LigBench: A Unified and Human-Aligned Benchmark for LLM-based Research Idea Generation

Anonymous ACL submission

Abstract

With the rapid advancement of large language models (LLMs), automated research idea generation has attracted growing interest. Recent approaches allow LLMs to retrieve relevant literature and propose novel ideas across diverse scientific domains. However, existing evaluation practices remain fragmented and lack unified standards, often relying on direct LLM-based scoring, which limits reliability and consistency across heterogeneous idea sets. To address these challenges, we propose **LigBench**, an automated benchmark for fine-grained, reliable, and human-aligned assessment of AI research ideas. LigBench ensures consistent evaluation by combining structured idea formalization, large-scale reference data, and iterative pairwise comparison. We also introduce **PAIR-IQ**, a curated dataset from real academic papers annotated with debiased OpenReview scores, supporting pairwise judgment model training and serving as a reference for objective comparative evaluation. Experiments show that LigBench delivers stable, interpretable, and consistent evaluations, improving agreement with expert judgments, while models trained on PAIR-IQ achieve higher ranking accuracy and robustness, establishing a scalable benchmark for objective research idea assessment.

1 Introduction

The rapid evolution of LLMs (Achiam et al., 2023; Comanici et al., 2025; Yang et al., 2025) has ushered in a new era of automated research idea generation. Building on the strong reasoning and knowledge synthesis capabilities of these models, a growing body of recent work (Zhou et al., 2024; Lu et al., 2024; Baek et al., 2025; Su et al., 2025) has proposed LLM-based frameworks for scientific ideation, enabling the generation of novel hypotheses, methodological variations, and promising research directions grounded in large-scale scientific literature. Such capabilities hold great promise for accelerating discovery across disciplines, guiding researchers toward unexplored directions, and reducing the time and cost associated with manual ideation. However, as the utility of these systems continues to grow, the lack of rigorous, objective, and reproducible evaluation methodologies remains a critical bottleneck for the systematic assessment of generated ideas.

Current evaluation strategies for idea generation are often fragmented. Some studies (Hu et al., 2024; Si et al., 2024) rely on direct scoring by LLMs(Zheng et al., 2023), where models assign quality scores to their own or peer-generated ideas, while others (Wang et al., 2024; Hu et al., 2024) depend on human experts to perform pairwise or ranking-based evaluations. Automated metrics such as novelty, diversity, or semantic similarity offer partial insights but fail to capture the multi-dimensional nature of research idea quality. Overall, these approaches lack consistency: they may be sensitive to prompt phrasing, suffer from inter-annotator variability, or provide only relative assessments. Consequently, there is no unified framework to benchmark idea generation systems comprehensively.

Pairwise comparison has emerged as a promising approach to sidestep some subjective biases inherent in absolute scoring. In this paradigm, two candidate ideas are presented side by side, and an evaluator (whether human or automated) selects the one deemed superior along specified criteria. Aggregating multiple pairwise comparisons can yield a ranking of ideas with higher consistency than solitary ratings. Nevertheless, pairwise evaluation alone does not constitute a full benchmarking solution: it requires a sufficiently large and diverse pool of test pairs, clear annotation protocols, and scalable tooling to collect judgments at scale.

To address these challenges, we introduce LigBench, an automated evaluation framework tailored for research idea assessment. LigBench oper-

ationalizes multi-dimensional quality criteria by orchestrating large-scale pairwise comparisons, metadata tracking, statistical aggregation, and automatic score updates into a cohesive pipeline. We drew on the thinking process experts use when evaluating an idea. Experts typically begin by searching for existing results related to the idea for reference or comparison. In addition, pairwise comparison alone is not sufficient to objectively and convincingly reflect the quality of an idea. Therefore, we have designed and provided a complete process based on pairwise comparison combined with the Elo mechanism to continuously update scores. We also present the PAIR-IQ dataset, a curated dataset of research papers' ideas containing structured idea representations. PAIR-IQ serves both as a gold-standard reference for comparative evaluation and as a training resource for models designed to predict pairwise judgments.

Our main contributions are summarized as follows:

- **Unified Evaluation Framework:** We introduce LigBench as a unified and extensible evaluation framework that systematically integrates data curation, idea retrieval, pairwise comparison, score propagation, and result aggregation within a single automated pipeline.

- **PAIR-IQ Dataset:** We release PAIR-IQ dataset, comprising over 11000 conference papers with debiased evaluation scores and formalized representations to support objective assessment of research ideas.

- **Open-Source Resources:** We will release the PAIR-IQ dataset together with the complete LigBench evaluation pipeline to facilitate reproducibility and enable community-driven extensions. The PAIR-IQ dataset is publicly available at `https://huggingface.co/datasets/USER3IjEBHj9/PAIR-IQ/tree/main`. The full LigBench evaluation pipeline will be released in the near future.

- **Benchmarking and Analysis:** We conduct extensive experiments with state-of-the-art LLMs, demonstrating that LigBench can discern subtle differences in idea quality and offering insights into model strengths and weaknesses.

2 Related Work

Current LLM-based ideation systems typically employ evaluation protocols that are tightly coupled to their own generation pipelines, reflecting framework-specific assumptions rather than a unified or standardized assessment methodology. Existing approaches primarily rely on either direct scoring by large language models or judgments from human experts, and are often designed to operate only within the context of a particular ideation system. As a result, these evaluation methods lack the flexibility and generality required for consistent assessment across different models, prompting strategies, and idea distributions. For instance, SciPIP (Wang et al., 2024) relies on evaluations conducted by human researchers, limiting scalability and reproducibility. AI Idea Bench (Qiu et al., 2025) introduces an evaluation protocol that is heavily dependent on the underlying idea generation framework and does not support standalone assessment of individual ideas. CoI (Li et al., 2024) proposes a pairwise comparison approach, but restricts evaluation to comparisons among generated ideas, making objective comparison with human-authored research ideas challenging. Consequently, as highlighted by a recent survey (Shahhosseini et al., 2025), evaluation remains the most critical bottleneck in LLM-based scientific ideation.

3 Method

To provide a unified, objective, and human-aligned evaluation of research ideas, we propose **LigBench**, an automated benchmark designed for systematic idea assessment. As illustrated in Figure 1, the evaluation of each idea is decomposed into four aspects: *rating*, *contribution*, *soundness*, and *novelty*, with each aspect scored on a **0–5 scale**. Here, *rating* represents the overall quality as perceived by reviewers, *contribution* reflects the significance of the idea within its research domain, *soundness* captures methodological rigor and validity, and *novelty* measures the originality of the proposed idea.

Following the pipeline shown in Figure 1, each idea is first transformed into a formalized representation and then iteratively compared against similar papers retrieved from the database using a large language model. In each comparison, the model evaluates the relative quality between idea pairs, and the resulting judgments are aggregated with the objective scores from the PAIR-IQ dataset to update the score of the target idea. This process

