Is this Idea Novel? An Automated Benchmark for Judgment of Research Ideas

Tim Schopf[1,2] **and Michael Färber**[1]
[1]TU Dresden & ScaDS.AI Dresden/Leipzig, Germany
[2]National Institute of Informatics, Tokyo, Japan
{tim.schopf, michael.färber}@tu-dresden.de

Abstract

Judging the novelty of research ideas is crucial for advancing science, enabling the identification of unexplored directions, and ensuring contributions meaningfully extend existing knowledge rather than reiterate minor variations. However, given the exponential growth of scientific literature, manually judging the novelty of research ideas through literature reviews is labor-intensive, subjective, and infeasible at scale. Therefore, recent efforts have proposed automated approaches for research idea novelty judgment. Yet, evaluation of these approaches remains largely inconsistent and is typically based on non-standardized human evaluations, hindering large-scale, comparable evaluations. To address this, we introduce RINoBench, the first comprehensive benchmark for large-scale evaluation of research idea novelty judgments. It comprises 1,381 research ideas derived from and judged by human experts as well as nine automated evaluation metrics designed to assess both rubric-based novelty scores and textual justifications of novelty judgments. Using this benchmark, we evaluate several state-of-the-art large language models (LLMs) on their ability to judge the novelty of research ideas. Our findings reveal that while LLM-generated reasoning closely mirrors human rationales, this alignment does not reliably translate into accurate novelty judgments, which diverge significantly from human gold standard judgments—even among leading reasoning-capable models. Data and code available at: https://github.com/TimSchopf/RINoBench.

Keywords: research idea novelty judgment, evaluation benchmark, scientific discovery, llm-as-a-judge

1. Introduction

Judging the novelty of research ideas is fundamental to fostering scientific discovery and ensuring that new works meaningfully advance a field rather than reproducing existing results with minor variations that contribute little new insight. Hence, effective novelty judgment helps researchers identify unexplored directions, develop original contributions, and ultimately drive scientific progress. However, manually judging the novelty of a research idea requires a comprehensive review of previous work to determine whether the same or similar ideas have already been explored. With the rapid growth of scientific literature (Fortunato et al., 2018), this manual process has become increasingly challenging for researchers in terms of both time and cognitive effort. Moreover, novelty judgments are inherently subjective. Experts can often identify when two ideas are similar (Picard et al., 2025) but struggle to articulate what makes an idea truly novel (Shahid et al., 2025). In addition, such judgments are influenced by an individual's prior knowledge, intuition, and familiarity with the relevant literature (Ahmed et al., 2018; Picard et al., 2025). To address these challenges, automated approaches have been proposed to support and enhance research idea novelty judgments.

Recent work has used large language models (LLMs) to automatically judge the novelty of research ideas (Lu et al., 2024; Si et al., 2025; Li et al., 2025a; Su et al., 2025; Gottweis et al., 2025). However, these approaches do not ground their rationales in prior literature and struggle with subtle linguistic variation, leading to the misclassification of well-established ideas as novel (Beel et al., 2025; Gupta and Pruthi, 2025; Wang et al., 2025b). Furthermore, many LLM-based approaches restrict their outputs to binary classifications (novel vs. not novel) (Lu et al., 2024; Li et al., 2025a; Shahid et al., 2025; Su et al., 2025), overlooking the nuanced and gradual nature of novelty judgments. Moreover, most automated approaches provide only final pre-

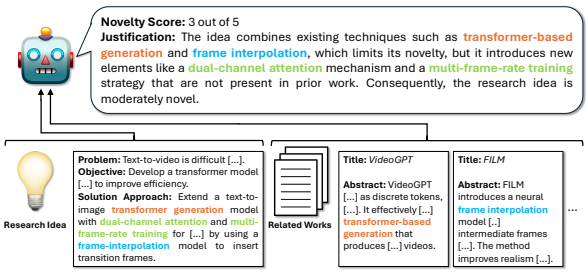

Figure 1: The task setup of RINoBench. Given a research idea and its related works, a model must judge the novelty of the idea according to a five-point rubric. In addition, the model must provide a textual justification for its judgment, grounded in a comparison between the proposed research idea and the related works.

dictions without offering interpretable explanations or justifications supporting their decisions. This lack of transparency reduces their practical utility, as researchers cannot review or learn from opaque judgments, hindering their ability to refine research ideas toward greater novelty. Finally, the fundamental limitations and differences between existing automated research idea judgment approaches make meaningful comparisons difficult. This problem is exacerbated by the fact that current evaluations of automated research idea judgment approaches are mainly based on non-standardized manual evaluations (Si et al., 2025; Gottweis et al., 2025, *inter alia*), hindering large-scale, systematic comparisons.

To address these limitations, we introduce the **R**esearch **I**dea **No**velty Judgment **Bench**mark (RINoBench), the first *comprehensive* and *reproducible* benchmark for the automatic evaluation of research idea novelty judgments. Using this benchmark, we conduct the first large-scale benchmarking study of current state-of-the-art LLMs, evaluating their ability to judge the novelty of research ideas. We reveal that *while LLMs often generate reasoning patterns similar to human experts, they fail to consistently translate these rationales into accurate novelty judgments*.

Our main contributions are:

- RINoBench, a comprehensive benchmark for systematically evaluating research idea novelty judgments, comprising **1,381 research ideas derived from and judged by human experts** as well as **nine automated evaluation metrics** designed to assess both rubric-based novelty scores and textual justifications of novelty judgments.
- A study investigating several state-of-the-art LLMs on their ability to judge the novelty of research ideas, involving a systematic analysis of the strengths and limitations of current state-of-the-art LLMs performing this task.

