Lissitz and Willhoft, 1985) is a commercially protected assessment tool consisting of six tasks: 1) asking a question about a picture; 2) guessing the cause of the action depicted in the image; 3) predicting the consequences of the action described in the image; 4) improving a product described in 2-3 sentences in the most interesting and unusual way; 5) suggesting interesting and unconventional uses for a given item; and 6) imagining what would happen if an improbable situation were to occur. Both AUT (Guilford, 1964) and rCAB (Runco, 2011) ask participants to generate as many uses as possible for a common object. The Remote Associates Test (RAT) (Mednick and Halpern, 1968) presents participants with three seemingly unrelated words and asks them to find a fourth word that connects all three. The Consensual Assessment Technique (CAT) (Amabile, 1982) evaluates creative products, such as stories, poetry, dramatic performances, and musical compositions, using a panel of domain experts. The Wallach-Kogan Creativity Tests (WCT) (Brody, 1966) consist of the AUT, Instances Test, and Similarities Test. The Scholarly Creativity Test (SCT) (Hu and Adey, 2002) measures scholarly creativity and process skills. The Divergent Association Task (DAT) (Olson et al., 2021) asks participants to name unrelated nouns and calculates the pairwise semantic distance between them. However, all these existing cognitive science benchmarks are not suited for evaluating LLMs' capability to assess novelty in scholarly publications, a gap our proposed benchmark addresses.

2.2 Creativity and Novelty Assessment

Traditional general novelty assessment methods use pre-defined metrics like the similarity to existing methods (Just et al., 2024) and the diversity of references (Shibayama et al., 2021) to score the novelty of a method or scholarly paper. To assess LLMs' capability of generating or assessing creativity and novelty, current studies employ different prompt strategies to interact with LLMs and collect responses for evaluation. Guzik et al. (2023) utilized a basic prompt to evaluate GPT-4 on the TTCT benchmark. Mehrotra et al. (2024) applied associative thinking (Mednick, 1962) in prompts designed for specific tasks like product design and marketing. Zhao et al. (2024) analyzed LLMs' responses to an expanded TTCT benchmark, applying diverse prompts, including basic prompts, instructive prompts, post-instructive prompts, and Chain of Thought (CoT) prompts. Stevenson et al. (2022) demonstrates that defining the role of LLMs as "scientist" can improve performance. Summers-Stay et al. (2023) improves the basic prompt method used in (Stevenson et al., 2022) by using multi-step reasoning to enhance GPT-3's performance on AUT. Similar to the multi-round interaction framework utilized in LLM Debate (Du et al., 2024), LLM Discussion (Lu et al., 2024) develops a role-play-enhanced LLM discussion framework to augment ChatGPT's performance on the WCT and SCT benchmarks. Unlike existing prompting methods, our proposed RAG-Novelty improves the LLM's performance by retrieving similar papers, assuming that novel papers should retrieve the latest publications.

2.3 LLM Performance Evaluation

Most existing studies (Summers-Stay et al., 2023; Stevenson et al., 2022; Guzik et al., 2023; Mednick, 1962) evaluate LLM performance on benchmarks (Section 2.1) using human assessments. For example, Guzik et al. (2023) evaluated LLM responses to the TTCT, which were scored by Scholastic Testing Services (STS). Other studies rely on LLMs to score responses from another LLM. Zhao et al. (2024) used a more powerful GPT-4 to evaluate the performance of smaller LLMs, while Lu et al. (2024) utilized ChatGPT to assess responses generated by GPT-4. Additionally, Lu et al. (2024) compared LLM-generated scores with human evaluations, finding that LLM evaluations correlated more closely with the average human score. Both Luchini et al. (2023) and (Organisciak et al., 2023) fine-tuned models on human-scored data to evaluate LLM responses. Since our benchmark provides ground-truth binary labels, evaluation is straightforward.

3 Scholarly Novelty Benchmark

Unlike the semantic novelty evaluated by the benchmarks from cognitive science (Section 2.1), novelty in scholarly publications refers to introducing new ideas, methods, or discoveries that have previously not been explored or established in the literature. Evaluating novelty is fundamentally an exercise in understanding the relationship between ideas across time rather than simply assessing new ideas or techniques. This understanding is crucial in determining the contribution of a research paper. The assumption can be made that later works are more novel than prior works, as they typically introduce new ideas and methodologies in the current research climate (Beaty and Silvia, 2012; Acar et al., 2019). In this paper, we apply this assumption to establish ground truth values for our created benchmark SchNovel.

3.1 Dataset Collection and Structure

The arXiv dataset[2] comprises approximately 2.5 million articles, with the earliest dating back to 1986. All articles are categorized into eight distinct fields[3], each of which has some sub-fields. We picked six out of eight fields: Computer Science (cs), Mathematics (math), Physics (physics), Quantitative Biology (q-bio), Quantitative Finance (q-fin), and Statistics (stat), as we did not collect enough papers in other fields. Figure 6 in Appendix A.1 shows the number of papers published each year for each field. To assess the ability of LLMs to assess the novelty of research papers, we sampled a subset of articles from each field, denoted as dataset $D = \{(f, g, s, x, y, label)_i\}_{i=1}^{N}$ where $N = 15000$, following the procedure outlined in Algorithm 1 in Appendix A.4, where f represent the field, x and y represent the paper ids, s represents the year in which paper x was published, g represents the number of years paper y was published before paper x and $label$ equals to paper x as we assume in the same field, later published paper is more novel.