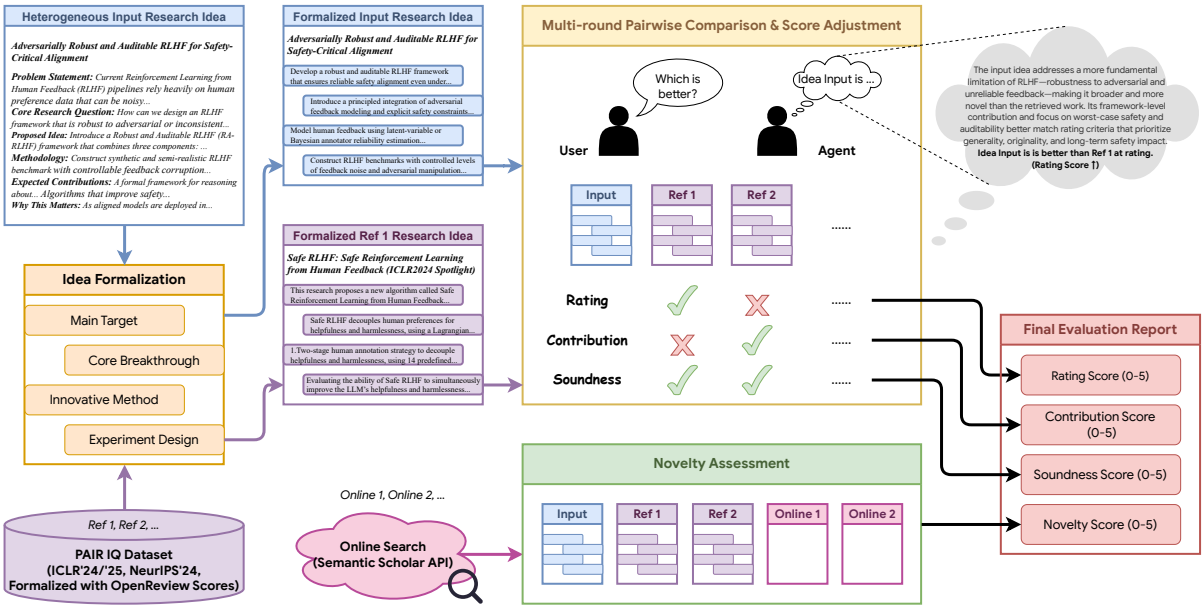


Figure 1: Overview of the LigBench evaluation pipeline. Given a target idea, LigBench performs idea formalization, retrieves related papers, conducts LLM-based pairwise comparisons, and iteratively updates scores using PAIR-IQ until convergence across multiple evaluation dimensions.

is repeated until convergence, producing a final score that is consistent across different sources and aligned with human preferences.

3.1 Idea Formalization

To begin with, a fundamental challenge in building a unified idea evaluation framework is the heterogeneity in the formats of the ideas under evaluation. Due to inherent differences in the distributions of ideas generated by different models and frameworks, directly comparing these ideas is prone to bias. For instance, when using large language models as evaluators, they often tend to favor ideas that are described in greater detail, potentially overlooking those that are more concise yet fundamentally more innovative or valuable. Moreover, different studies are conducted under diverse frameworks, leading to substantial variations in how final ideas are presented, including their structure and formatting (Chiang and Lee, 2023; Liu et al., 2023), which further complicates fair and consistent comparison across models and methods.

To mitigate these issues, it is necessary to standardize ideas across different distributions. Here, we focus on the core aspects and boundaries of each idea and decouple it into four distinct components:

- **Main Target**: a concise, one-sentence summary of the task and objective that the idea aims to address.
- **Core Breakthrough**: the key advancement or novel contribution introduced by the idea compared to prior work.
- **Innovative Methods**: the concrete methodological approaches or techniques proposed to achieve the target.
- **Experimental Design**: the experimental setup and evaluation plan designed to validate the effectiveness of the proposed methods.

Formally, let $I \in \mathcal{I}$ denote an input idea from the idea space. We define an LLM-based decomposition operator $\mathcal{D}_{\text{LLM}} : \mathcal{I} \to \mathcal{T} \times \mathcal{B} \times \mathcal{M} \times \mathcal{E}$ such that

$$(T, B, M, E) = \mathcal{D}_{\text{LLM}}(I), \tag{1}$$

where $T \in \mathcal{T}$ represents the main target, $B \in \mathcal{B}$ denotes the core breakthrough, $M \in \mathcal{M}$ corresponds to the innovative methods, and $E \in \mathcal{E}$ specifies the experimental design. This formulation casts idea formalization as a structured, LLM-driven mapping into a unified representation space, facilitating systematic comparison and subsequent quantitative evaluation. The concrete prompting strategy used for this formalization is provided in Appendix B.1.

By explicitly disentangling complementary semantic facets of an idea into structured components,

this LLM-driven decomposition yields a unified yet fine-grained representation that captures the substantive content while abstracting away superficial variations in expression and format. Consequently, it mitigates evaluation bias arising from heterogeneous idea sources and presentation styles, enabling more objective and consistent comparison across models and frameworks. Furthermore, this structured representation establishes a principled foundation for subsequent quantitative analyses, including assessments of soundness, novelty, contribution, and other relevant criteria.

3.2 PAIR-IQ Construct

To better evaluate research ideas according to human preferences, we collected and constructed the **PAIR-IQ dataset**, comprising over 11,000 papers from **ICLR 2024, ICLR 2025, and NeurIPS 2024**, including oral, spotlight, poster, and rejected papers. For each paper, we retrieved **ratings, contribution scores, and soundness scores from OpenReview**, and projected each score to a standardized 0–5 range to ensure comparability across different venues and scoring scales.

Figure 2 illustrates the composition of the PAIR-IQ dataset. As shown in Figure 4, the dataset draws from three major machine learning venues, with ICLR 2025 contributing the largest proportion (36.7%), followed by ICLR 2024 (32.4%) and NeurIPS 2024 (30.9%). Figure 2a presents the distribution across acceptance categories: poster papers constitute the majority (57.3%), followed by rejected submissions (36.6%), spotlight papers (4.3%), and oral presentations (1.8%). This distribution ensures that the dataset captures a wide spectrum of paper quality levels, from top-tier accepted works to borderline and rejected submissions. Figure 2b further depicts the topical diversity of the dataset, spanning 12 research areas. Large language models (LLM) represent the most prevalent topic (19.3%), followed by training methodologies (15.4%) and reinforcement learning (13.4%), reflecting the current research trends in the machine learning community. Detailed information and a specific example of the dataset are provided in Appendix C.

Following the idea formalization framework introduced in the previous subsection, we now describe in detail how we extract and formalize research ideas from full academic papers. For each paper, we parsed all textual content and employed a LLM to extract structured components, including Main Target, Core Breakthrough, Innovative Methods, and Experimental Design. These components provide a standardized representation of each research idea, enabling consistent downstream evaluation and facilitating subsequent scoring procedures. A more detailed description of the prompting strategy and extraction process is provided in Appendix B.2.

	Rating	Contribution	Soundness
ICLR 2024	2.740	2.689	2.472
ICLR 2025	2.821	2.687	2.498
NeurIPS 2024	2.918	2.914	2.721
all	2.825	2.758	2.559

Table 1: Average scores of papers from different venues in terms of rating, contribution, and soundness. The values highlight systematic differences across conferences and years, motivating the need for score debiasing.

Next, we address potential biases in the collected ratings. Since different venues and years may exhibit systematic differences in reviewer behavior, the resulting score distributions can vary significantly across the dataset. To ensure that the scores of all papers are comparable and the overall distribution is consistent, we perform **score debiasing** using a mean-shifting procedure.

Specifically, let $s_i^{(v)}$ denote the original score (rating, contribution, or soundness) of paper i from venue v, with venue mean μ_v and overall mean μ_{all}. The debiased score is computed as

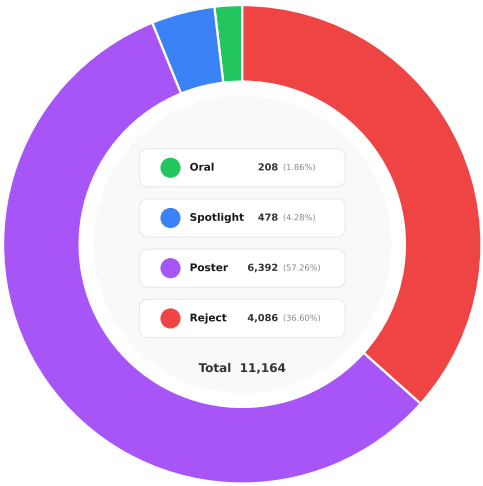
$$\hat{s}_i^{(v)} = s_i^{(v)} - \mu_v + \mu_{\text{all}}. \tag{2}$$

This procedure aligns the score distributions across different venues and years, ensuring that subsequent evaluations reflect the intrinsic quality of each idea rather than systematic biases in reviewer ratings. Additionally, the dataset was further processed through embedding computation to facilitate downstream evaluation tasks.

