# Literature-Grounded Novelty Assessment of Scientific Ideas

Anonymous ACL submission

## Abstract

Automated scientific idea generation systems have made remarkable progress, yet the automatic evaluation of idea novelty remains a critical and underexplored challenge. Manual evaluation of novelty through literature review is labor-intensive, prone to error due to subjectivity, and impractical at scale. To address these issues, we propose the **Idea Novelty Checker**, an LLM-based retrieval-augmented generation (RAG) framework that leverages a two-stage retrieve-then-rerank approach. The Idea Novelty Checker first collects a broad set of relevant papers using keyword and snippet-based retrieval, then refines this collection through embedding-based filtering followed by facet-based LLM re-ranking. It incorporates expert-labeled examples to guide the system in comparing papers for novelty evaluation and in generating literature-grounded reasoning. Our extensive experiments demonstrate that our novelty checker achieves approximately 13% higher agreement than existing approaches. Ablation studies further showcases the importance of the facet-based re-ranker in identifying the most relevant literature for novelty evaluation.

## 1 Introduction

Novelty evaluation is foundational for determining whether ideas in scientific research, product development, or creative ideation introduce meaningful innovation relative to prior work. Yet, as the volume of published literature grows exponentially, manual verification of originality becomes impractical. This is further complicated by the inherent subjectivity of novelty judgments, which is why experts can more easily decide on similarity of two ideas (Picard et al., 2023) and often struggle to articulate why one idea is more novel than another. Further the evaluation becomes subjective as it also depends on personal knowledge and intuition gained from scientific literature (Ahmed et al., 2018; Picard et al., 2023).

Automated systems attempt to address this challenge by defining novelty as differences observed while comparing new ideas against prior work with similarity measures, but they exhibit critical limitations. Prior work has evolved from using n-gram frequency and lexical metrics (TF-IDF, LDA) (Wang et al., 2019; Sarica et al., 2019) to semantic embeddings (G'omez-P'erez et al., 2022; Su et al., 2024) that capture similarity but don't capture paraphrased variations of ideas and papers. Moreover, while recent approaches have adopted LLM-augmented pipelines to generate numerical scores (1-10) (Bougie and Watanabe, 2024; Wang et al., 2024) or provide binary classifications (novel versus not novel) (Lu et al., 2024; Li et al., 2024; Si et al., 2024; Su et al., 2024), they do not ground the rationales in existing works and frequently fail to capture subtle variations in phrasing, resulting in the misclassification of well-documented ideas as novel (Beel et al., 2025; Gupta and Pruthi, 2025). This shortcoming makes it difficult for researchers to distinguish novel ideas from incremental contributions or subtle cases of plagiarism (Gupta and Pruthi, 2025).

Moreover, all these approaches hinge on the successful retrieval of relevant literature for a given idea, a task that remains inherently challenging (Mysore et al., 2022; Mysore et al.; Stevenson and Merlo, 2022; Eger et al., 2019; Xu et al., 2014; Freestone and Karmaker, 2024). Prior works (Si et al., 2024; Lu et al., 2024) extract keywords from the idea to search for papers, so important work can easily be missed if the relevant paper does not have the exact keyword. This undermines the reliability of the novelty evaluation process.

We address these gaps with Idea Novelty Checker, a retrieval-augmented LLM pipeline that assesses an idea's novelty by comparing it to a set of the most relevant papers. First, Idea Novelty Checkercollects a broad set of relevant papers using keyword and snippet-based retrieval, as well as by retrieving papers similar to any seed papers provided. Next, an embedding-based similarity search filters this large collection, and a facet-based LLM re-ranker (Sun et al., 2023b) further narrows the set by comparing idea facets (purpose, mechanism, evaluation, and application) with those in the retrieved papers. Finally, expert-annotated in-context examples of *novel* and *not novel* ideas guide the system in generating literature-grounded rationales, mitigating subjectivity in novelty judgments.

In our experiments, we compared Idea Novelty Checkerwith baselines such as zero-shot prompting, prompt optimization approaches (DSPY and TextGRAD), and expert-based OpenReview examples. Our results show that expert-annotated in-context examples significantly improve classification performance. Comparisons with systems like AI Scientist and AI Researcher further demonstrate that our Idea Novelty Checkerachieves higher agreement with expert judgments, and our ablation studies shows that the combined retrieval and two-stage re-ranking are critical for identifying the most relevant papers.

Our contributions are as follows:

- We introduce Idea Novelty Checker, a retrieval-augmented LLM pipeline that automatically evaluates the novelty of scientific ideas. We plan to release our code and expert-collected data[1] to support work in automatic scientific discovery and provide literature-grounded novelty evaluations.
- We conducted a formative study in which experts evaluated ideas for novelty. The study revealed two key challenges to consider for novelty evaluation: clarifying what constitutes novelty given its subjectivity, and identifying relevant literature to assess it. This directly shaped the design of Idea Novelty Checker.
- Our method integrates keyword-based and snippet-based retrieval, followed by a two-stage re-ranker with embedding similarity and facet-based LLM re-ranking to identify key literature related to the given idea.
- We present extensive evaluations, ablation studies, and qualitative analyses that demonstrate the effectiveness of our novelty checker over existing approaches. Additionally, we discuss prompt sensitivity in LLMs for novelty evaluation further highlighting the importance of clear novelty definitions.

[1] anonymous.4open.science/r/idea_novelty_checker

## 2 Related Work

Automated approaches to novelty assessment in scientific literature have evolved considerably. Early methods relied on lexical similarity metrics, such as TF-IDF, LSA, and LDA (Wang et al., 2019; Sarica et al., 2019), but these techniques struggled to capture paraphrased concepts. Semantic embedding methods (G'omez-P'erez et al., 2022) improved on this by identifying deeper relationships, yet they are confined to surface-level comparisons (Mysore et al., 2022; Mysore et al.).

Retrieval-augmented LLM systems have emerged as a promising alternative, evaluating novelty either on a numerical scale (e.g., 1–10) (Bougie and Watanabe, 2024; Wang et al., 2024) or with binary classification (Li et al., 2024; Lu et al., 2024). AI Researcher (Si et al., 2024) uses a Swiss-system tournament ranking to compare ideas pairwise for *similarity* and *novelty* against individual papers. If any comparison has sufficient similarity, the idea is not novel. Another notable work is AI Scientist (Lu et al., 2024) that employs an iterative process in which an LLM generates queries from a research idea to retrieve relevant papers via the Semantic Scholar API (Kinney et al., 2023). The LLM then compares the idea against these papers until a clear decision is reached or a preset iteration limit is met. However, this approach has several limitations. First, it depends on keyword-based retrieval methods to get the most relevant papers to an idea, which may fail if the relevant papers do not contain the exact keywords. Second, comparing an idea against a large number of retrieved papers (sometimes over 100) can introduce known issues that LLMs often overlook instructions within a prompt (Loya et al., 2023; Sclar et al., 2024; Joshi et al., 2024). Finally, the decision of novelty evaluation relies on string matching for phrases like "decision made: novel" or "decision made: not novel." If such a decision is not reached, the idea is automatically considered novel. Independent evaluations (Beel et al., 2025) have further highlighted challenges in AI Scientist's novelty assessments, noting that the system can misclassify well-established concepts (micro-batching for stochastic gradient descent) as novel.

Our work builds on these insights by combining retrieval-then-rerank methods (Zhou et al., 2022; Naik et al., 2021) and uses expert-annotated examples to ensure that our novelty evaluations are grounded in the relevant literature.