3.2 Tasks and Evaluation Metrics

We define the task as assessing which paper is more novel when given a pair of papers. Specifically, for each tuple $(f, g, s, x, y, label)_i$, the selected LLM is provided with the title, abstract, and optional metadata for each paper—information typically available to a reviewer. However, unlike a full review, the model does not have access to the full text, making the task more challenging. While the abstract offers a strong indication of a paper's content and key findings, important details may be missed. By limiting the context to the abstract and

[2]Available at https://www.kaggle.com/datasets/Cornell-University/arxiv

[3]See the full taxonomy at https://arxiv.org/category_taxonomy

metadata, we also improve efficiency in terms of token consumption and cost. We will discuss the potential limitations of this approach in Section 8. Various comparison methods, such as point-wise and pair-wise, can be employed, and we evaluate performance based on accuracy.

4 RAG-Novelty

Assessing the novelty in scholarly papers requires the model to have a good understanding of past and present works to accurately judge whether a paper is novel in the current research climate. However, once trained, LLMs are frozen in time, meaning that they are no longer updated with the latest information, so they lack this understanding of the field's current state. Inspired by RAG, we propose a novel method, RAG-Novelty, to further improve LLMs' capability to assess novelty in our benchmark. As shown in Figure 1, apart from the information, like abstract, that can be utilized for a paper, we apply the paper abstract as a query to retrieve top-K papers from the already built index, and then create a prompt based on the query paper and the retrieved papers to ask the LLM to score the novelty of the query paper from 0 to 10.

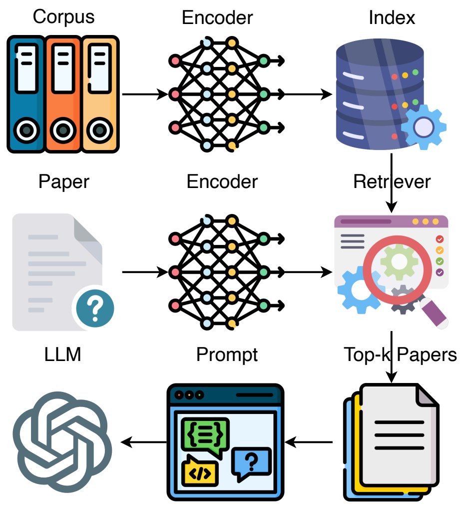

Figure 1: The overview of RAG-Novelty

4.1 Indexing and Retriever

To assess the novelty of a paper with the information provided by our SchNovel, such as title, abstract, and other metadata excluding the whole paper, an expert human reviewer in the same field may accurately score the novelty, a junior human reviewer, however, is likely not confident of scoring the novelty directly and instead will first review some similar papers and then assess the novelty. To mimic the review process taken by a human reviewer, we randomly sampled 500 papers from all years from 2000 to 2023, yielding 12000 papers for each field. Then, the abstracts of these papers are encoded into embeddings using OpenAI's *text-embedding-3-small*[4] model. The retrieval is the exact search method based on cosine similarity, as the number of candidates is very small. Our method can also handle huge candidate corpus by building an approximate nearest neighbor searching index using faiss (Douze et al., 2024; Johnson et al., 2019).

When a human reviewer conducts a literature search, it is naturally impossible to retrieve papers published after the query paper's publication date. To simulate this realistic constraint in our evaluation, we filtered out any papers published after the query paper and retrieved the top-k relevant papers from those published prior to or on the same date. However, in the context of pairwise comparisons, where we are assessing the novelty between two papers with different publication dates, it is reasonable to retrieve papers up to the publication date of the more recent paper. To prevent any leakage, we ensured that the papers themselves were excluded from the top-k retrieved documents. This approach mirrors a realistic scenario in which novelty is judged relative to the latest available knowledge at the time of publication. By implementing this strategy, we ensure that the novelty assessment remains fair and contextually appropriate, avoiding any temporal bias while maintaining the integrity of the comparison.

4.2 Prompt

We first compared the zero-shot, two-shot, and self-reflection prompts and found that the self-reflection prompt performed the best (Section 6.1). So, for RAG-Novelty, we built the prompt, shown in Appendix A.6, based on the self-reflection prompt, shown in Appendix A.3, by incorporating the information of the retrieved papers. Specifically, we added a "Contextual Data Analysis" instruction that assumes that the more recent papers are retrieved, the more novel this query paper is:

Average the published dates of the retrieved documents. Use this average date as additional context for your evaluation.

[4] https://platform.openai.com/docs/guides/embeddings

Consider that papers with an average date that is later or more recent in time are generally more novel.