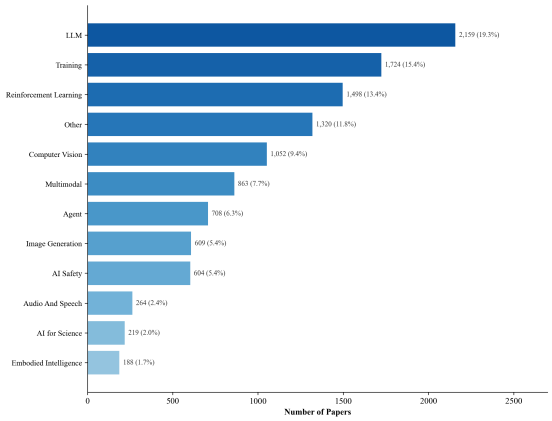
3.3 Pairwise Comparison

When a new research idea is submitted for evaluation, we first transform it into a formalized representation using the same procedure described in the previous subsection. This ensures that the target idea and reference papers are represented in a unified format.

Based on this representation, we retrieve semantically related papers from the database using a parallel retrieval strategy that combines keyword



(a) Distribution by acceptance category.



(b) Distribution by research topic.

Figure 2: Statistical overview of the PAIR-IQ dataset. The dataset comprises 11,164 papers collected from three major machine learning conferences (ICLR 2024, ICLR 2025, and NeurIPS 2024), spanning diverse acceptance categories and research topics.

matching and embedding-based similarity. The target idea is then compared pairwise with each retrieved paper using a LLM, which assesses relative quality along three dimensions: *rating*, *contribution*, and *soundness*. The specific score initialization and the comparison prompt are provided in Appendix B.3.

We update the target idea's scores using an adapted Elo rating algorithm(Elo, 1978). In standard Elo, the expected probability that idea A wins over idea B is:

$$E_A = \frac{1}{1 + 10^{(s_B - s_A)/d}}, \tag{3}$$

where s_A and s_B denote the current scores of ideas A and B, and d is a scaling parameter. Given the observed outcome $S_A \in \{0, 0.5, 1\}$, the score is updated as:

$$s_A^{\text{new}} = \mathcal{C}\left(s_A + K \cdot (S_A - E_A)\right), \tag{4}$$

where K is an adaptive update factor and $\mathcal{C}(\cdot)$ is a soft clamping function that constrains scores to $[0, 5]$.

Through multiple rounds of pairwise comparisons, the target idea's scores gradually converge. The evaluation terminates when score changes fall below a predefined threshold, yielding the final assessment. A detailed mathematical analysis of the algorithm, including convergence properties and parameter selection, is provided in Appendix D.

3.4 Novelty Assessment

Through the above procedures, we obtain objective evaluations of research ideas in terms of *rating*, *contribution*, and *soundness*. However, these dimensions do not explicitly capture the *novelty* of an idea. Unlike other evaluation dimensions that can be assessed through comparative analysis, novelty requires determining whether an idea introduces genuinely new insights compared to existing literature.

To address this challenge, we design a novelty assessment module that combines LLM-based initial estimation with similarity-based quantitative analysis. The evaluation proceeds in three stages: (1) obtaining an initial novelty estimate $s_{\text{novelty}}^{(0)}$ from the LLM evaluator; (2) retrieving semantically related papers via the Semantic Scholar API(Kinney et al., 2023) and computing a weighted similarity metric s_{weighted}, which is then mapped to a novelty score s_{sim} through an inverse sigmoid function; and (3) fusing both assessments into the final score:

$$s_{\text{novelty}}^{\text{final}} = \beta \cdot s_{\text{sim}} + (1 - \beta) \cdot s_{\text{novelty}}^{(0)}, \tag{5}$$

where $\beta = 0.7$ assigns higher weight to the similarity-based calculation, reflecting its objectivity, while retaining the LLM's judgment to capture aspects not fully represented by semantic similarity alone. A detailed description of the similarity computation and parameter selection is provided in Appendix E.

4 Experiments

4.1 Model Training

We also fine-tune models of smaller sizes to investigate whether the scientific "taste" of an area can be learned through knowledge distillation(Hinton et al., 2015). Following Section 4.2, we sample 15,000 thematically similar pairs while ensuring none of the selected papers appear in the test data. To help small models better understand the underlying domain knowledge, we adopt an additional chain-of-thought (Wei et al., 2023) generated by larger models based on ground-truth labels. We choose Qwen2.5-72B-Instruct (Qwen Team et al., 2025) as the teacher model, and smaller models from the same family as the student models. As for the training framework, we employ LLaMA-Factory (Zheng et al., 2024), and further details are presented in Appendix A.

4.2 Pairwise Evaluation

The accuracy of pairwise judgments made by the LLM serves as a key guarantee for the reliability of the entire evaluation framework. To assess this reliability, we conduct a dedicated pairwise evaluation on the three dimensions: *rating*, *contribution*, and *soundness*. The goal of this evaluation is to verify whether the LLM can correctly determine the relative quality between two research ideas under each criterion.

For this purpose, we randomly sample a held-out test set consisting of 269 pairs of thematically similar papers from the database, ensuring that none of the selected papers appear in the training data. Thematic similarity is determined using the same hybrid retrieval strategy based on keyword matching and embedding-based similarity. Each sampled pair is presented to the LLM in its formalized representation, and the model is asked to judge which idea performs better on each of the three dimensions.

The ground-truth labels for pairwise comparison are derived from the **debiased OpenReview scores**. Specifically, for each paper pair, we compare their debiased scores on *rating*, *contribution*, and *soundness*, and treat the resulting relative ordering as the reference outcome. The accuracy of the LLM's pairwise judgments is then measured by comparing its predictions against these ground-truth relationships, as summarized in Table 2.

Overall, stronger models consistently achieve higher accuracy across all three dimensions, indicating that pairwise judgment is a non-trivial capability that benefits from increased model capacity and reasoning ability. Among the off-the-shelf models, GPT-5 substantially outperforms other baselines, achieving accuracies above 0.80 on both *rating* and *soundness*, which suggests a high level of agreement with debiased human judgments.

Model	Rating	Contrib.	Sound.
Qwen2.5-7B	0.514	0.496	0.488
Qwen2.5-14B	0.506	0.598	0.541
GPT-4o	0.549	0.615	0.747
GPT-4.1	0.539	0.545	0.649
GPT-5	**0.820**	0.695	**0.826**
GPT-5.2	0.801	0.707	0.803
Gemini-2.5-pro	0.706	0.607	0.750
Gemini-3-pro	0.806	**0.713**	0.814
Qwen2.5-7B(trained)	0.714	0.710	0.712
Qwen2.5-14B(trained)	0.755	0.701	0.751

Table 2: Accuracy of LLM-based pairwise judgments on *rating*, *contribution*, and *soundness*, evaluated against debiased OpenReview ground-truth labels.

In contrast, smaller open-source models such as Qwen2.5-7B and Qwen2.5-14B exhibit limited performance, with accuracies close to random guessing in some dimensions, particularly on *rating* and *soundness*. This highlights the difficulty of reliably performing fine-grained comparative evaluation without sufficient model capacity or task-specific supervision.