## 3 Formative Study on Challenges in Evaluating Novelty

Evaluating idea novelty in scientific literature is inherently challenging because the criteria for novelty are subjective and can be defined in multiple ways. We conducted a formative study, referred to as the expert-annotated study throughout the paper, where the first and second authors reviewed the novelty of ideas based on the most relevant papers.

To assess idea novelty relative to existing literature, our study engaged experts who evaluated 51 ideas, comprising of 46 generated by the Scideator system (Radensky et al., 2024) and 5 adapted from accepted and rejected papers from OpenReview (ICLR 22, NeurIPS 23).[2] Each idea was classified into one of three categories: novel, moderately novel, or not novel. For every idea, we identified the most relevant papers through a two-step process: candidate papers were initially gathered using keyword-based queries and subsequently re-ranked using an LLM-based reranker (Sun et al., 2023b) according to their overall relevance to the idea.

The experts achieved a moderate agreement (Cohen's Kappa = 0.64). A key challenge identified was that experts sometimes relied on their broader domain knowledge rather than restricting their judgments to the most relevant papers, as the top papers alone were often not sufficient. Additionally, using three categories led to disagreements, as the distinction between novel and moderate novelty is itself subjective.

Building on these observations, we conducted a second study to minimize the influence of external knowledge. In this study, experts were instructed to base their judgments solely on the provided papers, and the categories were simplified to just two: novel and not novel.

Inspired by prior work (Portenoy et al., 2022; Kang et al., 2022; Chan et al., 2018; Suh et al., 2024; Srinivasan and Chan, 2024; Choi et al., 2024; Kang et al., 2024; Radensky et al., 2024) that categorizes research ideas into core facets such as purpose (the problem being addressed by the paper) and mechanism (the proposed solution to the problem), we define novelty as follows: An idea is considered novel if it differs from all retrieved papers in at least one core facet for the topic at hand—namely, purpose (i.e., a distinct objective), mechanism (i.e., a distinct technical approach), or evaluation (i.e., a distinct validation method). An idea is also considered novel if it uniquely combines these facets or applies them to a new application domain.

Using this controlled framework, we reannotated a set of ideas and evaluated 51 ideas, comprising of 34 new ones generated by Radensky et al. (2024) and 17 from the previous study where external knowledge had influenced novelty judgments. By narrowing the focus to the relevant papers alone, we observed fewer disagreements and achieved a higher agreement rate (Cohen's Kappa = 0.68). Of the 8 instances of disagreement, in 4 cases one expert overlooked details from the paper, in 2 cases the experts differed in their perception of subtle contributions to novelty, and in the remaining 2 cases no specific comments were provided.

This formative study highlights that a robust novelty checker depends critically on high-quality retrieval and a well-defined notion of novelty. These findings directly inform our methodology described in the following section.

## 4 Methodology: Idea Novelty Checker

Based on our formative study findings, our novelty checker is designed with two key components that address the two critical challenges: C1 ensuring high-quality retrieval of papers relevant to the idea for novelty assessment and C2 establishing clear criteria for judging novelty. The challenge C1 arises from the vast space of overlapping papers—there are hundreds of millions of potential matches. To address this, our system first filters the scientific literature to collect the most relevant papers for a given idea (see Section 4.1 and Step 1 and 2 in Figure 1). The input idea is then compared to each paper in this collection by prompting an LLM (see Section 4.2).

In tackling challenge C2, which arises from the inherent subjectivity and multiple definitions of novelty, the novelty checker leverages expert-labeled examples of *novel* and *not novel* ideas from the formative study. It generates reasoning grounded not only in comparisons against the most relevant papers but also in the standardized definition of novelty introduced earlier, which helps counteract subjectivity (see Step 3 in Figure 1). Below, we detail these two components.

[2] Fewer examples were taken from OpenReview since the primary focus was on evaluating ideas from Scideator.

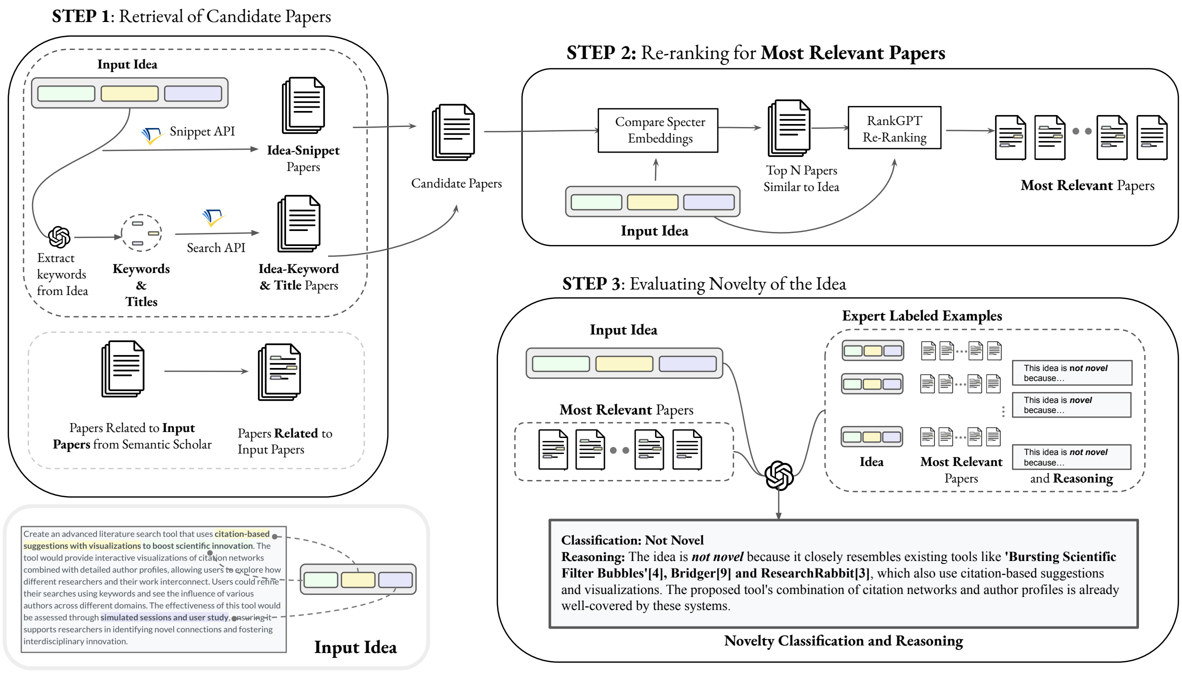


Figure 1: Our Idea Novelty Checker follows a retrieve-then-rerank approach for novelty evaluation. First, it gathers a broad set of papers relevant to an idea using query expansion (extracting keywords and titles from the idea) and snippet search (using the entire idea as input). Optionally, if seed papers are provided, we retrieve papers similar to these seed papers. Next, a two-stage re-ranking process is applied, where an embedding-based ranking strategy filters the large collection to top-$N$ papers, followed by a facet-based LLM re-ranker to identify the top-$k$ most relevant papers. Finally, these top-$k$ papers are used to assess the idea's novelty, guided by in-context examples that evaluate novelty with grounded reasoning.

### 4.1 Most Relevant Papers to Idea

Following established information retrieval practices (Gao et al., 2024; Nouriinanloo and Lamothe, 2024; Abdallah et al., 2025; Meng et al., 2024; Sun et al., 2023a; Baldelli et al., 2024), our pipeline uses a two-phase approach for identifying the most relevant papers to a given idea. First, we gather a broad set of candidate papers related to the idea. Then, we re-rank these candidates in two steps: first using embedding-based similarity, and then applying LLM-based re-ranking to facet-based similarity.