5 Experimental Setup

5.1 Baseline Methods

Zero-Shot as shown in Appendix A.2, involves providing the model with two research papers' titles, abstracts, and four-step instructions, guiding the LLM to leverage its internal knowledge to make an informed decision. We also conducted a pointwise comparison by revising the zero-shot prompt to instruct the LLM score on the novelty of each paper first and then compare which one is more novel.
Two-Shot We randomly sampled two example paper pairs and added them to the zero-shot prompt.
Chain of Thought (CoT) (Wei et al., 2023) elicits reasoning within models by giving the model time to "think". We achieved CoT by adding instructions to Zero-Shot guiding LLMs to provide demonstrations.
Self-Reflection (Renze and Guven, 2024) has shown several strides in improving LLMs' logical fallacies by prompting the model to reflect on its incorrect solutions. We adopted this strategy to design a prompt, which is shown in Appendix A.3.
Self-Consistency (Wang et al., 2023) assumes that ground truth answers can be achieved through different reasoning paths. We followed the original paper to sample 10 generated sequences and voted majority.
LLM Discussion (Lu et al., 2024) assigns LLMs with different roles and lets them discuss with each other before making the final decision. We adopted LLM Discussion to simulate the review process taken by human reviewers. Specifically, we assume the papers are submitted to a conference to be reviewed, and we designed four roles: (a) a professor; (b) a PhD student; (c) an editor of a prestigious journal; (d) the chair of the conference where the professor, PhD student, and editor are all reviewers and they have two round discussions and the chair make the final decision. The prompt is shown in Appendix A.5.

5.2 LLM Configuration

We adopted the default settings of API [5] for Zero-Shot, Two-Shot, CoT, and RAG-Novelty. We followed the Self-Consistency to adopt the temperature as 0.7 and set the number of reasoning paths as 10. For LLM discussion, we limit the max tokens to 200 to avoid overwhelming the model with long inputs in subsequent rounds of discussion. For Self-consistency, we limit the max tokens so that the response is concise, as long reasoning for this task is unnecessary because we're looking for consistency rather than depth. In both cases, we prompt the model to limit its output to 150 tokens to ensure that its response fits within the 200 token limit.

[5] https://platform.openai.com/docs/guides/chat-completions

Method	cs	math	physics	qbio	qfin	stat
Zero-Shot	0.64	0.55	0.57	0.54	0.55	0.63
Two-Shot	0.62	0.55	0.57	0.54	0.55	0.60
CoT	0.63	0.56	0.57	0.54	0.56	0.62
Self-Reflection	0.65	0.56	0.58	0.56	0.57	0.63
LLM Discussion	0.60	0.55	0.56	0.53	0.50	0.58
Self-Consistency	<u>0.66</u>	<u>0.57</u>	<u>0.59</u>	<u>0.58</u>	<u>0.60</u>	<u>0.64</u>
RAG-Novelty	†0.72⋆	0.58⋆	†0.62⋆	†0.65⋆	†0.73⋆	†0.68⋆

Table 1: RAG-Novelty vs. Baselines on SchNovel with GPT-4o-mini. Averaged accuracy is reported. † denotes statistically significant enhancements over the second-best result, with p-values < 0.05, as determined by the McNemar test. The best results across different methods are denoted with the symbol ⋆. The second-best results across different methods are underlined.

5.3 Research Questions

This study aims to address several key questions regarding the performance of LLMs on the SchNovel benchmark.

- **R1:** Which comparison approach yields better results: pointwise or pairwise?
- **R2:** How do different LLMs perform in assessing the novelty of research papers?
- **R3** How does the category of the research paper affect the performance of LLMs?
- **R4:** How does the publication start year influence the performance of LLMs?
- **R5:** What impact does the gap between the publication years of research papers have on LLMs' performance?
- **R6:** What are the effects of other metadata attributes on LLMs' performance?
- **R7:** Can RAG-Novelty outperform recent baselines?

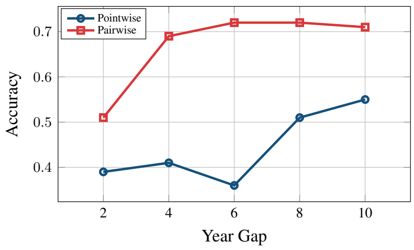

Figure 2: Pointwise vs. pairwise. The metrics above were obtained in the cs field with the start year $s = 2023$ and GPT-4o-mini.

6 Experimental Results

6.1 RAG-Novelty vs. Baseline Models (R7)

In this experiment, we evaluate the performance of RAG-Novelty against baseline methods. All methods use GPT-4o-mini, and the accuracy is averaged across different start years s and year gaps g. Pairwise comparison is applied to all methods, and we account for position bias by swapping the order of the two papers in the comparisons.

Two-Shot does not improve upon Zero-Shot as it typically does in other tasks. We attribute this to the complexity of the novelty assessment task, which requires deeper contextual understanding and comparison between papers—something that randomly selected examples may not effectively convey. Through iterative prompt refinement, Self-Reflection outperforms CoT in all fields except mathematics. LLM Discussion methods perform the worst, failing to even surpass Zero-Shot. Self-Consistency achieves the best results among baseline methods, demonstrating that obtaining answers through different reasoning paths helps improve performance. Our RAG-Novelty achieves the highest results overall, significantly outperforming the second-best method, except in the mathematics field. Across all methods, the improvement in mathematics is limited, possibly due to the slower progression of the field, the prevalence of symbols that LLMs struggle to interpret, or a lack of sufficient mathematical content in the training data compared to other fields.