Notably, models trained on the PAIR-IQ dataset show substantial improvements across all dimensions. Both trained variants achieve consistently higher accuracies than their untrained counterparts, demonstrating that targeted pairwise supervision effectively enhances the model's ability to align with human evaluation criteria. These results validate the design of the PAIR-IQ dataset and further support the use of pairwise judgment as a reliable component within the LigBench evaluation framework.

Although the pairwise judgment accuracy does not reach near-perfect levels, this outcome also reflects the intrinsic difficulty of the task, as distinguishing fine-grained differences between closely related research ideas is inherently challenging. Nevertheless, when applied within the LigBench framework, even imperfect pairwise judgments can still lead to reliable final evaluations. Starting from a strong base model, repeated score updates over a large number of pairwise comparisons allow the target scores to gradually converge toward reasonable

Rating	Contrib.	Sound.
0.114	0.196	0.138

Table 3: Mean debiased score differences for incorrectly judged paper pairs by **GPT-5** in the pairwise evaluation. The small score gaps indicate that errors have limited impact on iterative score updates.

Rating	Contrib.	Sound.
71%	79%	73%

Table 4: Agreement between LLM-based pairwise judgments and human expert evaluations across three assessment dimensions. Values denote the proportion of pairwise decisions consistent with PhD-level expert judgments.

values.

Further analysis of the error cases, summarized in Table 3, reveals that most incorrect judgments occur when the debiased scores of the two compared papers are very close. This observation suggests that even when the LLM makes an incorrect decision, the resulting update magnitude is limited. According to the score update mechanism described in the previous section, such small score differences lead to correspondingly small update steps, thereby preventing severe error accumulation during the iterative evaluation process.

4.3 Human Alignment

Although the pairwise judgment capability of the large language model has been validated using debiased OpenReview scores, it remains essential to perform human verification. This is because even debiased scores cannot guarantee a perfectly objective assessment, and expert judgment is required to ensure alignment with human evaluation standards.

To further verify human alignment, we recruited several PhD-level AI researchers to conduct pairwise comparisons on **100 randomly sampled idea pairs**. For each pair, experts selected the superior idea along the three dimensions: *rating*, *contribution*, and *soundness*. The resulting human judgments were then compared with the LLM's pairwise predictions. The alignment results are summarized in Table 4, indicating a strong consistency between the LLM and expert evaluations.

As shown in Table 4, the LLM exhibits substantial agreement with human experts across all three dimensions. In particular, the highest alignment is observed for *contribution*, indicating that the model is effective at assessing the relative significance of research ideas within a domain. The consistently strong agreement on *rating* and *soundness* further suggests that the LLM's pairwise judgments capture key aspects of overall quality and methodological rigor. These results provide additional evidence that LigBench produces evaluations that are well aligned with expert human preferences, complementing the automated assessments based on debiased review scores.

	Accept	Reject
Rating	2.547	2.043
Contribution	2.432	2.030
Soundness	1.853	1.542
Novelty	2.627	2.234

Table 5: Evaluation of 50 NeurIPS 2025 papers using LigBench. Each score represents the average assessment of the papers on *rating*, *contribution*, *soundness*, and *novelty*.

4.4 Paper Evaluation

We conducted an evaluation on papers from **NeurIPS 2025** to demonstrate the effectiveness of our framework. The evaluation applied the full LigBench procedure. The results of this evaluation are summarized in Table 5, highlighting the performance of the framework across the different assessment dimensions.

As shown in Table 5, papers that were accepted at NeurIPS 2025 consistently receive higher scores across all four evaluation dimensions compared to rejected papers. The largest differences are observed in *novelty* and *rating*, suggesting that both originality and overall perceived quality are strong determinants of acceptance. Differences in *contribution* and *soundness* are also notable, indicating that methodological rigor and domain impact are effectively captured by LigBench.

These results demonstrate that LigBench can meaningfully distinguish higher-quality research ideas from lower-quality ones, even when applied to a relatively small test set of 50 papers. This alignment with actual acceptance outcomes provides preliminary validation of the framework's effectiveness in reflecting human and peer-review judgments.

4.5 Benchmark and Model Evaluation

To comprehensively assess the quality of LLM-generated research ideas, we evaluate both stan-

Model	Rating	Contrib.	Sound.	Novelty
CoI	3.468	2.996	2.487	2.474
SciPIP	2.621	1.900	**3.243**	2.267
GPT-5.2	**3.974**	**3.392**	3.022	**2.583**
GPT-5	3.686	3.210	2.866	2.528
GPT-4.1	1.000	0.760	1.877	2.194
GPT-4o	0.672	0.514	1.264	1.877

Table 6: Evaluation of idea-generation frameworks and standalone LLMs using LigBench. Scores are averaged across 110 generated ideas per model. CoI(Li et al., 2024) and SciPIP(Wang et al., 2024) both use GPT-5 as the backbone model.

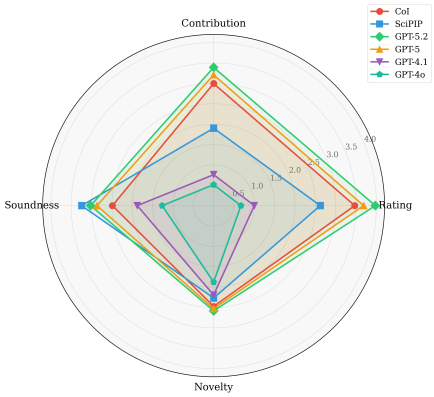


Figure 3: Radar chart visualization of Table 6. GPT-5.2 achieves the highest overall performance, while CoI demonstrates competitive results despite using GPT-5 as its backbone.

dalone LLMs and idea-generation frameworks using LigBench. Each system is prompted to freely generate 10 ideas across 11 research topics, including Reinforcement Learning, Image Generation, AI Safety, Agent, Computer Vision, Audio and Speech, Embodied Intelligence, Multimodal, AI for Science, Large Language Models, and Training. The final score for each model is computed as the average across all generated ideas.

We compare two representative idea-generation frameworks, CoI and SciPIP, against direct LLM baselines. Both frameworks employ GPT-5 as their underlying language model, enabling a controlled comparison that isolates the contribution of framework design from model capability. The evaluation results are summarized in Table 6.

Several observations emerge from these results. First, among standalone LLMs, we observe a clear performance hierarchy aligned with model capability: GPT-5.2 and GPT-5 substantially outperform GPT-4.1 and GPT-4o across all dimensions, with GPT-4o scoring below 1.0 on rating and contribution, indicating that earlier-generation models struggle to meet basic quality thresholds.

Second, an unexpected finding is that idea-generation frameworks do not consistently outperform direct prompting when applied to strong backbone models. Despite using GPT-5 as their underlying model, both CoI and SciPIP achieve lower scores than standalone GPT-5 on most dimensions. This suggests that these frameworks may have been optimized for enhancing weaker models, and their structured pipelines introduce constraints that limit the performance ceiling of already-capable models. Notably, SciPIP attains a higher score on the Soundness dimension than its underlying model GPT-5, and achieves the highest soundness score among all evaluated systems and models, indicating that its design contributes positively to methodological rigor despite the overall performance trade-offs.

Finally, although the novelty scores of the various models and frameworks exhibit an ordering that is broadly consistent with our expectations, the absolute differences between their scores are relatively small. This observation also suggests, from another perspective, that for current LLMs, generating a truly novel idea remains a non-trivial challenge.

5 Conclusion

We introduced **LigBench**, a unified and human-aligned benchmark for evaluating research ideas generated by large language models and agentic frameworks. By decomposing idea quality into four dimensions—*rating*, *contribution*, *soundness*, and *novelty*, LigBench enables consistent and fine-grained assessment across diverse idea sources.