**STEP 1: Retrieval of Candidate Papers**

To accurately assess the novelty of an idea, it is crucial to compare it against a comprehensive collection of papers that cover the various facets of the idea. For a given idea and its corresponding papers (if any) used to generate the idea, we find more related papers to these input seed papers using the Semantic Scholar API [3]. However, simple retrieval methods often overlook important aspects of an idea (Mysore et al., 2022; Wang et al., 2023). To improve the paper collection's coverage we follow (Lu et al., 2024; Si et al., 2024) and employ a query-based retrieval method, where search queries are generated corresponding to different keywords related to the idea, and queried through the Semantic Scholar Search API (Kinney et al., 2023). Corresponding to each search query, papers are added to the collection of relevant papers. We prompt the LLM ($LLM_{query}$) to generate these search queries based on the keywords and potential titles related to the idea.

Next, we also employ Semantic Scholar's snippet search[4], which is trained to identify similar snippets (approximately 500 words of text) in other papers. We leverage the context size of this retrieval mechanism by incorporating the entire idea into the snippet search. Finally, we combine the seed papers and their related works with the papers retrieved from the two Semantic Scholar based query-retrieval method. This combined set form the candidate papers for the ideation process.

[3] api.semanticscholar.org/api-docs/recommendations

[4] api.semanticscholar.org/api-docs/#tag/Snippet-Text

**STEP 2: Re-ranking for Most Relevant Papers** To identify the papers most likely to overlap with the candidate idea, we implement a two-stage re-ranking process that combines embedding-based filtering with an LLM-based re-ranking approach. To identify the papers most likely to overlap with the candidate idea, we implement a two-stage re-ranking process that combines embedding-based filtering with an LLM-based re-ranking approach.

First, we employ **embedding-based filtering** to compute the semantic similarity between the idea and each paper in our collection of papers from STEP 1. We select the top $N$ papers with the highest cosine similarity between their embeddings and the idea embedding. While this embedding-based ranking efficiently narrows down the collection of papers, it is limited in its capacity to capture deeper and more contextual relationships between different facets of the idea and the papers, in comparison to powerful state-of-art LLMs (Reimers and Gurevych, 2019).

To address these limitations we employ a popularly used **LLM-based re-ranker**, RankGPT (Sun et al., 2023b), which refines the initial ranking of candidate papers by examining how relevant each paper is to the idea. We change relevance criteria to match it with each key facet of the idea. RankGPT goes beyond simple surface similarities by comparing the papers against the idea's application domain, purpose, mechanism, and evaluation. It follows a clear set of priorities: first, it favors papers that match all key facets of the idea; then, it prefers those that align with the application domain and purpose; next, it considers papers that share similarities in purpose, mechanism, or evaluation; and finally, it ranks lower those that only partially match or address related facets. This approach ensures that the final ranking accurately reflects the relevance and depth of each paper in connection with the idea. We refer to the LLM used for RankGPT as ($LLM_{rankgpt}$).
This collection of $k$-most relevant papers is used by the novelty checker in the next step to evaluate the idea's novelty.

### 4.2 Idea Novelty Evaluation

To assess an idea's novelty, we prompt an LLM ($LLM_{novelty}$) with both the idea and its top-$k$ relevant papers. The LLM outputs a binary classification (novel or not novel) accompanied by reasoning based on the top-$k$ retrieved literature. To guide the LLM's judgment, we include $n_{examples}$ in-context examples drawn from our formative study, where $n_{examples}$ is treated as a hyperparameter. These examples reflect the experts' criteria for novelty: an idea is considered novel if it differs from all retrieved papers in at least one core facet—namely, purpose (i.e., objective of idea), mechanism (i.e., technical approach), evaluation (i.e., validation method), a unique combination of these facets, or if it applies the same facets to a new application domain.

## 5 Implementation & Baselines

**Dataset:** From our formative study, we collected 67 consensus-labeled examples (39 labeled as novel and 28 as non-novel). We split into training and test sets (35 for training and 32 for testing) with a balanced distribution of novel and non-novel ideas. Please refer to Table 4 in the Appendix for sample examples.

**Baselines:** We evaluated multiple baselines to benchmark our novelty assessment approach. First, we employed a zero-shot prompt as a straightforward baseline, and further refined this manually written prompt using Anthropic's prompt generator[5]. We also applied popular prompt optimization techniques such as DSPy (Khattab et al., 2023) and TextGRAD (Yuksekgonul et al., 2024), which optimize the prompt instructions using a train/validation split created from formative study examples.

As an alternative to using in-context examples from the formative study, we extracted reviews from ICLR and NeurIPS submissions via the OpenReview API (OpenReview). These reviews comprise aspects such as *strengths, presentation, limitations, soundness, weaknesses, questions, confidence, contribution, summary,* and *rating*. The input title and abstract were adapted to match the ideas in the training data using a style-change prompt[6]. After rigorous filtering, we identified approximately 8,156 submissions discussing idea novelty and manually selected reviews that specifically evaluated the core idea rather than the entire paper. From these, we randomly sampled 20 idea-review pairs to serve as an additional baseline with different in-context examples.

In addition to these baselines, we also compare our novelty checker '*prompt*' with that of AI Sci-

[5] https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/prompt-generator

[6] All prompts are provided in the anonymised codebase.

entist (Lu et al., 2024) (different from its paper reviewer) and AI Researcher (Si et al., 2024) on the same test set of ideas and fixed top 10 papers. We compare **only the prompts to assess novelty** of these two approaches with ours, rather than the entire system, because the test set containing the novelty judgments by experts were based on a fixed set of the 10 most relevant papers for each idea. Since different retrieval methods could introduce new papers and potentially change novelty classification, we standardize the most relevant papers to ensure a fair comparison of the prompts alone. Additionally, since both setups require a different style of input idea, we adapted the ideas to match the requirements of each system.

**Implementation Settings:** For our novelty evaluation system, we use SPECTER-2 (Cohan et al., 2020) as the default embedding model. Initially, we retrieve the top $N$ =100 papers using these embeddings, from which the top $k$ =10 most relevant papers are selected for comparison with the input idea. The default language model for the idea keyword extraction ($LLM_{query}$), re-ranking process ($LLM_{rankgpt}$), and novelty evaluation ($LLM_{novelty}$) is gpt-4o[7] . Expert-labeled data from the formative study is incorporated as in-context examples in the novelty checker. We experimented with various numbers of in-context examples (comprising idea-paper pairs along with their novelty class and reviews) and found that the best performance was achieved using 15 idea examples (random seed 100). For the OpenReview examples, the best setup involved 5 idea-review pairs. For DSPy we used 2 bootstrapped examples, and trained both DSPy and TextGRAD for 12 prompt iterations.

## 6 Experiments

In this section, we first compare different baselines on the dataset for novelty evaluations (Section 6.1). Next we present our findings from ablation studies that shows the imporance of each component in our approach (Section 6.2). ablations studies We supplement these findings with qualitative examples of expert-labeled ideas and compare our setup with recent novelty checkers (Section 6.3). We conclude with insights from prompt optimization experiments that highlights the sensitivity of LLMs to prompt variations for novelty evaluation tasks (Section 6.4).

[7]We used the model "gpt-4o" during August and September 2024.

### 6.1 Comparing Novelty Checker Prompts

Our experiments show that incorporating expert-annotated data as in-context examples significantly enhances novelty classification accuracy compared to zero-shot prompts, DSPY, TextGRAD, and setups using OpenReview examples (Table 1).Since OpenReview reviews do not reference the associated papers, we evaluated our expert-labeled examples both with and without including relevant papers to ensure a fair comparison. Notably, even when we excluded the relevant papers from the expert-labeled examples, our approach still outperformed the OpenReview baseline.