6.2 Pointwise vs Pairwise (R1)

As mentioned in Section 5.1, we revised the pairwise Zero-Shot prompt (Appendix A.2) to a pointwise one. We compared the two methods by evaluating them in the cs field with the start year 2023, crossing different year gaps. As shown in Figure 2, pairwise is consistently much better than pointwise across different year gaps. This significant difference highlights the importance of context. As with human evaluations, providing relevant context or reference points is crucial for accurate assessments (Yan et al., 2022), allowing reviewers to consider the broad implications of a paper within the current research landscape. Pairwise comparisons align with this process, simplifying the task of considering the relative merits of two papers side-by-side rather than evaluating each one in isolation. Thus, pairwise comparisons are used in the rest of the following experiments.

6.3 The Impact of Different Fields (R3)

In Figure 3, the cs category shows the highest accuracy across most year gaps (starting in 2023), likely due to the availability of data and well-defined evaluation metrics. In contrast, math and physics show lower accuracy, likely due to domain-specific challenges such as complex notation in mathematics and theoretical frameworks in physics.

One explanation is the lack of domain knowledge in ChatGPT's training data, which, being sourced from the internet, may not adequately cover specialized fields. Research has shown that LLMs exhibit biases in various prompts and tasks (Cheng et al., 2023; Stranisci et al., 2023), suggesting potential categorical biases in lesser-known or slower-growing domains. This has significant implications for using AI tools in academia and industry, particularly in automated scoring or ranking systems, where such biases could perpetuate inequalities.

6.4 The Impact of Different Start Years and Year Gaps (R4 & R5)

To better understand how different start years affect the performance of LLMs in evaluating novelty, we investigated the model's results for five distinct start years. As shown in Figure 4, the model's results for all five start years were relatively consistent across different year gaps. This suggests that the model's ability to evaluate novelty between two papers is more dependent on the year gap between them than the specific publication years.

For example, evaluating two papers with a 10-year gap from 2009 to 2019 should be equivalent in difficulty to evaluating two papers with a 10-year gap from 2013 to 2023. Regardless of the boundary years within those ranges (i.e., considering papers published at specific points like 2009 and 2019, versus 2013 and 2023), it's the decade-long gap between the papers' publication times that makes

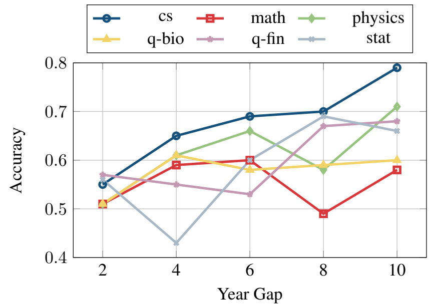

Figure 3: Comparison of fields. The metrics above were obtained using Self-Reflection in cs field with the start year $s = 2023$ with GPT-4o-mini.

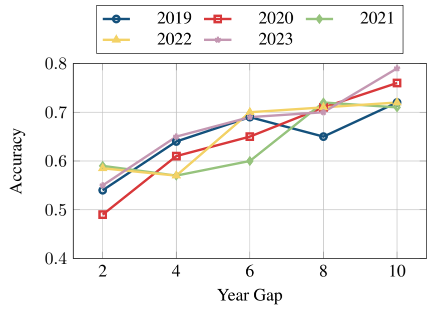

Figure 4: Comparison of Start Years. The metrics above were obtained using Self-Reflection in the cs field with GPT-4o-mini.

it easier for the model to make such a binary evaluation.

6.5 The Impact of Different LLMs (R2)

All LLMs can vary significantly depending on their training data and model architecture. With various different models available, it is essential to understand how they perform when assessing the originality of ideas presented in research papers. In this section, we examine the impact of using different LLMs on evaluating novelty.

Our findings in Table 2 reveal significant disparities in performance across different LLMs. GPT-4o-mini, GPT-3.5, and Gemma 2 performed more in line with expectations, achieving a more balanced distribution of predictions throughout all year gaps. Notably, GPT-4o-mini outperformed all other models, demonstrating a substantial advantage over smaller models like LLaMA 3.1-8b, Mistral 7b, and Gemma 2-9b.

Despite such success, even ChatGPT 4o-mini and ChatGPT 3.5 exhibit position bias, where the order of papers in the prompt affects their decision-making instead of content alone. This bias is magnified in smaller models, which lack extensive training compared to larger models. For example, Mistral 7b is heavily biased toward the last paper in the prompt. This aligns with known issues regarding LLMs' performance being best when relevant information appears towards the beginning or end of the prompt (Liu et al., 2024; Dai et al., 2024).

In contrast, LLaMA 3.1-8b exhibits a different bias, favoring the first paper that appears toward the middle of the prompt. According to Dubey et al. (2024), the LLaMA 3.1 models excel at "needle-in-the-haystack" tasks, where one needs to find specific information in large amounts of text (Kamradt, 2023), ultimately fixing the issues described in Liu et al. (2024). This is similar to skimming, which is efficient for finding specific information but may not facilitate deep understanding. Thus, while LLaMA 3.1-8b excels at retrieving specific information from anywhere in a context, this skillset is not ideal for evaluating novelty between two papers.