To support reliable evaluation, we constructed the **PAIR-IQ** dataset with standardized and debiased scores, enabling effective pairwise judgment and robust score updating. Experimental results show that LigBench produces stable evaluations aligned with expert preferences, even when individual pairwise judgments are imperfect.

We believe LigBench provides a practical foundation for benchmarking future idea-generation systems and advancing the automated evaluation of scientific creativity, with the potential to serve as a reliable reward signal for training idea-generation models through reinforcement learning, thereby enabling more effective, scalable, and human-aligned development of AI-driven research idea generation.

6 Limitations

Despite its effectiveness, LigBench has several limitations. The framework relies on the pairwise judgment capability of large language models to compare research ideas across multiple dimensions. While our experiments show strong agreement with debiased OpenReview scores and human expert judgments, pairwise comparison remains inherently challenging, particularly when differences between ideas are subtle or when ideas are closely related. In addition, LigBench depends on a static database of existing papers for reference-based evaluation and score propagation. Although the use of a large and diverse corpus improves robustness, ideas from newly emerging research areas or those that substantially deviate from existing literature may be less accurately assessed due to limited relevant references.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and 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.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Arpad E. Elo. 1978. *The Rating of Chessplayers, Past and Present*. Arco Publishing.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. 2024. Nova: An iterative planning and search approach to enhance novelty and diversity of llm generated ideas. *arXiv preprint arXiv:2410.14255*.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, and 1 others. 2023. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, and 1 others. 2024. Chain of ideas: Revolutionizing research via novel idea development with llm agents. *arXiv preprint arXiv:2410.13185*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and 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.

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

Yansheng Qiu, Haoquan Zhang, Zhaopan Xu, Ming Li, Diping Song, Zheng Wang, and Kaipeng Zhang. 2025. Ai idea bench 2025: Ai research idea generation benchmark. *arXiv preprint arXiv:2504.14191*.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Fatemeh Shahhosseini, Arash Marioriyad, Ali Momen, Mahdieh Soleymani Baghshah, Mohammad Hossein Rohban, and Shaghayegh Haghjooy Javanmard. 2025. Large language models for scientific idea generation: A creativity-centered survey. *arXiv preprint arXiv:2511.07448*.

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

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, and 1 others. 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.

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 preprint arXiv:2410.23166*.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. Hypothesis generation with large language models. *arXiv preprint arXiv:2404.04326*.

A Hyper-parameters for Fine-Tuning

Table 7: Hyper-parameters for Fine-Tuning.

Hyper-Parameter	Default Value
Finetuning Type	LoRA
LoRA Target	all
LoRA Rank	16
LoRA Alpha	16
LoRA Dropout	0.05
Cutoff Length	4,096
Gradient Accumulation Steps	16
Learning Rate	1×10^{-4}
Train Epochs	1.0
Learning Rate Scheduler	Cosine
Warmup Ratio	0.1

B Prompt Specification

B.1 Heterogeneous Idea Formalization

Heterogeneous Idea Formalization PROMPT

You are an expert agent specialized in research idea analysis and evaluation.
You will be provided with a raw research idea description. Your task is to formalize this idea into a structured representation that facilitates systematic comparison and evaluation.
Each research idea should be decomposed into the following four components: 1. Main Target: A concise, one-sentence summary of the core task and research objective that the idea aims to address. 2. Core Breakthrough: The key conceptual or technical advancement introduced by the idea, highlighting what fundamentally differentiates it from prior work. 3. Innovative Methods: The concrete methodological approaches, algorithms, or techniques proposed to realize the main target. 4. Experimental Design: The experimental setup and evaluation strategy used to validate the proposed methods, including datasets, benchmarks, baselines, and evaluation metrics when applicable.
Now, the new idea is: <new_idea>
Please formalize it into the four parts.
Output Format: A list of four strings in the following order: ["<main_target>", "<core_breakthrough>", "<innovative_method_1>; <innovative_method_2>; <innovative_method_3>; ...", "<experimental_design>"]
Example: ["To improve the robustness of large language models against distribution shifts", "Introducing a unified robustness-aware training objective that jointly models uncertainty and domain variance", "Uncertainty-guided representation learning; Domain-adaptive regularization; Robust contrastive pretraining", "Evaluation on multiple out-of-distribution NLP benchmarks with comparisons to standard fine-tuning and domain adaptation baselines"]

B.2 Paper Idea Formalization

When constructing the PAIR-IQ dataset, we need to extract standardized research ideas from existing papers, which differs from directly normalizing heterogeneous, free-form ideas. Specifically, we first parse the full textual content of each paper and use an LLM to map diverse and heterogeneous section headings into five canonical categories: *Abstract*, *Introduction*, *Related Work*, *Method*, and *Experiment*. Based on this unified structure, we perform targeted extraction for different components of an idea from the corresponding sections. The *Main Target* is primarily extracted from the *Abstract*. The *Core Breakthrough* is derived from the *Abstract*, *Introduction*, and *Related Work*. The *Innovative Methods* are extracted from the *Abstract* and *Method*. The *Experimental Design* is obtained from the *Experiment* section. The detailed prompts used for this process are provided below.

Section Mapping PROMPT

You will get a list of section names from a research paper. Please identify which section(s) correspond to each of the following: abstract, introduction, related work, method, and experiment.
Note: Sometimes a part (e.g., method) may be split across multiple sections. In such cases, please use a list to record all matching section names, e.g., "method": ["Methodology", "Approach"]. Note: The "related work" section may also describe existing problems in the field, not just previous approaches. The answer format MUST be a JSON object with keys "abstract", "introduction", "related_work", "method", and "experiment". Each value should be either a list of section names (if present) or null (if not present). For example: {{ "abstract": ["Abstract"], "introduction": ["Introduction"], "related_work": ["Related Work"], "method": ["Methodology", "Approach"], "experiment": ["Experiment"] }} Now, the section name list is: “‘{section_names}”’.

Main Target PROMPT

Here you will get the abstract part of a research paper. You need to return the main target of this paper in one sentence, such as "This research designs a new model for image classification." or "This research proposes a new dataset for object detection" or "This research introduces a new framework for idea generation" and so on.
The answer format MUST be a simple string without any explanation or extra content.
Now , the abstract is: “‘{abstract}”’.

Core Breakthrough PROMPT

Here you will get three sections of a research paper: abstract, introduction, and related work. You need to return the core breakthrough of this paper in one sentence, such as what existing problems are solved or what advances are made compared to previous work. The answer format MUST be a simple string without any explanation or extra content. Now, the abstract part is: “‘{abstract}”’. The introduction part is: “‘{introduction}”’. The related work part is: “‘{related_work}”’.

Innovative Methods PROMPT

Here you will get the abstract and method sections of a research paper. Please list three to four of the most innovative and core methods proposed in this paper, each as a separate numbered item. Keep your answer concise and focus on the key elements only. The answer format MUST be a simple numbered list (e.g., "1. ... 2. ...") without any explanation or extra content. Now, the abstract part is: “‘{abstract}”’. The method part is: “‘{method}”’.

Experimental Design PROMPT

Here you will get the abstract and experiment sections of a research paper. Please list the main experimental designs proposed in this paper, each as a separate numbered item. Ignore any results, numbers, or metrics mentioned in the experiment section; only describe what experiments were designed. The answer format MUST be a simple numbered list (e.g., "1. ... 2. ...") without any explanation or extra content. Now, the abstract part is: “‘{abstract}”’. The experiment part is: “‘{experiment}”’.