Additionally, we compared two configurations for DSPY, one with reasoning and one without. Our expert-labeled prompt consistently achieved higher performance than the prompt optimizations produced by these methods, and we posit that the number of examples for train/validation were not sufficient for prompt optimisers with gpt-4o. The TextGRAD prompt optimiser did not improve upon its initial system prompt. It provided valuable insights into the LLM's prompt sensitivity, which we further discuss in Section 6.4.

Our approach achieved over **10 times more agreement with expert-labeled examples** compared to AI Scientist, and **approximately 13% higher agreement** than AI Researcher, further validating the effectiveness of our novelty checker. It is important to note that AI Scientist defaults to "not novel" when it fails to reach a conclusion in novelty evaluation (18 out of 32 times), which may have impacted its agreement rates. We also present some qualitative examples in Figures 2, 3 and 4 of the Appendix, showcasing how these approaches evaluate the novelty of an idea.

### 6.2 Ablation Studies

**Setup:** To assess the contribution of each component in our novelty checker, we conducted ablation studies using 58 ideas (comprising 13 ‘not novel’ instances from our test set and 45 NLP papers from the literature). For this experiment, we focus on the ‘not novel’ cases, since the ideas labeled novel in expert-labeled test data can vary with different retrieved paper sets. In our ablations, we considered the following variations: (i) **Complete System:** Uses both keyword and snippet retrieval (each returning the top-$k$ documents based on Semantic Scholar's ranking), embedding filtering, and facet-based RankGPT re-ranking; (ii) **RankGPT Rele-**

| Models | Accuracy | Precision | Recall | F1 | Cohen Kappa |
|---|---|---|---|---|---|
| **Zero Shot Setting** | | | | | |
| Zero Shot | 0.68 | 0.76 | 0.64 | 0.65 | - |
| + improved prompt using Anthropic prompt generator | 0.68 | 0.70 | 0.64 | 0.64 | - |
| **Prompt Optimizers** | | | | | |
| **DSPy** | | | | | |
| - with idea, most relevant papers, class | 0.68 | 0.83 | 0.62 | 0.58 | - |
| - with idea, most relevant papers, class, reasoning | 0.66 | 0.82 | 0.58 | 0.52 | |
| **TextGRAD** | | | | | |
| - with idea, most relevant papers, class | 0.78 | 0.76 | 0.76 | 0.76 | - |
| **In-context Setting** | | | | | |
| **Open-Review Examples** | | | | | |
| - with idea & review (i.e., reasoning) | 0.59 | 0.55 | 0.51 | 0.43 | - |
| **Expert Labeled Examples** | | | | | |
| - with idea, reasoning | 0.75 | 0.76 | 0.77 | 0.75 | - |
| - with idea, most relevant papers, class | 0.78 | 0.77 | 0.76 | 0.77 | - |
| - with idea, most relevant papers, class, reasoning | 0.81 | 0.84 | 0.78 | 0.79 | 0.59 |
| **Other Novelty Checkers** | | | | | |
| **AI Scientist** (Lu et al., 2024) | 0.47 | 0.55 | 0.53 | 0.44 | 0.05 |
| **AI Researcher** (Si et al., 2024) | | | | | |
| - GPT-4O | 0.78 | 0.81 | 0.74 | 0.75 | 0.52 |
| - CLAUDE-3-5-SONNET | 0.56 | 0.63 | 0.61 | 0.56 | 0.19 |

Table 1: Experimental Results using gpt-4o on expert-annotated dataset.

**vance:** Used the same retrieval methods (keyword and snippet) plus embedding filtering, but replaced the facet-based RankGPT re-ranker with one based on general relevance (Sun et al., 2023b). This variation differs from the complete system only in the LLM re-ranking component, allowing us to assess the importance of facet-based re-ranking; (iii) **Embedding Filtering:** Omits the LLM re-ranker entirely, relying only on the embedding-based filtering. This setup allows us to assess the importance of the LLM re-ranking step; and (iv) **Snippet Retrieval** and **Keyword Retrieval**: Each of these setups returned the top-$k$ documents from their respective retrieval method (without embedding filtering or any LLM re-ranking), leveraging the inherent ranking/scoring provided by Semantic Scholar. This setup allows to assess the importance of both re-ranking steps. This structured setup enabled us to isolate the contribution of each component (retrieval method vs. re-ranking strategy) and evaluate whether they collectively brought key papers for novelty assessment into the top 10. We use o3-mini for evaluating novelty (Step 3) and gpt-4o for re-ranking (Step 2).

**Classification Analysis:** Table 2 shows that the complete system, which employs facet-based re-ranking in RankGPT, significantly outperforms its ablated variants in accuracy. The results demonstrate that methods relying only on keyword or snippet-based retrieval have much lower accuracy, and even alternate re-ranking strategies with a single embedding-based reranker or both embedding and general relevance RankGPT are insufficient to consistently bring key papers into the most relevant paper set. These findings show that combining facet-based reranking with embedding is critical for identifying the most relevant papers.

Table 2: Accuracy of predicting "not novel".

| Method | Accuracy |
|---|---|
| Complete System | 89.66% |
| - Relevance RankGPT | 13.79% |
| - Embedding Filtering | 10.34% |
| - Snippet Retrieval | 8.62% |
| - Keyword Retrieval | 5.17% |

**Analysis of the Most Relevant Papers:** Table 3 compares the top-10 most relevant papers retrieved under each ablation setting with those from the complete system. Approximately 30% of the papers differ when using either embedding-based or general relevance RankGPT. Additionally, notable rank shifts are observed between the facet-based and relevance-based LLM rerankers. In contrast, without the reranking steps, both snippet and keyword retrieval exhibit minimal overlap with the final system's top results, highlighting the importance of the reranker stage.

### 6.3 Qualitative Analysis

Table 4 in the Appendix shows examples from our training set, including an idea, its most relevant papers, and the corresponding expert reasoning.

Table 3: Comparing rank and overlap in retrieved papers with each variant to the complete system. *Overlap* indicates how many papers overlap on average with the complete system top-10 papers. *Rank Shift* measures the average absolute difference in rank positions (only among overlapping papers).

| **Method** | **Overlap** (↑) | **Rank Shift** (↓) |
|---|---|---|
| Relevance RankGPT | 7.97 | 0.67 |
| Embedding Filtering | 7.93 | 0.84 |
| Snippet Retrieval | 2.88 | 1.85 |
| Keyword Retrieval | 1.17 | 1.39 |

While assessing novelty, we add both the titles and abstracts of the most relevant papers for each idea.

Figures 2, 3, and 4 in the Appendix qualitatively compare novelty evaluations by AI Scientist, AI Researcher, and Idea Novelty Checker (ours) on two research ideas. Idea Novelty Checker provides concise justifications for its novelty decisions by referencing key similarities and differences with existing works. For example, in Example 1, it correctly identifies the idea as 'novel' by highlighting these aspects. In contrast, AI Researcher evaluates each paper individually, classfying an idea as 'not novel' if any paper is considered citable; but in our examples, none of the papers were flagged as citable despite sharing similar purposes, leading to a 'novel' classification. Due to space constraints, we show insights only from the first paper for each example. Figure 4 indicates that while AI Scientist's judgments generally align with the ground truth and offer actionable suggestions, it sometimes misinterprets the idea—as in Figure 3, where its focus shifts from the idea to the accompanying code.