6.6 The Impact of Metadata (R6)

Previously, our experiments evaluated novelty based solely on a paper's title and abstract. However, human evaluations often take into account various metadata that can subtly influence reviewers' decisions. This metadata-induced bias has significant implications for research evaluations and highlights the need for more anonymous reviewal processes, leading to solutions such as double-blind reviewal processes. A pairwise comparison was applied for all the experiments in this section, and we accounted for position bias by swapping the order of the two papers in the comparisons.

6.6.1 Adding a TLDR Summary

We utilized the SciTLDR model (Cachola et al., 2020) from the Semantic Scholar API (Kinney et al., 2023) to generate TLDRs for our dataset, expecting this additional information to enhance accuracy by helping the model generalize and better understand the paper. As shown in Table 3, adding TLDRs decreases the accuracy across all year gaps. Nevertheless, incorporating such data did mitigate position bias, as evidenced by the negligible difference between ascending and descending year accuracies across nearly all year gaps.

6.6.2 Adding Author

We then added the author to the prompt, expecting that this additional information would not affect the model performance as the authors should not influence the novelty assessment. To our surprise, adding such information did help mitigate some

Year Gap	ChatGPT4o-mini			ChatGPT3.5			LLaMA 3.1-8b			Mistral-7b			Gemma-2-9b		
	Asc Yr	Desc Yr	Acc.	Asc Yr	Desc Yr	Acc.	Asc Yr	Desc Yr	Acc.	Asc Yr	Desc Yr	Acc.	Asc Yr	Desc Yr	Acc.
2	0.44	0.66	0.55	0.46	0.62	0.54	0.03	0.98	0.51	1.00	0.00	0.50	0.66	0.38	0.52
4	0.58	0.72	0.65	0.58	0.57	0.58	0.02	0.97	0.50	1.00	0.00	0.50	0.70	0.48	0.59
6	0.63	0.75	0.69	0.67	0.60	0.64	0.01	0.99	0.50	1.00	0.01	0.51	0.69	0.41	0.55
8	0.63	0.77	0.70	0.63	0.68	0.66	0.01	0.99	0.50	0.99	0.00	0.50	0.76	0.46	0.61
10	0.79	0.78	0.79	0.67	0.71	0.69	0.05	0.97	0.51	0.99	0.01	0.50	0.80	0.43	0.62
Average	0.61	0.74	0.68	0.60	0.64	0.62	0.02	0.98	0.50	0.996	0.004	0.50	0.72	0.43	0.58

Table 2: Comparison of different LLMs. The metrics above were obtained using Self-Reflection in the cs field with the start year $s = 2023$. "Asc Yr" indicates that the older paper is presented first in the prompt, while "Desc Yr" means the newer paper is presented first.

Year Gap	Zero-Shot			Self-Reflection			Self-Reflection w/ tldr			Self-Reflection 2 w/ author		
	Asc Yr	Desc Yr	Acc.	Asc Yr	Desc Yr	Acc.	Asc Yr	Desc Yr	Acc.	Asc Yr	Desc Yr	Acc.
2	0.41	0.60	0.51	0.44	0.66	0.55	**0.53**	**0.51**	0.52	**0.57**	**0.55**	0.56
4	0.63	0.74	0.69	0.58	0.72	0.65	**0.64**	**0.64**	0.64	0.67	0.62	0.65
6	0.64	0.79	0.72	0.63	0.75	0.69	**0.66**	**0.69**	0.68	0.75	0.62	0.69
8	0.66	0.77	0.72	0.63	0.77	0.70	0.69	0.61	0.65	**0.68**	**0.67**	0.68
10	0.64	0.78	0.71	0.78	0.79	0.79	**0.76**	**0.76**	0.76	0.80	0.75	0.78
Average	0.60	0.74	0.67	0.61	0.74	0.68	0.66	0.64	0.65	0.69	0.64	0.67

Table 3: The impact of metadata. The metrics above were obtained using Self-Reflection in the cs field with the start year $s = 2023$ and GPT-4o-mini. "Asc Yr" indicates that the older paper is presented first in the prompt, while "Desc Yr" means the newer paper is presented first.

of the position bias, as seen in the bold results in Table 3, but overall, it decreased the performance slightly.

6.6.3 Adding Affiliation

We selected two universities, one of which is a top research university and the other a teaching university, to study whether affiliation bias exists in LLMs' assessment of novelty.[6] Specifically, we first assigned the top research university as the affiliation of the more recently published paper and the teaching university to the earlier published paper, with the results shown in blue. Then, we swapped the affiliations, and the results are shown in red. As illustrated in Figure 5, the top research university starts with similar accuracy to the teaching university at a year gap of $g = 2$, but as the year gap increases, the top research university consistently outperforms the teaching university. This suggests that affiliation bias exists in LLMs' novelty assessments, with a tendency to "trust" papers from top research universities. However, although we observed LLMs' preference for choosing the top research university, the top research university experiments are undertaken without affiliation. This unexpected result raises questions about how LLMs process affiliation information, which warrants further investigation to better understand and mitigate such biases.