2. Related Work

Automated methods for judging novelty in scientific literature have advanced significantly in recent years. Early approaches measured novelty via atypical combinations of cited references (Uzzi et al., 2013), constructed historical co-occurrence matrices and derived journal vectors, where lower cosine similarity indicates greater novelty (Wang et al., 2017), and relied on lexical similarity (Wang et al., 2019; Sarica et al., 2020). However, such approaches are inherently limited in their ability to capture paraphrased, conceptually equivalent, or closely related ideas, as they reduce novelty to patterns of statistical co-occurrence or lexical overlap rather than accounting for their semantic relationships. Subsequent work using semantic embeddings (Gómez-Pérez et al., 2022) enhanced the ability to identify deeper conceptual relations, but remains constrained to surface-level semantic comparisons (Mysore et al., 2022). More recently, Wu et al. (2025b) combined human and LLM knowledge for novelty evaluation. In addition, retrieval-augmented LLM approaches have emerged as promising alternatives (Yang et al., 2024; Bougie and Watanabe, 2024; Lu et al., 2024; Radensky et al., 2024; Si et al., 2025; Liu et al., 2025; Su et al., 2025; Wang et al., 2025a; Baek et al., 2025; Zhang et al., 2025; Tang et al., 2025; Li et al., 2025a, *inter alia*). However, these approaches typically treat novelty judgment of research ideas as an intermediate step within a broader AI-assisted scientific discovery pipeline. As a result, and further exacerbated by the lack of automated benchmarks, these works either omit a dedicated and systematic evaluation of their novelty judgments or rely on costly and often small-scale human evaluations. Complementary to this, Wen et al. (2025) introduce a large-scale benchmark dataset designed to predict which of two research ideas performs better on a given set of benchmarks, but they do not address the task of novelty judgment. In addition, Wu et al. (2025a) examine which sections of research papers are most informative for novelty judgment. The work most closely related to ours provides the only publicly available evaluation dataset to date for judging the novelty of research ideas (Shahid et al., 2025). However, this dataset is limited to 51 manually annotated research ideas with binary labels (novel vs. not novel) and does not include any evaluation of textual novelty judgment justifications.

3. On the Notion of Novelty

Novelty is a fundamental concept in scientific research, which has been extensively characterized in existing literature. Arts et al. (2021) consider novelty as the uniqueness of specific knowledge elements, whereby the introduction of previously unknown elements indicates novel information. Further, Foster et al. (2021) define novelty as the extent to which a proposed contribution diverges from the existing scientific literature. Importantly, novelty is not limited to entirely new knowledge. An idea can also be considered novel if it represents a previously unobserved combination of known knowledge elements or applies them to new contexts (Boudreau et al., 2016; Shahid et al., 2025).

Closely related to novelty is the concept of *originality*. It refers to the generation of new ideas, methods, conclusions, or other valuable outputs that deviate from existing paradigms and can inspire further innovation (Shibayama and Wang, 2019; Hou et al., 2022). In practice, however, distinguish-

ing originality from novelty is challenging (Guetzkow et al., 2004), leading to the frequent interchangeable use of these terms (Wu et al., 2025a).

In essence, novelty, often used interchangeably with originality, is a fundamental driver of scientific progress, providing the foundation for both innovation and disruptive advances. While a novel idea frequently entails introducing a previously unseen element of knowledge, it can also emerge from a previously unexplored combination of existing knowledge in innovative ways.

4. Benchmark

RINoBench unifies approaches for judging the novelty of research ideas by formalizing the task, illustrated in Figure 1, as the process of comparing a proposed research idea with existing work to identify meaningful differences. Further, the task requires predicting a rubric-based novelty score (1–5) alongside a textual justification that grounds the judgment in related literature.

4.1. Data

Collecting a comprehensive dataset through dedicated workshops or user studies is prohibitively expensive and practically infeasible due to the complexity of the task. Human experts would need to generate novel research ideas, and other experts would then need to evaluate them. Both tasks impose a high cognitive load and require substantial domain expertise, meaning that each instance demands significant time and effort. Moreover, the pool of qualified participants is small, making recruitment difficult. Consequently, prior data collection efforts of this kind have been limited in scale, typically yielding only around 50 human-generated research ideas (Si et al., 2025; Shahid et al., 2025).

To overcome these limitations, we adopt a different strategy by leveraging publicly available data from OpenReview. Specifically, peer reviews from ICLR 2022 and ICLR 2023 provide a rich source of information: human experts have already submitted papers based on their research ideas, which have been explicitly evaluated by other human experts using rubric-based novelty scores and corresponding textual justifications. By processing and enriching this peer review data, we construct a high-quality dataset for studying research idea novelty judgment.

4.1.1. Data Collection & Processing

We collect all publicly available ICLR 2022 and ICLR 2023 submissions and their corresponding reviews from OpenReview, yielding 6,410 papers with associated reviewer feedback. Each submission was evaluated by approximately three expert reviewers, who rated the novelty of the research using a rubric-based numerical scale and provided brief textual justifications. Specifically, reviewers assessed two novelty dimensions: "*Technical Novelty and Significance*" and "*Empirical Novelty and Significance*". We use both dimensions in our dataset. Since human novelty judgments are inherently subjective and may vary substantially, we filter out all submissions where the maximum disagreement between reviewers exceeds one point within and across both novelty dimensions. This results in a filtered dataset of 3,535 paper submissions with high inter-reviewer agreement.

To obtain a single gold-standard novelty score for each paper, we average all reviewer scores across both novelty dimensions. The resulting decimal values, however, are difficult to interpret and predict. To address this, we transform them into whole numbers on a unified 1–5 scoring rubric by binning the averaged values into five intervals. This 1-5 scoring rubric, as shown in Table 1, offers an intuitive and standardized measure of novelty, featuring a clear midpoint, balanced polarity, and nuanced gradation, consistent with conventions commonly used in user studies.

Score	Degree of Novelty
1	The idea is not novel. All aspects already exist in prior work.
2	The idea is marginally novel. It represents only a minor variation of existing work.
3	The idea is somewhat novel. Aspects already exist in prior work. However, it might combine known approaches in new ways, apply them to new contexts, or propose incremental updates.
4	The idea is novel. It introduces new aspects not present in existing work.
5	The idea is highly innovative and novel. It is not present in existing work and potentially encourages new thinking or opens up new research directions.