### 6.4 Prompt Sensitivity

In our experiments with TextGrad, we investigated how specific prompt instructions influence an LLM's ability to classify the novelty of an idea. Figures in Appendices 5, 6, and 7 present the accuracy of various prompts optimized with TextGrad on our dataset (train=25, validation = 10, test = 32).

Prompts with both non-zero and zero validation accuracy included various instructions for evaluating the novelty of ideas, such as assessing the uniqueness of methods and their comparison to existing research. Through this prompt optimization process, we observed interesting ways in which LLMs may evaluate novelty, like considering historical context, frequency of similar studies, comparative analysis with existing works, examining arguments for both novel and non-novel perspectives. However, prompts without these specific instructions also influenced accuracy, suggesting the complexity of novelty evaluation with LLMs.

Notably, some prompts with similar instructions showed different performance on validation data. For example, both prompt 3 (accuracy = 0) and prompt 9 (accuracy = 0.6) include instructions to evaluate if the idea introduces unique methodologies, and how it compares to existing work. However, the difference in their performance suggests that subtle variations in wording and instruction framing can significantly impact the classification performance. It remains unclear why certain prompts perform better despite having similar instructions.

Our analysis highlights the LLM's sensitivity to prompt design when assessing novelty of an idea. Even minor variations in wording and structure can lead to substantial performance changes, emphasizing the need for careful prompt engineering and well-chosen in-context examples to guide the LLM for idea novelty evaluation.

## 7 Conclusion

In this work, we propose Idea Novelty Checker, a retrieval-augmented pipeline for evaluating the novelty of scientific ideas and generating literature-grounded rationales. Our formative study highlighted two main challenges in evaluating novelty: (1) retrieving the most relevant papers from a vast corpus, and (2) establishing a fixed notion of novelty due to its inherent subjectivity. To address the latter, we incorporate expert-annotated examples in our novelty checker where we consider an idea to be novel within a given topic domain if it (1) differs from all retrieved papers in at least one core facet—namely, purpose (a new objective), mechanism (a distinct technical approach), or evaluation (a distinct validation method); (2) uniquely combines these facets; or (3) applies them to a new application domain.

Our experiments on an expert-annotated dataset demonstrate that Idea Novelty Checker outperforms two well-known recent baselines, and our ablation studies confirm the importance of each component in our system. Furthermore, qualitative comparisons and analyses of prompt sensitivity provide additional insights into novelty evaluation.

## 8 Limitations & Future Work

While Idea Novelty Checker is superior in many aspects, it also has some limitations. For instance, due to context size constraints (with fifteen in-context examples for both *novel* and *not novel* categories), our analysis is restricted to the top 10 retrieved papers, which may disproportionately influence the overall novelty assessment. Additionally, our definition of novelty relies on expert annotations, and the same annotators who provided the in-context examples also classified the test ideas. This could potentially give our approach an advantage in understanding our view of novelty. Moreover, many of the ideas used for testing were generated by the same system (Radensky et al., 2024) that produced the in-context examples, although some ideas were sourced from OpenReview.

In future work, we aim to address these limitations by expanding the literature scope using tools such as DeepResearch and ScholarQA and further refining our novelty evaluation to view novelty as a continuum rather than binary classification.

## References

Abdelrahman Abdallah, Bhawna Piryani, Jamshid Mozafari, Mohammed Ali, and Adam Jatowt. 2025. Rankify: A comprehensive python toolkit for retrieval, re-ranking, and retrieval-augmented generation.

Faez Ahmed, Sharath Kumar Ramachandran, Mark D. Fuge, Samuel T. Hunter, and Scarlett R. Miller. 2018. Interpreting idea maps: Pairwise comparisons reveal what makes ideas novel. *Journal of Mechanical Design*.

Davide Baldelli, Junfeng Jiang, Akiko Aizawa, and Paolo Torroni. 2024. Twolar: A two-step llm-augmented distillation method for passage reranking. *ArXiv*, abs/2403.17759.

Joeran Beel, Min-Yen Kan, and Moritz Baumgart. 2025. Evaluating sakana's ai scientist for autonomous research: Wishful thinking or an emerging reality towards 'artificial research intelligence' (ari)?

Nicolas Bougie and Narimasa Watanabe. 2024. Generative adversarial reviews: When llms become the critic. *ArXiv*, abs/2412.10415.

Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. Solvent: A mixed initiative system for finding analogies between research papers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–21.

DaEun Choi, Sumin Hong, Jeongeon Park, John Joon Young Chung, and Juho Kim. 2024. Creative-connect: Supporting reference recombination for graphic design ideation with generative ai. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–25.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *ArXiv*, abs/2004.07180.

DeepResearch. Deepresearch. https://openai.com/index/introducing-deep-research/.

Steffen Eger, Andreas Rücklé, and Iryna Gurevych. 2019. Pitfalls in the evaluation of sentence embeddings. *ArXiv*, abs/1906.01575.

Matthew Freestone and Shubhra (Santu) Karmaker. 2024. Revisiting word embeddings in the llm era.

Jingtong Gao, Bo Chen, Xiangyu Zhao, Weiwen Liu, Xiangyang Li, Yichao Wang, Zijian Zhang, Wanyu Wang, Yuyang Ye, Shanru Lin, Huifeng Guo, and Ruiming Tang. 2024. Llm-enhanced reranking in recommender systems. *ArXiv*, abs/2406.12433.

Jos'e Manuel G'omez-P'erez, Andr'es Garc'ia-Silva, Rosemarie Leone, Mirko Albani, Moritz Fontaine, Charles Poncet, Leopold Summerer, Alessandro Donati, Ilaria Roma, and Stefano Scaglioni. 2022. Artificial intelligence and natural language processing and understanding in space: A methodological framework and four esa case studies. *ArXiv*, abs/2210.03640.

Tarun Gupta and Danish Pruthi. 2025. All that glitters is not novel: Plagiarism in ai generated research.

Ishika Joshi, Simra Shahid, Shreeya Venneti, Manushree Vasu, Yantao Zheng, Yunyao Li, Balaji Krishnamurthy, and Gromit Yeuk-Yin Chan. 2024. Coprompter: User-centric evaluation of llm instruction alignment for improved prompt engineering. *ArXiv*, abs/2411.06099.

Hyeonsu B Kang, David Chuan-En Lin, Nikolas Martelaro, Aniket Kittur, Yan-Ying Chen, and Matthew K Hong. 2024. Biospark: An end-to-end generative system for biological-analogical inspirations and ideation. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Hyeonsu B Kang, Xin Qian, Tom Hope, Dafna Shahaf, Joel Chan, and Aniket Kittur. 2022. Augmenting scientific creativity with an analogical search engine. *ACM Transactions on Computer-Human Interaction*, 29(6):1–36.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T.

Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler C. Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. 2023. The semantic scholar open data platform. *ArXiv*, abs/2301.10140.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xinxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Deli Zhao, Yu Rong, Tian Feng, and Li Bing. 2024. Chain of ideas: Revolutionizing research via novel idea development with llm agents. *ArXiv*, abs/2410.13185.

Manikanta Loya, Divya Sinha, and Richard Futrell. 2023. Exploring the sensitivity of LLMs' decision-making capabilities: Insights from prompt variations and hyperparameters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3711–3716, Singapore. Association for Computational Linguistics.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob N. Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *ArXiv*, abs/2408.06292.

Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. 2024. Ranked list truncation for large language model-based re-ranking. *ArXiv*, abs/2404.18185.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-vector models with textual guidance for fine-grained scientific document similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4453–4470.