[6]The real names of the universities are not used to ensure objectivity and to avoid any unintended bias or implications.

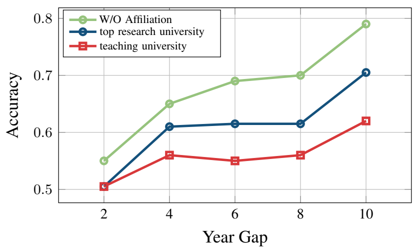

Figure 5: Comparison of different organizations. The metrics above were obtained using Self-Reflection in the cs field with start year $s = 2023$ and GPT-4o-mini.

7 Conclusion and Future Work

To evaluate LLMs' ability to assess novelty in scholarly publications, we introduce SchNovel, a benchmark consisting of 15,000 pairs of papers across six fields. We conducted extensive experiments to understand how various factors influence LLM performance on SchNovel. To enhance LLMs' capability to assess novelty, we propose RAG-Novelty, which significantly outperforms strong baseline models in comprehensive experiments. For future work, we plan to expand SchNovel by including more papers and covering additional fields to evaluate LLMs on a larger scale. Another promising direction is investigating which part of a paper best represents the whole for novelty assessment by LLMs. Additionally, studying how LLMs process affiliation and addressing biases in novelty evaluation, such as position and affiliation bias, is an important area for further research.

8 Limitations

Our study evaluates an LLM's ability to assess novelty using a research paper's title, abstract, and metadata. While the abstract provides a strong indication of a paper's content and key findings, it may not fully capture the novelty of the research compared to the complete text. Abstracts often summarize the main ideas but may omit important technical details. Although this approach streamlines the evaluation process, it could occasionally limit the depth of the novelty assessment due to the absence of a more comprehensive context.

Additionally, the exclusive use of arXiv data is limiting. We selected arXiv as an initial step for its broad, publicly accessible range of publications. Future work can improve robustness using peer-reviewed publications and sampling papers from more sources.

References

Selcuk Acar, Ahmed M. Abdulla Alabbasi, Mark A. Runco, and Kenes Beketayev. 2019. Latency as a predictor of originality in divergent thinking. *Thinking Skills and Creativity*.

Teresa M Amabile. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of personality and social psychology*, 43(5):997.

Barrett R. Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition, Chicago, IL, USA, June 23-26, 2024*, pages 413–425. ACM.

Roger E. Beaty and Paul Silvia. 2012. Why do ideas get more creative across time? an executive interpretation of the serial order effect in divergent thinking tasks. *Psychology of Aesthetics, Creativity, and the Arts*, 6:309–319.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.

Viola A Brody. 1966. Modes of thinking in young children: A study of the creativity-intelligence distinction. *Archives of General Psychiatry*, 14(3):329–331.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S. Weld. 2020. Tldr: Extreme summarization of scientific documents. *Preprint*, arXiv:2004.15011.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 30:1–30:34. ACM.

Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2023. Creativity support in the age of large language models: An empirical study involving emerging writers. *CoRR*, abs/2309.12570.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6437–6447, New York, NY, USA. Association for Computing Machinery.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

J Pv Guilford. 1964. Some new looks at the nature of creative processes. *Contributions to mathematical psychology. New York: Holt, Rinehart & Winston*.

Erik E Guzik, Christian Byrge, and Christian Gilde. 2023. The originality of machines: Ai takes the torrance test. *Journal of Creativity*, 33(3):100065.

Weiping Hu and Philip Adey. 2002. A scientific creativity test for secondary school students. *International Journal of Science Education*, 24(4):389–403.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Julian Just, Thomas Ströhle, Johann Füller, and Katja Hutter. 2024. Ai-based novelty detection in crowdsourced idea spaces. *Innovation*, 26(3):359–386.

Gregory Kamradt. 2023. Llmtest_needleinahaystack. `https://github.com/gkamradt/LLMTest_NeedleInAHaystack/blob/main/README.md`. Accessed: 2024-09-08.

R. M. Kinney, C. Anastasiades, R. Authur, I. Beltagy, and et al. J. Bragg. 2023. The semantic scholar open data platform. *ArXiv*, abs/2301.10140.

Robert W Lissitz and Joseph L Willhoft. 1985. A methodological study of the torrance tests of creativity. *Journal of Educational measurement*, 22(1):1–11.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung-yi Lee, and Shao-Hua Sun. 2024. LLM discussion: Enhancing the creativity of large language models via discussion framework and role-play. *CoRR*, abs/2405.06373.

Simone Luchini, Nadine T Maliakkal, Paul V DiStefano, John D Patterson, Roger Beaty, and Roni Reiter-Palmon. 2023. Automatic scoring of creative problem-solving with large language models: A comparison of originality and quality ratings.

Martha T Mednick and Sharon Halpern. 1968. Remote associates test. *Psychological Review*.

Sarnoff Mednick. 1962. The associative basis of the creative process. *Psychological review*, 69(3):220.