Table 1: Novelty Judgment Rubric

Our next data processing step focuses on transforming submitted papers into concise research ideas by systematically identifying and reformulating their core ideas and contributions. Specifically, we provide an LLM with paper titles, abstracts, and reviewer summaries as context to distill the most salient information and produce structured and concise representations of the underlying research ideas. For this and all subsequent LLM-based data processing steps, we use the *GPT-OSS-120B* (OpenAI et al., 2025) model. In this step, the model is prompted to analyze the provided context, identify the key elements that define a paper's research idea, and output a structured JSON representation

capturing the core facets of the research idea. A central challenge in this process involves obtaining a reproducible and comparable representation of research ideas that contains all the information necessary for novelty understanding and judging. To this end, we adapt existing research idea templates (Si et al., 2025; Shahid et al., 2025) to structured presentations consisting of the following aspects:

- **Problem statement:** A detailed description of the core research problem(s) or question(s) addressed.
- **Objective:** A clear articulation of the research aim(s) or intended outcomes.
- **Solution approach:** A detailed description of the proposed methods or approaches designed to solve the problem and achieve the stated objectives.

Following this, we synthesize the individual reviewer justifications for their novelty scores into a single, coherent justification aligned with each gold-standard novelty judgment score. To achieve this, we provide an LLM with the reviewers' textual comments, the generated research idea, the assigned gold novelty score, and the novelty judgment rubric as context. The model is then prompted to identify the reviewers' arguments justifying their given novelty scores for a research idea and to integrate them into a unified, coherent justification that explains the rationale behind the gold novelty score.

Finally, after involving an LLM in earlier stages of data processing and accounting for their tendency to produce hallucinations and inaccuracies, our final step focuses on two objectives: enriching research ideas with relevant related works to enable grounded novelty judgments, and enforcing strict quality control to ensure that only high-quality samples are included in the final dataset. To this end, we first obtain related works by retrieving the titles and abstracts of publications cited in the introduction and related work sections of paper submissions. We extract the relevant paper sections from the PDF submissions using Nougat (Blecher et al., 2024) and obtain the works cited therein via Semantic Scholar (Kinney et al., 2025). Citations in other sections are ignored, as they typically include references to evaluation metrics or datasets that are irrelevant for novelty assessment. Next, we apply a quality filtering step. Research ideas are excluded if fewer than five related works are retrieved, typically due to missing indexing in Semantic Scholar. We then use an LLM to verify the formal correctness of the research idea, ensuring it is written in the first person, not as a summary of multiple reviews, and contains no explicit numerical novelty scores. Additionally, we assess whether all arguments in the synthesized novelty justifications are fully grounded in the corresponding research ideas and related works. Ungrounded justifications may arise not only from LLM-induced hallucinations during the synthesis of reviewer arguments, but also when reviewers use arguments derived from related works that are not cited in the corresponding submitted paper, or when related works are not available via Semantic Scholar. To assess the grounding of these justifications, we use an LLM to verify that every argument in the textual novelty justification is grounded in either the research idea or in the retrieved related works. In particular, the LLM extracts all arguments from the novelty justification pertaining to the novelty aspects of the research idea and verifies whether each novelty argument is grounded in the idea. In addition, the LLM extracts all arguments involving comparisons to related work and ensures that each argument is grounded in at least one retrieved title and abstract. Only samples with a formally correct research idea and a fully grounded novelty justification are included in the final data set.

4.1.2. Dataset

Our final dataset consists of 1,381 research ideas, each paired with rubric-based novelty scores, corresponding textual novelty judgment justifications, and an average of 25.23 titles and abstracts of related works. We perform a stratified 80:20 train-test split on the dataset, yielding the data distribution presented in Table 2.

Novelty Score	#training	#test	$\sum$
1	60	15	75
2	239	60	299
3	349	87	436
4	322	81	403
5	134	34	168
$\sum$	1,104	277	1,381

Table 2: Class distributions within RINoBench.

4.2. Evaluation Metrics

This section outlines the metrics used in RINoBench to evaluate both the predicted novelty scores and the generated textual justifications for the novelty judgments.

4.2.1. Novelty Score Metrics

$\mathbf{F_1}$ We treat novelty score prediction as a classification task and use F_1 as the primary evaluation metric. Specifically, we employ macro-averaged F_1 to ensure that each degree of novelty is weighted equally. This approach disregards the actual category imbalance and treats all categories as equally important, consistent with how they are valued in practice. Additionally, we report the F_1 scores for each novelty category individually to evaluate how

a model performs across different categories, identifying where it performs particularly well or poorly.

Mean Absolute Error (MAE) Since novelty scores in this task are not limited to discrete categories but also represent rubric-based values, MAE allows us to evaluate the magnitude of deviation between predicted and gold scores. This provides insight not only into whether correct novelty scores are predicted, but also into how far the predictions diverge from the gold standard. By evaluating the average distance of predictions from the gold scores, we can determine if a model's predictions are reasonably close or significantly misaligned with the expected outcomes.

4.2.2. Justification Metrics

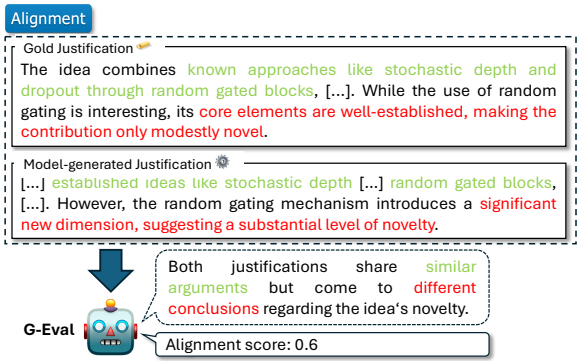

Figure 2: Evaluation of justification *alignment* for novelty judgments using the G-Eval framework, which produces textual reasoning and a numerical score. We use only the numerical score for evaluation.