Sheshera Mysore, Tim O'Gorman, Andrew McCallum, and Hamed Zamani. Csfcube–a test collection of computer science research articles for faceted query by example.

Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Lu Wang, and Tom Hope. 2021. Literature-augmented clinical outcome prediction. *ArXiv*, abs/2111.08374.

Baharan Nouriinanloo and Maxime Lamothe. 2024. Re-ranking step by step: Investigating pre-filtering for re-ranking with large language models. *ArXiv*, abs/2406.18740.

OpenReview. Openreview. `https://openreview.net/`.

Cyril Picard, Kristen M. Edwards, Anna C. Doris, Brandon Man, Giorgio Giannone, Md Ferdous Alam, and Faez Ahmed. 2023. From concept to manufacturing: Evaluating vision-language models for engineering design. *ArXiv*, abs/2311.12668.

Jason Portenoy, Marissa Radensky, Jevin D West, Eric Horvitz, Daniel S Weld, and Tom Hope. 2022. Bursting scientific filter bubbles: Boosting innovation via novel author discovery. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S. Weld. 2024. Scideator: Human-llm scientific idea generation grounded in research-paper facet recombination. *ArXiv*, abs/2409.14634.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

Serhad Sarica, Jianxi Luo, and Kristin L. Wood. 2019. Technology knowledge graph based on patent data. *Expert Syst. Appl.*, 142.

ScholarQA. Scholarqa. `https://scholarqa.allen.ai/chat/`.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

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

Arvind Srinivasan and Joel Chan. 2024. Improving selection of analogical inspirations through chunking and recombination. In *Proceedings of the 16th Conference on Creativity & Cognition*, pages 374–397.

Suzanne Stevenson and Paola Merlo. 2022. Beyond the benchmarks: Toward human-like lexical representations. *Frontiers in Artificial Intelligence*, 5.

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. 2024. Many heads are better than one: Improved scientific idea generation by a llm-based multi-agent system.

Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured generation and exploration of design space with large language models for human-ai co-creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–26.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023a. Is chatgpt good at search? investigating large language models as re-ranking agent. *ArXiv*, abs/2304.09542.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023b. Is chatgpt good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937.

Jianyou Wang, Kaicheng Wang, Xiaoyue Wang, Prudhviraj Naidu, Leon Bergen, and Ramamohan Paturi. 2023. Doris-mae: scientific document retrieval using multi-level aspect-based queries. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 38404–38419.

Kai Wang, Boxiang Dong, and Junjie Ma. 2019. Towards computational assessment of idea novelty. In *Hawaii International Conference on System Sciences*.

Wenxiao Wang, Lihui Gu, Liye Zhang, Yunxiang Luo, Yi Dai, Chen Shen, Liang Xie, Binbin Lin, Xiaofei He, and Jieping Ye. 2024. Scipip: An llm-based scientific paper idea proposer. *ArXiv*, abs/2410.23166.

Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, G. Wang, X. Liu, and Tie-Yan Liu. 2014. Rc-net: A general framework for incorporating knowledge into word representations. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*.

Yucheng Zhou, Tao Shen, Xiubo Geng, Chongyang Tao, Can Xu, Guodong Long, Binxing Jiao, and Daxin Jiang. 2022. Towards robust ranker for text retrieval. *ArXiv*, abs/2206.08063.

Table 4: **Expert-labeled examples from annotation study**

**Example 1**

**Idea:** Develop a **natural language processing classifier designed to improve scientific paper revisions** by automatically identifying and categorizing reviewer comments that are most likely to lead to substantial and actionable revisions. The system would be trained on a **manually-labeled dataset analysis** of scientific review comments and the corresponding paper edits, leveraging features such as linguistic cues, sentiment, and comment specificity to predict the likelihood of a comment being acted upon. This classifier could then be used to prioritize reviewer feedback, helping authors focus on the most impactful suggestions first.

**Most Relevant Papers:**

1. ARIES: A Corpus of Scientific Paper Edits Made in Response to Peer Reviews
2. Can large language models provide useful feedback on research papers?
3. A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications
4. arXivEdits: Understanding the Human Revision Process in Scientific Writing
5. Characterizing Text Revisions to Better Support Collaborative
6. Can We Automate Scientific Reviewing?
7. DeepReviewer: Collaborative Grammar & Innovation Neural Network for Automatic Paper Review
8. Aspect-based Sentiment Analysis of Scientific Reviews
9. Aspect-based sentiment analysis of online peer reviews and prediction of paper acceptance
10. ReviVal: Towards Automatically Evaluating the Informativeness of Peer Reviews

**Reasoning:** The idea is **novel** because it uniquely focuses on prioritizing reviewer comments for actionable revisions, which is not explicitly addressed in ARIES[1] or other related works like ReviVal [10].

**Example 2**

**Idea:** Develop a **systematic review-based framework** designed **to align LLM evaluation with human preferences**, ensuring that evaluation criteria are continuously refined based on comprehensive reviews of user feedback and emerging model behaviors. This framework will utilize **content analysis of user interactions and feedback** to identify patterns and areas of improvement. The effectiveness of this framework will be assessed through a **qualitative study** involving iterative cycles of user feedback and criteria refinement.

**Most Relevant Papers:**

1. EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria
2. Humanely: Human evaluation of LLM yield, using a novel web-based evaluation tool
3. Evaluation of Code Generation for Simulating Participant Behavior in Experience Sampling Method by Iterative In-Context Learning of a Large Language Model
4. Human-Centered Evaluation and Auditing of Language Models
5. Aligning Model Evaluations with Human Preferences: Mitigating Token Count Bias in Language Model Assessments
6. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences
7. Human-Centered Design Recommendations for LLM-as-a-judge
8. CheckEval: Robust Evaluation Framework using Large Language Model via Checklist
9. Discovering Language Model Behaviors with Model-Written Evaluations
10. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models

**Reasoning:** The idea is **not novel** because it closely resembles existing frameworks like EvalLM[1] and HumanELY[2], which already align LLM evaluations with human preferences using user-defined criteria and human feedback.

Figure 2: Two example ideas used as the basis for comparison in subsequent figures, evaluated by Idea Novelty Checker (Ours), AI Scientist, and AI Researcher.

**Example 1**

**Idea:** Develop a system that uses a *faceted author representation* of digital learning resource (DLR) creators based on their educational materials and inferred teaching personas. This system would aim *to support ubiquitous learning* by helping learners discover novel educators and materials that offer innovative perspectives. *Usability testing of learning resources* would be conducted to ensure the system enhances the learning experience by balancing relevance and novelty, thus boosting the accessibility and discoverability of diverse educational content.

**Most Relevant Papers:**

(1) Bursting Scientific Filter Bubbles: Boosting Innovation via Novel Author Discovery
(2) Bridger: Toward Bursting Scientific Filter Bubbles and Boosting Innovation via Novel Author Discovery
(3) Novel Algorithmic Recommendation Engine for Diverse Content Discovery
(4) ComLittee: Literature Discovery with Personal Elected Author Committees
(5) Explanations in Open User Models for Personalized Information Exploration
(6) AMiner: Mining Deep Knowledge from Big Scholar Data
(7) Similar researcher search in academic environments
(8) VeTo-web: A Recommendation Tool for the Expansion of Sets of Scholars
(9) From Who You Know to What You Read: Augmenting Scientific Recommendations with Implicit Social Networks
(10) DiscipLink: Unfolding Interdisciplinary Information Seeking Process via Human-AI Co-Exploration

**Expert-Labeled Class:** Novel

**Example 2**

**Idea:** Develop a **Co-Creative Interaction Framework** for LLM-assisted evaluations to **align llm evaluation with human preferences**. This framework will map out detailed interaction models between human evaluators and LLMs, including turn-taking, communication protocols, and iterative feedback loops. The framework will be evaluated through **qualitative study** involving user testing and thematic analysis of evaluator experiences, capturing how well the framework supports alignment with human preferences and identifying areas for iterative improvement.