B.3 Pairwise Comparison

Initial Scoring with Examples PROMPT

You are an intelligent agent who is expert in scientific hypothesis evaluation.
A research idea can be decomposed into four part: Main Target, Core Breakthrough, Innovative Approach, and Experimental Design.
Here you will get a new idea and three published papers' ideas. You need to score the new idea in four dimensions: Rating, contribution, Soundness, and Novelty.
Dimension Definitions:
- Rating: The overall quality of the idea and a comprehensive evaluation
- Contribution: Whether the outcomes will be useful for future work
- Soundness: The logical and methodological integrity of the research
- Novelty: The originality and potential to contribution new insights or approaches
We have scored these three published papers in four dimensions. The "comments" record the reasons for the scores.
Please read these three published papers' scores and comments carefully, and then score the new idea.
New Idea: “‘ Main Target: {new_idea_main} Core Breakthrough: {new_idea_breakthrough} Innovative Approach: {new_idea_approach} Experimental Design: {new_idea_experiment}
”’
Calibration Example 1: “‘ {example1} ”’
Calibration Example 2: “‘ {example2} ”’
Calibration Example 3: “‘ {example3} ”’
Important: Think carefully, as a new idea may appear highly novel and compelling but lack sufficient supporting details.
Output Format: A JSON list of four floating-point numbers (0-5): [<rating>, <contribution>, <soundness>, <novelty>]
Example: [3.7, 4.1, 2.6, 2.9]

Pairwise Comparison PROMPT

You are an intelligent agent who is expert in **scientific hypothesis evaluation**. A research idea can be decomposed into four part: Main Target, Core Breakthrough, Innovative Approach, and Experimental Design.
Here you will get two research ideas; you need to compare them on three dimensions: Rating, Soundness and Contribute, to judge which one is better on each dimension.
Dimensions definitions: - Rating: The overall quality of the idea. This is a comprehensive evaluation that summarizes the strengths and weaknesses of the idea as a whole. - Soundness: The logical and methodological integrity of the research. Assess whether the idea is based on solid reasoning, uses appropriate methods, and avoids major flaws or unsupported assumptions. - Contribute: Whether the outcomes will be useful for future work. Evaluate if the idea can advance scientific knowledge, provide valuable insights, or serve as a foundation for further research in the field.
Your answer format should be a think process between <think> and </think>, containing three steps: '1. ... 2. ... 3. ...', where 1 is the reasoning for Rating, 2 is the reasoning for Soundness, and 3 is the reasoning for Contribution, and a final list of three numbers (0, 1, or 0.5) in the order [Rating, Soundness, Contribution], where - '1' means the first idea is better, - '0.5' means the two ideas are equal, - '0' means the second idea is better. For example, if the first idea wins on Rating, loses on Soundness, and is equal on Contribution, your answer should be: '<think>...</think>[1, 0, 0.5]' When considering the rating, you may use the Main Target and Innovative Approach as the most important basis. When considering the soundness, you may use the Main Target, Innovative Approach and Experimental Design as the most important basis. When considering the contribution, you may use the Main Target and Core Breakthrough as the most important basis.
Now: First idea: “‘The Main Target is {idea_A_main}. And the Core Breakthrough is {idea_A_breakthrough}. The Innovative Approach is {idea_A_approach}. The Experimental Design is {idea_A_experiment}.”’
Second idea: “‘The Main Target is {idea_B_main}. And the Core Breakthrough is {idea_B_breakthrough}. The Innovative Approach is {idea_B_approach}. The Experimental Design is {idea_B_experiment}.”’
Please give your answer directly according to the above format.

C PAIR-IQ extra information

Figure 4 illustrates the dataset statistics of PAIR-IQ, showing the distribution of collected papers across different conference venues.

An example of the standardized paper idea representation used in the PAIR-IQ dataset is shown below.

Listing 1: Example of a Paper Idea in PAIR-IQ

```
{
  "title": "A Generative Model of Symmetry
      ↪ Transformations",
  "url": "https://openreview.net/forum
        ?id=aFP24eYpWh",
  "conference": "NeurIPS",
  "year": 2024,
  "label": "accept-poster",
  "abstract": "Correctly capturing the symmetry
      ↪ transformations of data can lead to
```

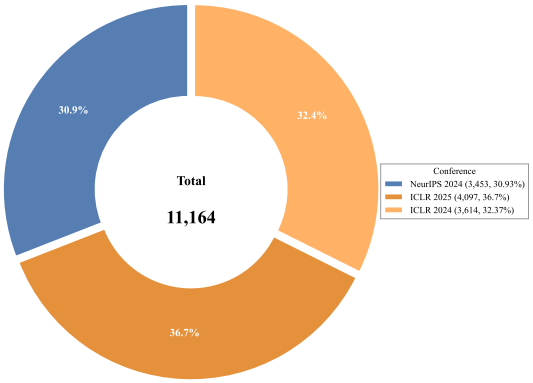


Figure 4: Distribution by conference venues.


```
      ↪ efficient models with strong
      ↪ generalization capabilities ...",
  "keywords": "approximate symmetries, invariances,
      ↪ deep generative models",
  "primary area": "Generative models",
  "rating": 2.9070274219833774,
  "contribute": 2.587674461477269,
  "soundness": 2.843680907598996,
  "novelty": "0",
  "analysis": {
    "main_target": "This research constructs a
        ↪ generative model that learns to capture
        ↪ the approximate symmetries of data from
        ↪ a prespecified set of possible
        ↪ symmetries.",
    "core_breakthrough": "This paper introduces a
        ↪ Symmetry-aware Generative Model (SGM)
        ↪ that learns to capture the approximate
        ↪ symmetries present in data, leading to
        ↪ improved data efficiency and higher
        ↪ marginal test-log-likelihoods when
        ↪ combined with standard generative
        ↪ models.",
    "innovative_methods": "1. A generative model
        ↪ that decomposes the data distribution
        ↪ into a distribution over prototypes and
        ↪ a distribution over transformation
        ↪ parameters.\n2. A self-supervised
        ↪ learning scheme to train an equivariant
        ↪ transformation inference function.\n3.
        ↪ A method for learning the distribution
        ↪ over transformation parameters using
        ↪ maximum likelihood, allowing the model
        ↪ to capture the extent of symmetries
        ↪ present in the data.\n4. Composing
        ↪ transformations in a transformation
        ↪ parameter space to minimize the number
        ↪ of applications and reduce
        ↪ discretization errors.",
    "experimental_design": "1. Exploring the
        ↪  S G M s  ability to learn symmetries by
        ↪ producing valid prototypes and
        ↪ generating plausible samples from the
        ↪ data distribution given those
        ↪ prototypes.\n2. Leveraging the SGM to
        ↪ improve data efficiency in deep
        ↪ generative models.\n3. Conducting
        ↪ experiments on dSprites, MNIST, and
        ↪ GalaxyMNIST..."
  },
  "embedding": [1024]
}
```


D Elo-based Score Update: Mathematical Analysis

This appendix provides a comprehensive mathematical treatment of the Elo-based score update mechanism used in LigBench, including theoretical foundations, convergence analysis, and empirical parameter optimization.

D.1 Algorithm Overview

The pairwise comparison module employs an adapted Elo rating system to iteratively update the scores of target ideas. The complete procedure is formalized in Algorithm 1.