Alignment Evaluating the alignment of novelty judgment justifications is essential to ensure that a model's decision-making process mirrors human-like judgment, both in terms of logic and argumentation. This metric evaluates whether the reasoning behind a predicted novelty judgment is consistent with the reasoning in the gold standard. Specifically, it verifies whether a model-generated justification follows the same line of argumentation, presents similar supporting arguments, and reaches the same conclusion as the human gold justification. As illustrated in Figure 2, we utilize the G-Eval framework (Liu et al., 2023) to prompt an LLM for alignment evaluation, generating an alignment score that ranges from 0 (worst) to 1 (best).

Known Aspects Recall We measure the extent to which arguments in the gold novelty justification that pertain to "*known aspects*" (see Figure 3) are captured in a model-generated justification. Following the FActScore approach (Min et al., 2023), an LLM first extracts all arguments from both the model-generated and gold justifications. It then verifies whether each argument extracted from the model-generated justification is supported by the gold justification. The final metric is computed as:

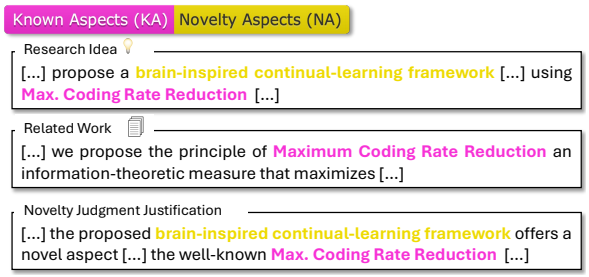

Figure 3: Example illustrating *known* and *novelty* aspects in novelty judgment justifications. *Known aspects* refer to elements in a justification that highlight already established concepts or findings from previous work in a research idea. *Novelty aspects* denote elements in a justification that highlight new contributions of a research idea, which do not exist in prior work.

$$\text{Recall}_{\text{Args.}} = \begin{cases} \min\left(1, \frac{N_{\text{supp}}}{N_{\text{gold}}}\right) & \text{if } N_{\text{gold}} > 0, \\ 0 & \text{if } N_{\text{gold}} = 0. \end{cases} \quad (1)$$

where N_{supp} is the number of model-generated known-aspect arguments supported by the gold justification, and N_{gold} is the total number of known-aspect arguments in the gold justification. Figure 4 provides an illustrated example.

Novelty Aspects Recall Analogous to above, we measure the extent to which arguments in the gold novelty justification related to "*novelty aspects*" (see Figure 3) are captured by a model-generated justification. This is done using the same LLM-based argument extraction and validation approach as in *Known Aspects Recall*, with the metric computed using Equation 1. In this case, N_{supp} represents the number of model-generated novelty-aspect arguments supported by the gold justification, while N_{gold} denotes the total number of novelty-aspect arguments in the gold justification.

Additional Known Aspects Ratio We measure the extent to which a model generates additional known-aspect arguments, which are not present in the gold justification but grounded in the associated related works. To evaluate this, we again use the same LLM-based argument extraction and validation approach as in *Known Aspects Recall*. Additionally, the LLM verifies whether the known-aspect

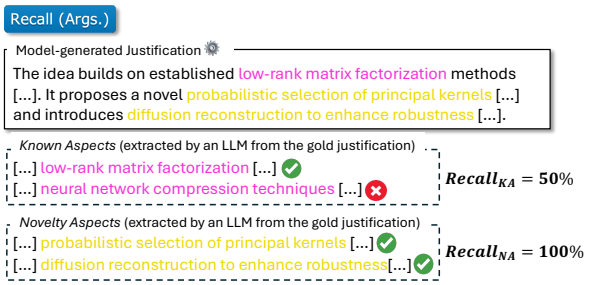

Figure 4: Example of *Known Aspects Recall* and *Novelty Aspects Recall* for evaluation of novelty judgment justifications.

arguments extracted from the model-generated justification are grounded in the corresponding related works. The metric is then computed as:

$$\text{Ratio}_{\text{Additional}} = \frac{N_{\text{additional}}}{\max(N_{\text{gold}}, 1)} \quad (2)$$

where $N_{\text{additional}}$ is the number of model-generated known-aspect arguments unsupported by the gold justification but grounded in the related works and N_{gold} is the total number of known-aspect arguments in the gold justification. Figure 5 provides an illustrated example.

Additional Novelty Aspects Ratio Similarly to above, we assess the extent to which a model generates additional novelty-aspect arguments that are not present in the gold justification but are grounded in the respective research idea. This is achieved using the same LLM-based argument extraction and validation approach as in *Known Aspects Recall*, with an added LLM-based step to verify whether the extracted novelty-aspect arguments from the model-generated justification are grounded in the corresponding research idea. The metric is computed using Equation 2, where $N_{\text{additional}}$ is the number of model-generated novelty-aspect arguments unsupported by the gold justification but grounded in the research idea and N_{gold} is the total number of novelty-aspect arguments in the gold justification.