**Most Relevant Papers:**

(1) EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria
(2) Human-Centered Design Recommendations for LLM-as-a-judge
(3) CheckEval: Robust Evaluation Framework using Large Language Model via Checklist
(4) Human-Centered Evaluation and Auditing of Language Models
(5) Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models
(6) Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences
(7) Evaluation of Code Generation for Simulating Participant Behavior in Experience Sampling Method by Iterative In-Context Learning of a Large Language Model
(8) Discovering Language Model Behaviors with Model-Written Evaluations
(9) LLM Comparator: Visual Analytics for Side-by-Side Evaluation of Large Language Models
(10) Do LLMs Meet the Needs of Software Tutorial Writers? Opportunities and Design Implications

**Expert-Labeled Class:** Novel

Figure 3: Reviews corresponding to idea 1 in Figure 2.

| Method | Review |
|---|---|
| **Scideator (Ours)** | The idea is **novel** because it introduces the concept of faceted author representation specifically for digital learning resource (DLR) creators, focusing on their educational materials and inferred teaching personas to support ubiquitous learning. This approach is distinct from related works like Bridger[1][2], which focus on scientific author discovery, and ComLittee[4], which supports authorcentric exploration in research. The proposed system aims to enhance the discoverability of diverse educational content, balancing relevance and novelty, which is not explicitly addressed in the related papers. |
| **AI Scientist** | THOUGHT: The proposed research presents an innovative approach to language models. The rough goal is to train a model to predict the next character in a sequence in order to be able to capture the semantic meaning of the entire sequence, which can be used in language models. This approach is not uncommon, though, as there are many areas where we must generate predictions of the next likely character, which can be used in order to understand the entire meaning of the text. This model aims to improve upon traditional models by locally training the model on the use of characters rather than words. Of particular note is the fact that the model uses a scaled gradient descent method which changes how the weights are updated, a method that is very new and, thus, is a key point of innovation. While that is extremely novel, a model that works on the character level is not. The primary new technique here is the use of a momentum of the gradient, which is used to improve learning. The model uses a technique similar to many current models. It is possible that it could be used with different parameters, though, to learn more or to learn in a more effective way than other models. This approach has been suggested many times over the years, but it is promising to pursue. It is, however, a novel approach to the existing technology.<br>However, we must consider the existing literature and how it might overlap with this proposition. I will now proceed to search for related publications that may be relevant in order to establish a definitive argument regarding the novelty of the proposed model. RESPONSE: “‘json\n\n "Query": "Character-level language models"\n\n“‘\n I believe this should identify any previous work or existing lines of research that are most similar to the proposed model. This will confirm whether this is actually a novel concept. My primary concern is understanding if this is a novel or an existing work, as this is a model that could potentially be used for many purposes. There is a good chance that it might have already been proposed. I hope that this search will bring more clarity to the argument.<br>**If decision is not made, default is: not novel.** |
| **AI Researcher** | Proposed Research Problem and Approach:<br>Project Proposal: The research problem is the limited accessibility and discoverability of innovative educational content due to the lack of effective representation and utilization of diverse teaching personas and educational materials of DLR creators. The approach involves developing a faceted author representation based on creators' educational materials and inferred teaching personas to enhance the discovery of novel educational content.<br>Paper Abstract: The research problem is the isolation of scientific research and information overload, which limit awareness and hinder innovation. The approach involves constructing a faceted representation of authors from their papers and inferred personas to balance relevance and novelty in discovering new scholars and their work.<br>Summary and Differences: - Project Proposal Summary: Enhancing the discovery of digital learning resources by developing a faceted author representation based on educational materials and teaching personas. - Paper Abstract Summary: Facilitating the discovery of new scholars and their work by constructing a faceted author representation from scientific papers and inferred personas to balance relevance and novelty. - Differences: - The project proposal focuses on digital learning resources and educational content, aiming to improve the learning experience. - The paper abstract focuses on scientific research and scholars, aiming to boost innovation and bridge diverse scientific communities.<br>No<br>........ *The comparison with other papers has been truncated due to space limitations. All papers in the most relevant set are marked as "Binary Judgment: No", which implies it **should not be cited as similar work to the idea.***<br>**Classification: novel** |

Figure 4: Reviews corresponding to idea 2 in Figure 2.

| Method | Review |
| --- | --- |
| **SCIDEATOR (Ours)** | The idea is **novel** because it introduces a CoCreative Interaction Framework specifically designed to align LLM evaluation with human preferences through detailed interaction models, including turntaking, communication protocols, and iterative feedback loops. This approach is distinct from related works like EvalLM[1], which focuses on iterative refinement of prompts, and EvaluLLM[2], which leverages LLMs as customizable judges but does not map out detailed interaction models. Additionally, CheckEval[3] and Prometheus 2[5] focus on evaluation frameworks and opensource evaluators, respectively, without the cocreative interaction aspect. The proposed framework's emphasis on qualitative studies and thematic analysis to capture evaluator experiences further differentiates it from existing methods. |
| **AI Scientist** | THOUGHT:\n The proposed idea involves designing a framework for co-creative interaction between human evaluators and Large Language Models (LLMs), focusing on aspects like turn-taking and communication protocols. It also emphasizes qualitative study and thematic analysis. This is a very specific angle that combines human-computer interaction and AI evaluation. While several existing papers discuss LLM evaluation frameworks, human-centered design, and human involvement in evaluation (e.g., EvalLM, EvaluLLM, and other papers from the results), the specific focus on co-creative interaction frameworks with detailed interaction models and iterative feedback loops may not have been explored thoroughly.\n\n Additionally, the idea looks at aligning evaluations with human preferences which is a nuanced approach. Therefore, a more targeted query to ascertain if any existing literature discusses similar co-creative frameworks or detailed interaction models involving human evaluators and LLMs is necessary. The current results do not seem to cover this specific angle in depth.<br>**Decision made: novel**<br>RESPONSE:\n“‘json\n\n "Query": "Co-creative interaction framework LLM evaluation"\n\n“‘ |
| **AI Researcher** | Proposed Research Problem and Approach:<br>- Project Proposal: - Problem: Current LLM evaluation methods often fail to align with human preferences. - Approach: Develop a Co-Creative Interaction Framework (CCIF) that includes detailed interaction models between human evaluators and LLMs, facilitating turn-taking, communication protocols, and iterative feedback loops to better align LLM evaluations with human preferences. - Paper Abstract: - Problem: Developers must iteratively revise prompts by evaluating outputs to diagnose weaknesses, which is effort-intensive and context-specific. - Approach: Present EvalLM, an interactive system for iteratively refining prompts by evaluating multiple outputs on user-defined criteria, using an LLM-based evaluator to provide feedback and improve prompts. - Differences: - The project proposal focuses on aligning LLM evaluations with human preferences through a co-creative interaction framework involving detailed interaction models. - The paper abstract focuses on refining LLM prompts through an interactive system that evaluates outputs based on user-defined criteria.<br>No<br>........ *The comparison with other papers has been truncated due to space limitations. All papers in the most relevant set are marked as "Binary Judgment: No", which implies it* ***should not be cited as similar work to the idea.***<br>**Classification: novel** |