Pronita Mehrotra, Aishni Parab, and Sumit Gulwani. 2024. Enhancing creativity in large language models through associative thinking strategies. *CoRR*, abs/2405.06715.

Jay A Olson, Johnny Nahas, Denis Chmoulevitch, Simon J Cropper, and Margaret E Webb. 2021. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, 118(25):e2022340118.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49:101356.

William Orwig, Emma R Edenbaum, Joshua D Greene, and Daniel L Schacter. 2024. The language of creativity: Evidence from humans and large language models. *The Journal of creative behavior*, 58(1):128–136.

Antoine Bellemare Pépin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A. Olson, Yoshua Bengio, and Karim Jerbi. 2024. Divergent creativity in humans and large language models. *CoRR*, abs/2405.13012.

Matthew Renze and Erhan Guven. 2024. Self-reflection in llm agents: Effects on problem-solving performance. *Preprint*, arXiv:2405.06682.

M.A. Runco. 2011. Runco creativity assessment battery (rcab). *Creativity Testing Services*.

Mark A Runco and Garrett J Jaeger. 2012. The standard definition of creativity. *Creativity research journal*, 24(1):92–96.

Sotaro Shibayama, Deyun Yin, and Kuniko Matsumoto. 2021. Measuring novelty in science with word embedding. *PloS one*, 16(7):e0254034.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

Claire Stevenson, Iris Smal, Matthijs Baas, Raoul P. P. P. Grasman, and Han L. J. van der Maas. 2022. Putting gpt-3's creativity to the (alternative uses) test. In *Proceedings of the 13th International Conference on Computational Creativity, Bozen-Bolzano, Italy, June 27 - July 1, 2022*, pages 164–168. Association for Computational Creativity (ACC).

Marco Antonio Stranisci, Rossana Damiano, Enrico Mensa, Viviana Patti, Daniele Radicioni, and Tommaso Caselli. 2023. WikiBio: a semantic resource for the intersectional analysis of biographical events. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12370–12384, Toronto, Canada. Association for Computational Linguistics.

Douglas Summers-Stay, Clare R Voss, and Stephanie M Lukin. 2023. Brainstorm, then select: a generative language model improves its creativity score. In *The AAAI-23 Workshop on Creative AI Across Modalities*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Xinyi Yan, Chengxi Luo, Charles L. A. Clarke, Nick Craswell, Ellen M. Voorhees, and Pablo Castells. 2022. Human preferences as dueling bandits. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 567–577, New York, NY, USA. Association for Computing Machinery.

Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, Qi Guo, Ling Li, and Yunji Chen. 2024. Assessing and understanding creativity in large language models. *CoRR*, abs/2401.12491.

A Appendix

A.1 Statistics of arXiv

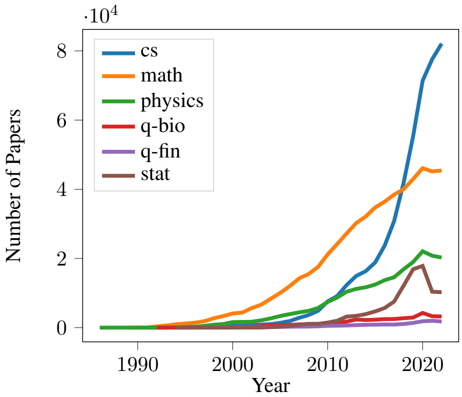

Figure 6: Number of Papers for Each Field (Up to 2023)

A.2 Zero-shot

You will be provided with the title and abstract of two research papers. Please determine which of the two articles is more novel. Follow these steps for evaluation.

Step 1: Identify the problem and solution that the research paper attempts to solve.

Step 2: Determine how unique the solution is given the current research landscape in 2024. Does the paper introduce a new idea, theory, or concept that has not been previously discussed in the literature?

Step 3: Determine how creative the solution is given the current research landscape in 2024. Does it apply a known idea in a completely new context or in a way that has not been done before?

Step 4: Using the findings from Steps 1-3, determine which paper is more novel.

In your response, please only state which paper is more novel (e.g., 1 if Paper 1 is more novel; 2 if Paper 2 is more novel).

User Prompt:

- Paper 1 Title: [paper_1_title]
- Paper 1 Abstract: [paper_1_abstract]
- Paper 2 Title: [paper_2_title]
- Paper 2 Abstract: [paper_2_abstract]

A.3 Self-Reflection

You are an advanced language model tasked with determining the novelty of research papers in 2024. Your goal is to evaluate and compare the novelty of two research papers based on their titles, abstracts, and any other given metadata.
The order in which the papers are presented is random and should not influence your evaluation.
Step 1: Independent Evaluation
Analyze each research paper's title and abstract **independently**. Treat each paper as if it is the only one under review at that moment.
Consider the following aspects for each paper:

- **Novelty of Methodology**: Are the methods used new and innovative?
- **Surprisingness of Findings**: Are the findings unexpected or counterintuitive?
- **Impact on Existing Knowledge**: How does the research challenge or expand current scientific understanding?
- **Potential for Future Research**: Does the paper open up new directions for research?
- **Relevance to 2024 Scientific Understanding**: How well does the paper align with or push the boundaries of current trends?