Known Aspects Hallucination Rate We quantify the extent to which model-generated justifications contain hallucinated known-aspect arguments (i.e., those not supported by any of the corresponding related works). To this end, we adopt the same LLM-based argument extraction and validation approach used in *Known Aspects Recall* and compute the metric as:

$$\text{Hall. Rate} = \begin{cases} \frac{N_{\text{hallucinated}}}{N_{\text{generated}}} & \text{if } N_{\text{generated}} > 0, \\ 0 & \text{if } N_{\text{generated}} = 0. \end{cases} \quad (3)$$

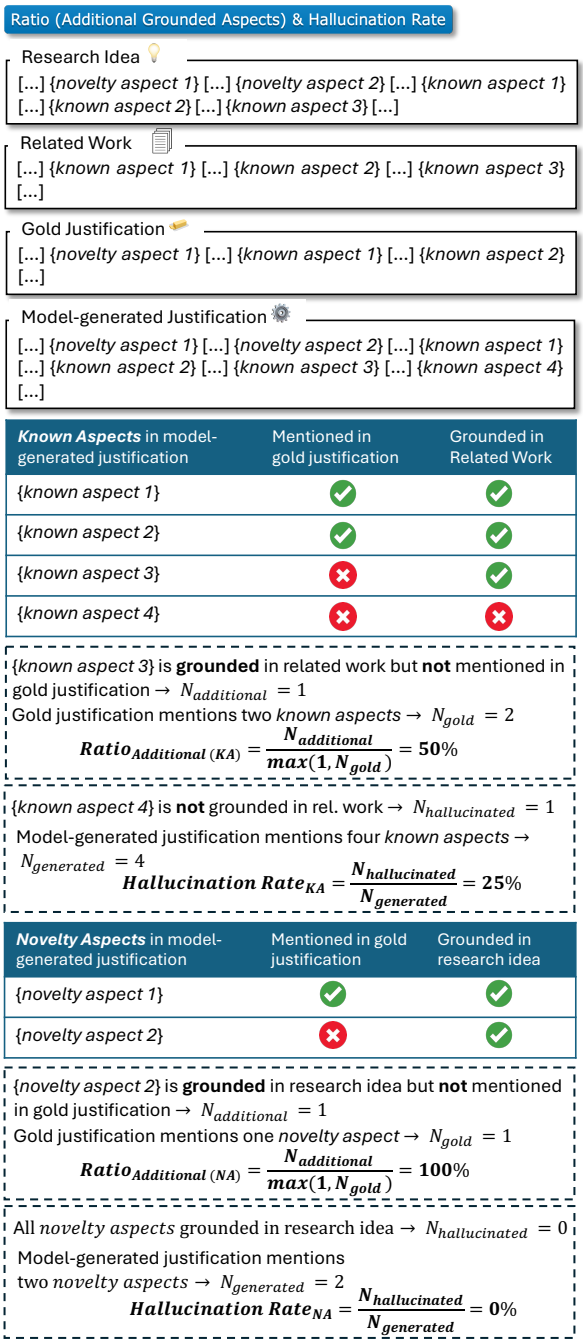

Figure 5: Example evaluation of a model-generated novelty judgment justification using *Additional Ratio* and *Hallucination Rate* for *known aspects* and *novelty aspects* respectively.

where $N_{\text{hallucinated}}$ is the number of model-generated known-aspect arguments unsupported by any of the corresponding related works and $N_{\text{generated}}$ is the total number of known-aspect arguments in the model-generated justification. Figure 5 provides an illustrated example.

Novelty Aspects Hallucination Rate Similarly, we quantify the extent to which model-generated

justifications contain hallucinated novelty-aspect arguments (i.e., those unsupported by the corresponding research idea). Using the same approach as in *Known Aspects Hallucination Rate*, we compute this metric with Equation 3, where $N_{\text{hallucinated}}$ represents the number of model-generated novelty-aspect arguments unsupported by the research idea and $N_{\text{generated}}$ is the total number of novelty-aspect arguments in the model-generated justification.

For all LLM-based evaluations, it is crucial to use a high-performing model to ensure the accuracy of the various metrics. Accordingly, we use the GPT-4.1 (OpenAI, 2025b) model for all LLM-based evaluations. Additionally, each presented justification metric is computed per individual sample. To provide a comprehensive evaluation in RINoBench, we average the computed justification metric scores across multiple samples.

5. Benchmarking LLMs as Judges of Research Idea Novelty

In this section, we present a benchmarking study examining several state-of-the-art LLMs on their ability to judge the novelty of research ideas. To this end, we follow recent works that frame the novelty judgment of research ideas as zero-shot task by directly giving the review criteria and prompting LLMs for a final score (Yang et al., 2024; Lu et al., 2024; Si et al., 2025; Li et al., 2025b; Baek et al., 2025). Accordingly, we instruct each LLM to perform the RINoBench task illustrated in Figure 1 to generate a numerical novelty score and a textual justification. Thereby, the LLM is provided with the novelty judgment rubric detailed in Table 1, alongside a research idea and its related works. The LLM is then asked to analyze the research idea, identify its key contributions, and compare it to the provided related works. Based on this analysis and comparison, the model is finally tasked to generate a suitable novelty score according to the rubric, accompanied by a brief justification explaining its reasoning for the predicted score. The exact instructions are shown in Figure 6.

For this study, we select a diverse set of LLMs, encompassing a range of sizes and reasoning capabilities. Specifically, we include non-reasoning models including *Llama-3.1-8B* (Grattafiori et al., 2024), *Llama-3.3-70B* (Grattafiori et al., 2024), and *Llama-4-Scout-17B-16E* (Meta, 2025), as well as reasoning-capable models including *DeepSeek-R1* (DeepSeek-AI et al., 2025), *GPT-OSS-120B* (OpenAI et al., 2025), *o3* (OpenAI, 2025c), and *GPT-5* (OpenAI, 2025a). This selection enables a comprehensive evaluation of model performance across different architectures and reasoning abilities.

Novelty Judgment Performance Table 3 shows the evaluation results. We observe that all models show very low novelty judgment abilities, with none of them achieving significant F_1 scores and the highest macro-average being 17.2. Notably, no model successfully predicted the novelty category 1, as indicated by the 0.0 F_1 scores for this category across all models. This suggests a strong bias against predicting ideas as "*not novel*", indicating that the models tend to avoid judging ideas as lacking novelty altogether. Interestingly, novelty categories 2 and 5 are occasionally predicted correctly, but the models predominantly predicted novelty categories 3 and 4. This suggests that the models tend to avoid assigning extreme values of novelty (such as "*marginally novel*" or "*highly novel and innovative*"). Instead, they consistently attempt to find at least some aspect of novelty in a research idea, even if it is not present. The MAE values are relatively low and consistently hover around 1, indicating that while the models may often make errors in their novelty judgments, these do not deviate drastically from the gold standard.