Algorithm 1 Elo-based Pairwise Score Update

Require: Target idea I, Reference paper set $\mathcal{R}$, Initial score s_0, Convergence threshold ϵ, Maximum iterations $T_{\max}$

Ensure: Final score s^*

$s \leftarrow s_0$; $t \leftarrow 0$
while $t < T_{\max}$ **do**
  Sample reference paper $R \sim \mathcal{R}$
  Compute expected outcome: $E = \frac{1}{1+10^{(s_R - s)/d}}$
  Obtain LLM judgment: $S \in \{0, 0.5, 1\}$
  Compute adaptive K-factor: $K_t = K_{\max} \cdot \gamma^t$
  Update raw score: $s^{\text{raw}} = s + K_t \cdot (S - E)$
  Apply soft clamping: $s^{\text{new}} = \mathcal{C}(s^{\text{raw}})$
  if $|s^{\text{new}} - s| < \epsilon$ **then**
    break
  end if
  $s \leftarrow s^{\text{new}}$; $t \leftarrow t + 1$
end while
return $s^* = s$

D.2 Theoretical Foundation

D.2.1 Expected Outcome Function

The expected outcome function is derived from the Bradley-Terry model (Bradley and Terry, 1952), which assumes that the probability of item A being preferred over item B follows a logistic function of their latent quality difference:

$$P(A \succ B) = \frac{1}{1 + 10^{(s_B - s_A)/d}}. \tag{6}$$

This formulation satisfies several desirable properties:

- **Symmetry**: $P(A \succ B) + P(B \succ A) = 1$

- **Transitivity**: If $s_A > s_B > s_C$, then $P(A \succ C) > P(A \succ B)$
- **Monotonicity**: $\frac{\partial P(A \succ B)}{\partial s_A} > 0$

D.2.2 Score Update as Gradient Descent

The score update rule can be interpreted as a stochastic gradient descent step on a cross-entropy loss function. Let $\mathcal{L}(s_A)$ denote the expected loss:

$$\mathcal{L}(s_A) = -\mathbb{E}_{S_A}[S_A \log E_A + (1-S_A)\log(1-E_A)]. \tag{7}$$

Taking the gradient with respect to s_A:

$$\frac{\partial \mathcal{L}}{\partial s_A} = \frac{\ln 10}{d} \cdot (E_A - S_A). \tag{8}$$

Thus, the update rule $s_A \leftarrow s_A + K(S_A - E_A)$ corresponds to gradient descent with learning rate $\eta = \frac{Kd}{\ln 10}$.

D.3 Parameter Selection

D.3.1 Scaling Parameter d

In standard chess Elo systems, $d = 400$ with ratings spanning approximately $[0, 3000]$. Our framework operates on a normalized score range of $[0, 5]$. To preserve the relative sensitivity of the expected outcome function, we rescale d proportionally:

$$d_{\text{new}} = d_{\text{chess}} \cdot \frac{\Delta s_{\text{new}}}{\Delta s_{\text{chess}}} = 400 \cdot \frac{5}{3000} \approx 0.67. \tag{9}$$

However, empirical experiments revealed that this theoretical value leads to overly aggressive score changes. We conducted a grid search over $d \in \{0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0\}$ using a validation set of 500 idea pairs with known ground-truth scores.

d	MAE ↓	Rank Corr. ↑	Conver. Rate ↑
0.50	0.412	0.723	0.89
0.75	0.356	0.761	0.92
1.00	0.298	0.789	0.94
1.25	0.267	0.812	0.96
1.50	**0.251**	**0.831**	**0.97**
1.75	0.263	0.819	0.95
2.00	0.289	0.798	0.93

Table 8: Parameter sensitivity analysis for scaling parameter d.

Based on these results, we set $d = 1.5$, which achieves the optimal balance between score discrimination sensitivity and convergence stability.

D.3.2 Adaptive K-factor Strategy

The K-factor controls the magnitude of score adjustments. We employ an exponential decay strategy:

$$K_t = K_{\max} \cdot \gamma^t, \tag{10}$$

where $K_{\max}$ is the initial K-factor, $\gamma \in (0, 1)$ is the decay rate, and t is the iteration number.

This adaptive strategy addresses two competing objectives: (1) **Fast initial convergence**: Large K values enable rapid movement toward the true score when the initial estimate is poor. (2) **Fine-grained calibration**: Small K values prevent oscillation around the equilibrium and enable precise final positioning.

We optimized $(K_{\max}, \gamma)$ through grid search:

$K_{\max}$	γ	Iterations ↓	Final MAE ↓	Stability ↑
0.3	0.90	28.4	0.287	0.91
0.3	0.95	35.2	0.264	0.94
0.5	0.90	18.7	0.271	0.89
0.5	**0.95**	**22.3**	**0.251**	**0.97**
0.5	0.98	31.6	0.249	0.96
0.7	0.90	14.2	0.298	0.82
0.7	0.95	17.8	0.273	0.88

Table 9: Grid search results for K-factor parameters.

The optimal configuration is $K_{\max} = 0.5$ and $\gamma = 0.95$, balancing convergence speed with final accuracy.

D.4 Soft Clamping Mechanism

Since our scores are constrained to $[0, 5]$, we apply a smooth boundary function using the hyperbolic tangent:

$$\mathcal{C}(s) = 2.5 + 2.5 \cdot \tanh\left(\frac{s - 2.5}{2.5}\right). \tag{11}$$

D.4.1 Properties of the Clamping Function

The clamping function $\mathcal{C} : \mathbb{R} \to (0, 5)$ satisfies:

1. $\mathcal{C}(2.5) = 2.5$ (fixed point at center)
2. $\lim_{s\to\infty} \mathcal{C}(s) = 5$ and $\lim_{s\to-\infty} \mathcal{C}(s) = 0$
3. $\mathcal{C}'(s) > 0$ for all $s \in \mathbb{R}$ (strictly monotonic)
4. $\mathcal{C}'(2.5) = 1$ (identity mapping near center)

Proof. Property 1: $\mathcal{C}(2.5) = 2.5 + 2.5 \cdot \tanh(0) = 2.5$.

Property 2: Since $\lim_{x\to\pm\infty} \tanh(x) = \pm 1$, we have $\lim_{s\to\infty} \mathcal{C}(s) = 2.5 + 2.5 = 5$ and $\lim_{s\to-\infty} \mathcal{C}(s) = 2.5 - 2.5 = 0$.

Property 3: $\mathcal{C}'(s) = \text{sech}^2\left(\frac{s-2.5}{2.5}\right) > 0$.

Property 4: $\mathcal{C}'(2.5) = \text{sech}^2(0) = 1$. □

The soft clamping preserves the relative ordering of scores while ensuring boundary constraints are satisfied smoothly, unlike hard clipping which introduces discontinuities in the gradient.

The tanh-based clamping was selected because it provides symmetric behavior around the midpoint ($s = 2.5$), maintains near-identity mapping for scores in the typical range $[1, 4]$, and gracefully compresses extreme values.

D.5 Convergence Analysis

D.5.1 Convergence Conditions

[Convergence of Elo Updates] Let $\{s_t\}_{t=0}^{\infty}$ be the sequence of scores generated by the Elo update rule with adaptive K-factor $K_t = K_{\max} \cdot \gamma^t$ where $\gamma \in (0, 1)$. Assume the LLM judgment is consistent with the ground-truth ordering with probability $p > 0.5$. Then $\{s_t\}$ converges almost surely to a fixed point s^*.

Proof Sketch. The proof relies on three key observations:

Step 1 (Bounded Updates): The update magnitude is bounded by $|K_t(S_A - E_A)| \leq K_t$ since $S_A, E_A \in [0, 1]$. Combined with the soft clamping, scores remain in $[0, 5]$.

Step 2 (Summable Learning Rates): The series $\sum_{t=0}^{\infty} K_t = K_{\max} \sum_{t=0}^{\infty} \gamma^t = \frac{K_{\max}}{1-\gamma} < \infty$, ensuring diminishing step sizes.

Step 3 (Martingale Convergence): Define the expected score change $\Delta_t = \mathbb{E}[s_{t+1} - s_t | s_t]$. Under the consistency assumption, Δ_t has the same sign as $(s^* - s_t)$ where s^* is the true score. By the martingale convergence theorem, $s_t \to s^*$ almost surely. □

D.5.2 Stability Under Noisy Judgments

In practice, LLM judgments contain noise. We analyze robustness by introducing artificial noise:

Noise Level	Final MAE	Rank Correlation	Convergence (%)
0% (ideal)	0.251	0.831	100%
10%	0.267	0.812	99.2%
20%	0.298	0.784	97.6%
30%	0.342	0.751	94.1%

Table 10: Robustness to judgment noise.