Step 2: Quantitative Assessment

- Assign a score from 1-10 to each research paper for its novelty, with 10 being the most novel. This score should be based solely on the content of the title and abstract.
- Provide a brief justification for the score, using specific quotes and context.

Step 3: Final Comparison

- After independently scoring each paper, compare the scores.
- Determine which paper exhibits greater novelty based on the higher score, and provide the identifier (X or Y) of the more novel paper.

Important: The order of presentation is random and should not influence your decision. Evaluate each paper strictly on its content and merit.

User Prompt:

- Paper X Title: [paper_x_title]
- Paper X Abstract: [paper_x_abstract]
- Paper Y Title: [paper_y_title]
- Paper Y Abstract: [paper_y_abstract]

A.4 SchNovel

Algorithm 1 Data Sampling Algorithm

$Fields \leftarrow [cs, math, physics, qbio, qfin, stat]$
$startYear \leftarrow [2019, 2020, 2021, 2022, 2023]$
$yearGap \leftarrow [2, 4, 6, 8, 10]$
$sampleNum \leftarrow 100$
$N \leftarrow 0$
$Dataset \leftarrow []$

```
for f in Fields do
    for s in startYear do
        for g in yearGap do
            while N ≠ sampleNum do
                x ← paper published in s from f
                y ← paper published in s-g from f
                label ← x
                Dataset ← (f, g, s, x, y, label)
                N ← N + 1
            end while
            N ← 0
        end for
    end for
end for
```

A.5 LLM Discussion

You are a [Role] with expertise across all areas of [Category]. You will be provided with the titles and abstracts of two research papers. Your task is to determine which of the two articles is more novel by evaluating their originality, contribution to the field, and potential impact. Focus on aspects such as new methodologies, unexplored problems, innovative solutions, and how the work advances the state of the art. Follow these steps for evaluation.

Step 1: Identify the problem and solution that the research paper attempts to solve.

Step 2: Determine how unique the solution is given the current research landscape in 2024. Does the paper introduce a new idea, theory, or concept that has not been previously discussed in the literature?

Step 3: Determine how creative the solution is given the current research landscape in 2024. Does it apply a known idea in a completely new context or in a way that has not been done before?

Step 4: Using the findings from Steps 1-3, determine which paper is more novel.

Please limit your response to 150 tokens max. In your response please conclude with: "The more novel and impactful paper is [Paper X or Paper Y]

User Prompt:

- Paper X Title: [paper_x_title]
- Paper X Abstract: [paper_x_abstract]
- Paper Y Title: [paper_y_title]
- Paper Y Abstract: [paper_y_abstract]
- *(Round 2 Discussion add on)* [previous_response] These are responses from other reviewers. Please revise your response if necessary... [other_responses]
- *(Round 3 Discussion add on)* These are responses from other reviewers. Please determine which paper is more novel... [other_responses]

A.6 RAG-Novelty

You are an advanced language model tasked with determining the novelty of research papers in 2024. Your goal is to evaluate and compare the novelty of two research papers based on their titles and abstracts.
The order in which the papers are presented is random and should not influence your evaluation.
Step 1: Independent Evaluation
Analyze each research paper's title and abstract **independently**. Treat each paper as if it is the only one under review at that moment.
Retrieve similar abstracts from a vector database based on the provided abstracts.
Contextual Date Analysis: Average the published dates of the retrieved documents. Use this average date as additional context for your evaluation. Consider that papers with an average date that is later or more recent in time are generally more novel.
Consider the following aspects for each paper:

- **Novelty of Methodology:** Are the methods used new and innovative?
- **Surprisingness of Findings:** Are the findings unexpected or counterintuitive?
- **Impact on Existing Knowledge:** How does the research challenge or expand current scientific understanding?
- **Potential for Future Research:** Does the paper open up new directions for research?
- **Relevance to 2024 Scientific Understanding:** How well does the paper align with or push the boundaries of current trends?

Step 2: Quantitative Assessment

- Assign a score from 1-10 to each research paper for its novelty, with 10 being the most novel. This score should be based on the content of the title and abstract, as well as the contextual information from the average published dates.
- Provide a brief justification for the score, using specific quotes and context.

Step 3: Final Comparison

- After independently scoring each paper, compare the scores.
- Determine which paper exhibits greater novelty based on the higher score, and conclude with: "The more novel and impactful paper is [Paper X or Paper Y].

Important: The order of presentation is random and should not influence your decision. Evaluate each paper strictly on its content and merit, incorporating the additional context from the vector database as described.

User Prompt:

- Paper X Average Cosine Similarity: [paper_x_avg_cosine_similarity]
- Paper X Average Contextual Date: [paper_x_avg_contextual_date]
- Paper Y Average Cosine Similarity: [paper_y_avg_cosine_similarity]
- Paper Y Average Contextual Date: [paper_y_avg_contextual_date]
- Paper X Title: [paper_x_title]
- Paper X Abstract: [paper_x_abstract]
- Paper Y Title: [paper_y_title]
- Paper Y Abstract: [paper_y_abstract]