Quality and Coverage of Justifications The justification metrics provide interesting insights into how the models substantiate their novelty judgments. It is noteworthy that all models exhibit relatively high recall, implying that the LLMs frequently use arguments similar to those found in the human-annotated gold standard justifications. This finding is consistent with the results of Afzal et al. (2026), who reported high alignment between LLM and human novelty reasoning. Moreover, the high additional ratios (Add. Ratio), often exceeding 100%, suggest that the LLMs tend to generate more elaborate justifications than humans and frequently find more arguments to justify their novelty predictions. This may be because, for humans, one or two well-chosen arguments are often sufficient to judge the novelty of a research idea, while LLMs appear to strive to provide a comprehensive set of arguments, likely in an attempt to satisfy the user.

Hallucinations and Judgment–Justification Gap Despite generating many arguments, the hallucination rate is low across all models, suggesting that the models' justifications are largely grounded in the provided context. This indicates that, while the models may sometimes over-elaborate in their novelty judgment justifications, the arguments they generate are mostly reliable and supported by evidence. This stands in contrast to their novelty score predictions, which are more dissimilar from the human-annotated gold novelty scores, pointing to a gap in the models' ability to accurately judge the novelty of research ideas, even if their justifications seem to contain plausible arguments.

Model	Novelty Score Metrics							Justification Metrics						
	$\mathbf{F_1}$						MAE	ALI	Recall		Add. Ratio		Hall. Rate	
	Macro	1	2	3	4	5			KA	NA	KA	NA	KA	NA
Non-Reasoning Models														
Llama-3.1-8B	14.6	0.0	0.0	26.2	41.3	**5.4**	1.00	0.58	85.5	75.3	62.8	111.6	4.2	3.4
Llama-3.3-70B	9.5	0.0	0.0	2.2	45.0	0.0	1.04	0.55	88.9	78.3	86.3	115.3	1.1	1.4
Llama-4-Scout	13.0	0.0	0.0	17.1	42.7	5.1	1.01	0.58	89.8	81.9	89.0	**120.3**	**0.0**	1.1
Reasoning Models														
GPT-OSS-120B	14.6	0.0	3.0	30.1	40.7	0.0	0.96	0.64	88.1	77.8	79.0	92.4	0.9	0.5
DeepSeek-R1	12.3	0.0	0.0	16.1	45.6	0.0	0.99	0.67	87.8	81.1	115.7	112.4	0.6	**0.2**
o3	16.2	0.0	5.6	**35.6**	**39.7**	0.0	**0.93**	**0.72**	**90.4**	85.6	**139.9**	74.0	1.3	1.7
GPT-5	**17.2**	0.0	**16.7**	32.2	37.3	0.0	**0.93**	0.71	89.9	**85.7**	122.1	91.8	0.6	0.5

Table 3: Evaluation results of novelty judgments on the RINoBench test set. As described in Section 4.2.1, the reported novelty score metrics include F_1 macro averaged and for each rubric category (1-5), as well as *MAE*. Further, as outlined in Section 4.2.2, the justification metrics include *Alignment (ALI)*, as well as *Recall*, *Additional Ratio* (in %), and *Hallucination Rate* (in %) for *Known Aspects (KA)* and *Novelty Aspects (NA)* respectively.

Reasoning vs. Non-Reasoning Models When comparing model performance, we observe that reasoning models outperform their non-reasoning counterparts, albeit by a small margin. The GPT-5 model achieves the highest macro-averaged F_1 score with 17.2, closely followed by o3 (16.2). These models, designed for more complex reasoning tasks, outperform the non-reasoning models, which generally exhibit lower F_1 scores and demonstrate worse novelty judgment abilities. This indicates that incorporating additional reasoning and deeper thinking during the generation process helps models make more accurate judgments regarding the novelty of research ideas.

Takeaway Overall, the results show that the LLM-generated novelty judgment justifications closely align with those of human experts, whereas the predicted novelty judgment scores diverge substantially from the human-annotated gold scores, pointing to a clear gap: *although LLMs can generate plausible and well-supported novelty justifications, they fail to translate their reasoning into accurate novelty judgments*. Further, the models tend to avoid extreme predictions and avoid judging ideas as *not novel at all* nor *highly novel and innovative*. Instead, they constantly strive to find a middle ground. Additionally, while the models' justifications are often grounded in the provided context, they tend to be more elaborate than human justifications, reflecting a difference in how humans and models approach novelty judgment. Despite these differences, the models' predictions are only slightly different from human annotations (see MAE scores), suggesting that their beliefs about novelty do not differ fundamentally, but are instead somewhat inaccurate or imprecise.

6. Conclusion

This work introduces RINoBench, the first automated benchmark for evaluating novelty judgments of research ideas. It includes 1,381 research ideas derived from and judged by human experts. Further, the benchmark comprises nine automatic metrics to assess the accuracy of predicted novelty scores and to compare model-generated textual justifications with human-authored gold justifications. Our work bridges the gap between current, largely manual and incomparable human evaluations, towards reproducible and comparable evaluations.

Further, we investigate the capability of several state-of-the-art LLMs to judge the novelty of research ideas. Our findings reveal that current LLMs face substantial challenges in accurately judging the novelty of research ideas. Notably, while LLMs refrain from judging ideas as completely lacking novelty, they tend to seek a middle ground, aiming to attribute at least some degree of novelty while simultaneously avoiding judgments of high novelty and innovation. Interestingly, as indicated by the strong correspondence between LLM-generated and human-authored justifications, LLM reasonings align closely with human rationales for research idea judgment. However, this alignment does not translate to the LLM-predicted novelty scores, which diverge considerably from human-assigned scores. Finally, while all LLMs examined exhibit difficulties in effective novelty judgment, our experiments indicate that reasoning-capable models consistently outperform non-reasoning ones, suggesting that longer thinking and deeper engagement with the input and task instructions improve LLM-based novelty judgment of research ideas.