The algorithm maintains reasonable performance even with 20-30% judgment error, demonstrating robustness to imperfect LLM evaluations.

E Novelty Assessment: Detailed Methodology

This appendix provides a comprehensive description of the novelty assessment module, including the mathematical formulation, parameter selection rationale, and empirical validation.

E.1 Three-Stage Evaluation Pipeline

The novelty assessment proceeds through three sequential stages, each contributing a distinct aspect to the final evaluation.

E.1.1 Stage 1: LLM-based Initialization

We first obtain an initial novelty estimate $s_{\text{novelty}}^{(0)}$ from the LLM evaluator. This provides a baseline assessment based on the model's understanding of the idea's originality and potential to introduce new insights. The LLM is prompted to rate novelty on a 0-5 scale, considering whether the approach incorporates eye-catching aspects or methods not commonly seen in the field.

The initialization prompt instructs the model to evaluate: (1) whether the core idea represents a departure from existing approaches; (2) the presence of unconventional methodological choices; and (3) the potential to open new research directions.

E.1.2 Stage 2: Similarity-Based Quantification

To objectively measure novelty against the current research landscape, we retrieve semantically related papers using the Semantic Scholar API. Based on the formalized representation of the target idea, we search for the most relevant publications and compute their cosine similarities $\{s_1, s_2, \ldots, s_N\}$ with the target idea using pre-trained embedding models.

Weighted Similarity Metric. We construct a weighted similarity metric that balances maximum and average similarity:

$$s_{\text{weighted}} = \alpha \cdot \max(\{s_i\}) + (1 - \alpha) \cdot \frac{1}{N} \sum_{i=1}^{N} s_i, \tag{12}$$

where α controls the emphasis on the most similar existing work.

Rationale for Weighting. The weighting scheme reflects the intuition that novelty is fundamentally constrained by the most similar existing work: even if an idea differs from most papers, high similarity

Table 11: Effect of weighting parameter α on novelty assessment quality.

α	Hum.Corr.↑	Discrim.↑	Stab.↑
0.0 (avg only)	0.612	0.534	0.91
0.3	0.658	0.587	0.93
0.5	0.691	0.623	0.94
0.6	**0.724**	**0.651**	**0.95**
0.7	0.718	0.642	0.94
0.8	0.687	0.598	0.92
1.0 (max only)	0.621	0.512	0.89

to a single prior publication significantly limits its novelty. Setting $\alpha = 0.6$ emphasizes this constraint while still accounting for the broader similarity landscape.

Inverse Sigmoid Mapping. The weighted similarity is mapped to a novelty score through an inverse sigmoid function:

$$s_{\text{sim}} = 5 \cdot \left(1 - \frac{1}{1 + e^{-k(s_{\text{weighted}} - \mu)}}\right), \quad (13)$$

where k controls the steepness of the transition and μ defines the inflection point. This continuous mapping ensures that high similarity to existing work leads to low novelty scores, while low similarity indicates high novelty.

E.1.3 Stage 3: Continuous Fusion

The final novelty score integrates both the LLM's subjective assessment and the objective similarity-based measurement:

$$s_{\text{novelty}}^{\text{final}} = \beta \cdot s_{\text{sim}} + (1 - \beta) \cdot s_{\text{novelty}}^{(0)}, \quad (14)$$

where β determines the relative contribution of each component.

E.2 Parameter Selection

E.2.1 Weighting Parameter α

We conducted experiments to determine the optimal value of α in Equation 12. The results are shown in Table 11.

The results show that $\alpha = 0.6$ achieves the best correlation with human novelty judgments while maintaining stable assessments.

E.2.2 Sigmoid Parameters k and μ

The parameters k and μ in Equation 13 control the shape of the similarity-to-novelty mapping.

Table 12: Similarity distribution statistics from PAIR-IQ.

Statistic	Value
Mean similarity	0.68
Median similarity	0.71
Standard deviation	0.12
25th percentile	0.61
75th percentile	0.79

Table 13: Effect of steepness parameter k on score distribution.

k	Range↑	Entropy↑	Hum.Corr.↑
6	2.1	1.82	0.687
9	3.2	1.91	0.712
12	**4.1**	**1.98**	**0.724**
15	4.5	1.94	0.718
18	4.7	1.87	0.703

Table 14: Ablation study on fusion weight β.

β	Hum.Corr.↑	Robust.↑	Interp.
0.0 (LLM only)	0.651	0.72	High
0.3	0.689	0.81	High
0.5	0.712	0.87	Medium
0.7	**0.724**	**0.93**	Medium
0.9	0.718	0.95	Low
1.0 (Sim only)	0.698	0.96	Low

Inflection Point μ. The inflection point $\mu = 0.72$ was calibrated based on empirical analysis of similarity distributions in the PAIR-IQ dataset. At this similarity level, ideas typically exhibit moderate overlap with existing work, justifying a neutral novelty score of 2.5. Table 12 shows the similarity distribution statistics.

Steepness Parameter k. The steepness parameter $k = 12$ was selected to ensure meaningful score differentiation across the typical similarity range $[0.5, 0.9]$. Table 13 shows the effect of different k values.

E.2.3 Fusion Weight β

The fusion weight $\beta = 0.7$ was determined through ablation studies comparing different weighting schemes, as shown in Table 14.

Setting $\beta = 0.7$ prioritizes the objective similarity-based measurement while retaining sufficient LLM contribution to capture semantic nuances not reflected in embedding similarity.

E.3 Mathematical Properties

E.3.1 Properties of the Inverse Sigmoid Mapping

The mapping function defined by Equation 13 satisfies several desirable properties.

The mapping function $f : [0, 1] \to [0, 5]$ satisfies:

1. f is strictly monotonically decreasing
2. $f(\mu) = 2.5$ (midpoint property)
3. $\lim_{s\to 0} f(s) = 5$ and $\lim_{s\to 1} f(s) = 0$
4. f is smooth (C^∞) on $(0, 1)$

Proof. Property 1: Taking the derivative,

$$f'(s) = -5 \cdot \frac{k \cdot e^{-k(s-\mu)}}{(1 + e^{-k(s-\mu)})^2} < 0 \tag{15}$$

for all $s \in (0, 1)$, confirming strict monotonic decrease.

Property 2: At $s = \mu$,

$$f(\mu) = 5 \cdot \left(1 - \frac{1}{1 + e^0}\right) = 5 \cdot \left(1 - \frac{1}{2}\right) = 2.5. \tag{16}$$

Property 3: As $s \to 0$, $e^{-k(s-\mu)} \to e^{k\mu} \to \infty$, so $f(s) \to 5$. As $s \to 1$, $e^{-k(s-\mu)} \to 0$, so $f(s) \to 0$.

Property 4: f is a composition of smooth functions (exponential and rational), hence smooth on the interior. □

E.3.2 Sensitivity Analysis

The sensitivity of the novelty score to changes in weighted similarity is given by:

$$\left| \frac{\partial s^{\text{final}}_{\text{novelty}}}{\partial s_{\text{weighted}}} \right| = \beta \cdot \frac{5k \cdot e^{-k(s_{\text{weighted}}-\mu)}}{(1 + e^{-k(s_{\text{weighted}}-\mu)})^2}. \tag{17}$$

This sensitivity is maximized at $s_{\text{weighted}} = \mu$, where small changes in similarity lead to the largest score adjustments, and diminishes toward the extremes, providing natural robustness to outliers.