Figure 5: **Performance trends of test accuracy across prompts during prompt optimization with TextGRAD.**
Highlighted text shows unique instructions used to evaluate the novelty of ideas. The final test accuracy was 0.78125, showing that none of the optimized prompts (1 to 12) improved over the original.

| Prompt Number | Validation Accuracy | Prompt Text |
|---|---|---|
| **Starting Prompt** | 0.8 | You are a classifier. Given a research idea and related papers, classify the idea as ‘novel’ or ‘not novel’ Respond in two lines:<br>Review: <This is the review you will generate after comparing idea with related papers.><br>Classification:<novel or not novel label according to your review/rationale> |
| **1** | 0 | You are a classifier. Given a research idea and related papers, classify the idea as ‘novel’ or ‘not novel’.<br>A novel idea introduces a unique, groundbreaking concept or approach not previously covered in the literature. A not novel idea reiterates or slightly modifies existing research. Consider the historical context and frequency of similar studies when making your classification. Focus on identifying unique, groundbreaking elements that differentiate the idea from existing research. Do not rely solely on keywords or the mention of a controlled setting to determine novelty. If the classification is ambiguous, indicate uncertainty and suggest a human review.<br>Respond with only the classification label: ‘novel’ or ‘not novel’. |
| **2** | 0.7 | You are a classifier. Classify the research idea as ‘novel’ or ‘not novel’ based on the related papers. Provide a brief review and directly state the classification.<br>Review: <This is the review you will generate after comparing the idea with related papers.> Classification: <novel or not novel label according to your review/rationale><br>Example: Review: The idea is unique as it combines adaptive interfaces with AI explanations, which is not covered in the provided papers. Classification: novel. |
| **3** | 0 | You are a classifier. Given a research idea and related papers, classify the idea as ‘novel’ or ’not novel.’<br>Respond in two lines:<br>Review: Provide a detailed review comparing the idea with related papers. Include specific examples and details from the referenced papers to justify your classification. Highlight both similarities and differences between the proposed idea and existing methodologies. Ensure your review is concise and precise, focusing on the main arguments.<br>Classification: Based on the review, classify the idea as ‘novel’ or ‘not novel’ according to the following criteria:<br>- Uniqueness of the approach<br>- Originality of the application<br>- Novelty of the results<br>Provide specific references or evidence from the papers mentioned to support your classification. Use assertive language to clearly convey your classification. |

Figure 6: contd. TextGrad Prompt Optimisation.

| Prompt Number | Validation Accuracy | Prompt Text |
|---|---|---|
| **4** | 0.7 | You are a classifier. Classify the research idea as 'novel' or 'not novel' based on the related papers. Respond concisely:<br>Review: <brief review> Classification: <novel or not novel> |
| **5** | 0 | You are a classifier. Given a research idea and related papers, classify the idea as 'novel' or 'not novel.'<br>Respond in two lines:<br>1. **Review**: Provide a detailed review comparing the idea with related papers. Include specific examples and details from the referenced papers to justify your classification. Highlight both similarities and differences between the proposed idea and existing methodologies. Ensure your review is concise and precise, focusing on the main arguments.<br>2. **Classification**: Based on the review, classify the idea as 'novel' or 'not novel' according to the following criteria:<br>- Uniqueness of the approach<br>- Originality of the application<br>- Novelty of the results<br>- Provide specific references or evidence from the papers mentioned to support your classification.<br>- Use assertive language to clearly convey your classification. |
| **7** | 0.3 | You are a classifier. Given a research idea and related papers, classify the idea as 'novel' or 'not novel' based on the following criteria:<br>1. **Definition of Novelty** : - A 'novel' idea introduces a unique, groundbreaking concept, methodology, or significant improvement over existing work. - A 'not novel' idea closely aligns with existing work without significant innovation.<br>2. **Contextual Instructions** : - If the idea involves common methodologies or well-known techniques, explicitly mention these aspects in your review. - Consider the historical context and frequency of similar studies when making your classification.<br>3. **Comparative Analysis** : - Compare the proposed idea with existing systems or technologies mentioned in the related papers. Highlight similarities to justify the classification.<br>4. **Evidence and Examples** : - Provide specific examples or evidence from the related papers that demonstrate the lack of novelty.<br>5. **Structured Format** : - Use a structured format with sections such as 'Introduction,' 'Evaluation Criteria,' 'Comparative Analysis,' and 'Conclusion' to present your review.<br>Respond in two lines:<br>Review: <This is the review you will generate after comparing the idea with related papers.> Classification: <novel or not novel label according to your review/rationale> |
| **8** | 0 | You are a classifier. Given a research idea and related papers, classify the idea as 'novel' or 'not novel'. Respond in two lines:<br>1. Review: Provide a detailed review that includes specific references to the related papers, highlighting similarities and differences. Include direct quotes or specific sections from the related papers that support your classification.<br>2. Classification: Use precise terminology to classify the idea as 'novel' or 'not novel' based on your review/rationale. Avoid vague terms and be specific in your justification. |

Figure 7: contd. TextGrad Prompt Optimisation.

| Prompt Number | Validation Accuracy | Prompt Text |
|---|---|---|
| **9** | 0.6 | You are a classifier. Given a research idea and related papers, classify the idea as 'novel' or 'not novel'.<br>Definition of Novel: An idea is 'novel' if it introduces a new concept, methodology, or significant improvement that is not already well-documented in the provided papers.<br>Criteria for Evaluation:<br>1. Uniqueness of the approach.<br>2. Combination of elements.<br>3. Presence of similar frameworks in the literature.<br>Instructions:<br>1. Compare the proposed idea's methodology, scope, and application with those described in the related papers.<br>2. Highlight specific aspects of the idea and compare them with the related papers.<br>3. Provide a detailed review based on the comparison.<br>4. Conclude with a classification of 'novel' or 'not novel' based on this comparison.<br>Respond in the following format: Review: <This is the review you will generate after comparing the idea with related papers.> Classification: <novel or not novel label according to your review/rationale> |
| **10** | 0.6 | You are a classifier. Given a research idea and related papers, classify the idea as 'novel' or 'not novel'. A novel idea introduces a unique, groundbreaking concept or approach not previously covered in the literature. A not novel idea reiterates or slightly modifies existing research. Consider the historical context and frequency of similar studies when making your classification.<br>Respond in two lines: Review: <one-sentence review> Classification: <novel or not novel><br>Ensure your response is concise and uses simple language. Avoid unnecessary details. |
| **11** | 0.6 | You are a classifier. Given a research idea and related papers, classify the idea as 'novel' or 'not novel'. Respond with: 1. Review: Provide a concise review in no more than two sentences, comparing the idea with related papers. Ensure your review includes a clear rationale for why the idea is classified as 'novel' or 'not novel'. Avoid using uncertain terms like 'appears' or 'seems' .<br>2. Classification: Use the term 'novel' or 'not novel' consistently based on your review.<br>Example:<br>Review: The proposed idea of developing a human-centric explainable AI system is novel because it uniquely combines explainable AI techniques with iterative improvement through human feedback and predictive models. Classification: novel |
| **12** | 0.5 | You are a classifier. Given a research idea and related papers, classify the idea as 'novel' or 'not novel'. Respond in two lines:<br>Review: <Provide a detailed review comparing the idea with related papers, including specific examples and reasons for your classification. Mention existing tools or research that cover similar capabilities.> Classification: <novel or not novel label according to your review/rationale. Use the term 'novel' consistently in both your review and classification. Ensure your response is detailed yet concise, avoiding unnecessary verbosity.> |