7. Limitations

RINoBench is derived exclusively from ICLR 2022 and 2023 submissions, limiting its domain, epistemic, and cultural diversity. Because it reflects a single conference ecosystem centered on machine learning, the benchmark captures the reviewing norms and novelty criteria of that community. Fields with different epistemological assumptions and evaluation practices—such as many areas in the humanities and social sciences—are not represented. Consequently, the novelty dimensions emphasized in RINoBench, focused on technical or methodological innovation, may underrepresent theoretical or discovery-oriented contributions. Findings based on RINoBench should therefore be interpreted within this specific context and validated on broader, more heterogeneous datasets.

The dataset relies on peer review scores, which are inherently subjective and shaped by disciplinary conventions. While this enables modeling real-world novelty judgments, it also operationalizes a particular reviewing culture rather than a universal notion of novelty, potentially reflecting individual or systemic biases.

Because the data originates from predominantly English-language OpenReview submissions, linguistic and rhetorical conventions may influence how novelty is expressed and assessed. Furthermore, although LLMs are used to extract structured ideas and synthesize justifications with careful filtering, they may introduce hallucinations, normalization effects, or discourse-sensitive distortions.

Finally, RINoBench focuses solely on novelty and does not capture other dimensions of research quality, such as rigor, significance, or reproducibility.

8. Ethics Statement

All data in RINoBench is derived from publicly available OpenReview submissions and papers indexed by Semantic Scholar. No private reviewer or author information is included. We emphasize that the dataset is intended for research and educational purposes only. Users should not use models trained on this data to make formal or high-stakes judgments of research ideas, as novelty judgments are inherently subjective and context-dependent.

9. Broader Impact

RINoBench is designed to advance AI-assisted scientific discovery by enabling models to reason about and explain novel contributions in research. It provides a valuable resource for researchers, educators, and students to understand and teach research idea novelty judgment and serves as a benchmark for developing models with explainable reasoning capabilities.

At the same time, automated predictions of research idea novelty should not replace human expert judgment. The dataset reflects inherently subjective human opinions and may contain biases from peer reviews. Models developed on RINoBench should be intended as tools to support, rather than replace human judgments of research ideas.

10. Acknowledgements

The authors acknowledge the financial support by the Federal Ministry of Research, Technology and Space of Germany and by Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus in the programme Center of Excellence for AI-research „Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig“, project identification number: ScaDS.AI.

The first author was supported by a scholarship of the German Academic Exchange Service (DAAD).

We used AI-based assistance tools to support language editing, minor formatting, and coding tasks. These tools did not contribute to the intellectual content or scientific conclusions. All content was reviewed by the authors, who assume full responsibility for the publication.

11. Bibliographical References

Osama Mohammed AfzalPreslav Nakov, Tom Hope, Iryna Gurevych, . 2026. Beyond "not novel enough": Enriching scholarly critique with llm-assisted feedback.

Faez AhmedSharath Kumar Ramachandran, Mark Fuge, Samuel Hunter, Scarlett Miller, . 2018. Interpreting idea maps: Pairwise comparisons reveal what makes ideas novel. *Journal of Mechanical Design*, 141(2):021102.

Sam ArtsJianan Hou, Juan Carlos Gomez, . 2021. Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*, 50(2):104144.

Jinheon BaekSujay Kumar Jauhar, Silviu Cucerzan, Sung Ju Hwang, . 2025. ResearchAgent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the Nations of the*

Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6709–6738, Albuquerque, New Mexico. Association for Computational Linguistics.

Joeran BeelMin-Yen Kan, Moritz Baumgart, . 2025. Evaluating sakana's ai scientist: Bold claims, mixed results, and a promising future? *SIGIR Forum*, 59(1):1–20.

Lukas BlecherGuillem Cucurull, Thomas Scialom, Robert Stojnic, . 2024. Nougat: Neural optical understanding for academic documents. In *The Twelfth International Conference on Learning Representations*.

Kevin J. BoudreauEva C. Guinan, Karim R. Lakhani, Christoph Riedl, . 2016. Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science*, 62(10):2765–2783.

Nicolas BougieNarimasa Watanabe, . 2024. Generative adversarial reviews: When llms become the critic.

DeepSeek-AIDaya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 180 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Santo FortunatoCarl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, Albert-László Barabási, . 2018. Science of science. *Science*, 359(6379):eaao0185.

Jacob G. FosterFeng Shi, James Evans, . 2021. Surprise! measuring novelty as expectation violation.

Juraj GottweisWei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, and 14 others. 2025. Towards an ai co-scientist.

Aaron GrattafioriAbhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, and 541 others. 2024. The llama 3 herd of models.

Joshua GuetzkowMichèle Lamont, Grégoire Mallard, . 2004. What is originality in the humanities and the social sciences? *American Sociological Review*, 69(2):190–212.

Tarun GuptaDanish Pruthi, . 2025. All that glitters is not novel: Plagiarism in AI generated research. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25721–25738, Vienna, Austria. Association for Computational Linguistics.

José Manuel Gómez-PérezAndrés García-Silva, Rosemarie Leone, Mirko Albani, Moritz Fontaine, Charles Poncet, Leopold Summerer, Alessandro Donati, Ilaria Roma, Stefano Scaglioni, . 2022. Artificial intelligence and natural language processing and understanding in space: A methodological framework and four esa case studies.

Jianhua HouDongyi Wang, Jing Li, . 2022. A new method for measuring the originality of academic articles based on knowledge units in semantic networks. *Journal of Informetrics*, 16(3):101306.

Rodney KinneyChloe 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 Graham, Fangzhou Hu, Regan Huff, and 28 others. 2025. The semantic scholar open data platform.

Long LiWeiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Yu Rong, Deli Zhao, Tian Feng, Lidong Bing, . 2025a. Chain of ideas: Revolutionizing research via novel idea development with LLM agents. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8971–9004, Suzhou, China. Association for Computational Linguistics.

Ruochen LiTeerth Patel, Qingyun Wang, Xinya Du, . 2025b. Mlr-copilot: Autonomous machine learning research based on large language models agents.

Yan LiuZonglin Yang, Soujanya Poria, Thanh-Son Nguyen, Erik Cambria, . 2025. Harnessing large language models for scientific novelty detection.

Yang LiuDan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, Chenguang Zhu, . 2023. G-eval:

NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Chris LuCong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, David Ha, . 2024. The ai scientist: Towards fully automated open-ended scientific discovery.

Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. Accessed: 2025-10-24.

Sewon MinKalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, Hannaneh Hajishirzi, . 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Sheshera MysoreArman Cohan, 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, Seattle, United States. Association for Computational Linguistics.

OpenAI:, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, and 107 others. 2025. gpt-oss-120b & gpt-oss-20b model card.

OpenAI. 2025a. Gpt-5 system card. Accessed: 2025-10-24.

OpenAI. 2025b. Introducing gpt-4.1 in the api | openai. Accessed: 2025-10-24.

OpenAI. 2025c. Openai o3 and o4-mini system card. Accessed: 2025-10-24.

Cyril PicardKristen M. Edwards, Anna C. Doris, Brandon Man, Giorgio Giannone, Md Ferdous Alam, Faez Ahmed, . 2025. From concept to manufacturing: evaluating vision-language models for engineering design. *Artificial Intelligence Review*, 58(9):288.

Marissa RadenskySimra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, Daniel S. Weld, . 2024. Scideator: Human-llm scientific idea generation grounded in research-paper facet recombination.

Serhad SaricaJianxi Luo, Kristin L. Wood, . 2020. Technet: Technology semantic network based on patent data. *Expert Systems with Applications*, 142:112995.

Simra ShahidMarissa Radensky, Raymond Fok, Pao Siangliulue, Daniel S Weld, Tom Hope, . 2025. Literature-grounded novelty assessment of scientific ideas. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 96–113, Vienna, Austria. Association for Computational Linguistics.

Sotaro ShibayamaJian Wang, . 2019. Measuring originality in science. *Scientometrics*, 122(1):409–427.

Chenglei SiDiyi Yang, Tatsunori Hashimoto, . 2025. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *The Thirteenth International Conference on Learning Representations*.

Haoyang SuRenqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, Philip Torr, Bowen Zhou, Nanqing Dong, . 2025. Many heads are better than one: Improved scientific idea generation by a LLM-based multi-agent system. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28201–28240, Vienna, Austria. Association for Computational Linguistics.

Jiabin TangLianghao Xia, Zhonghang Li, Chao Huang, . 2025. AI-researcher: Autonomous scientific innovation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Brian UzziSatyam Mukherjee, Michael Stringer, Ben Jones, . 2013. Atypical combinations and scientific impact. *Science*, 342(6157):468–472.

Jian WangReinhilde Veugelers, Paula Stephan, . 2017. Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46(8):1416–1436.

Kai WangBoxiang Dong, Junjie Ma, . 2019. Towards computational assessment of idea novelty. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Wenxiao WangLihui Gu, Liye Zhang, Yunxiang Luo, Yi Dai, Chen Shen, Liang Xie, Binbin Lin, Xiaofei He, Jieping Ye, . 2025a. Scipip: An llm-based scientific paper idea proposer.

Xin WangJiyao Liu, Yulong Xiao, Junzhi Ning, Lihao Liu, Junjun He, Botian Shi, Kaicheng Yu, . 2025b. The-tree: Can tracing historical evolution enhance scientific verification and reasoning?

Jiaxin WenChenglei Si, Chen Yueh-Han, He He, Shi Feng, . 2025. Predicting empirical AI research outcomes with language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Wenqing WuChengzhi Zhang, Tong Bao, Yi Zhao, . 2025a. Sc4anm: Identifying optimal section combinations for automated novelty prediction in academic papers. *Expert Systems with Applications*, 273:126778.

Wenqing WuChengzhi Zhang, Yi Zhao, . 2025b. Automated novelty evaluation of academic paper: A collaborative approach integrating human and large language model knowledge. *Journal of the Association for Information Science and Technology*, 76(11):1452–1469.

Zonglin YangXinya Du, Junxian Li, Jie Zheng, Soujanya Poria, Erik Cambria, . 2024. Large language models for automated open-domain scientific hypotheses discovery. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13545–13565, Bangkok, Thailand. Association for Computational Linguistics.

Yiming ZhangHarshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, Daphne Ippolito, . 2025. Noveltybench: Evaluating creativity and diversity in language models. In *Second Conference on Language Modeling*.

A. Appendix

Research Idea Novelty Judgment Prompt

```
You are an expert in machine
learning research evaluation. You
will be given two inputs:

1. A research idea with objective,
 problem statement, and solution
approach.
2. A list of related works, each
with a title and abstract.

Your task is to **assess the
novelty of the research idea**
compared to the related works.

### Instructions:
- Analyze the research idea and
summarize its key contributions.
- Compare it with the related
works to identify overlaps and
differences.
- Specifically, assess whether the
 idea introduces **significant new
 aspects** not present in existing
 work, or if it is largely a
variation on known approaches.
- Provide your output as a **JSON
object only**, with:
  - "reasoning": a short paragraph
   (2-4 sentences) explaining the
  reasoning behind the novelty
  score.
  - "novelty_score": an integer
  between 1-5 where: {
  novelty_rubric}

### Inputs:

**Research Idea:**
{research_idea}

**Related Works:**
{related_works}

### Output Format:
```json
{{
  "reasoning": <short explanation
  >,
  "novelty_score": <1|2|3|4|5>
}}
```

Figure 6: Zero-shot instructions for judging the novelty of research ideas.
```