GUIDE: Towards Scalable Advising for Research Ideas

Yaowenqi Liu*, **Bingxu Meng***, **Rui Pan***, **Yuxing Liu, Jerry Huang, Jiaxuan You, Tong Zhang**
University of Illinois Urbana-Champaign
{yl140, bingxum2, ruip4, yuxing6, jerry8, jiaxuan, tozhang}@illinois.edu

Abstract

The field of AI research is advancing at an unprecedented pace, enabling automated hypothesis generation and experimental design across diverse domains such as biology, mathematics, and artificial intelligence. Despite these advancements, there remains a significant gap in the availability of scalable advising systems capable of providing high-quality, well-reasoned feedback to refine proposed hypotheses and experimental designs. To address this challenge, we explore key factors that underlie the development of robust advising systems, including model size, data reweighting, context length, confidence estimation, and structured reasoning processes. Our findings reveal that a relatively small model, when equipped with a well-compressed literature database and a structured reasoning framework, can outperform powerful general-purpose language models such as Deepseek-R1 in terms of acceptance rates for self-ranked top-30% submissions to ICLR 2025. Moreover, when limited to high-confidence predictions, our system achieves an acceptance rate exceeding 90% on the ICLR 2025 test set, underscoring its potential to significantly enhance the quality and efficiency of hypothesis generation and experimental design.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable progress in tasks ranging from text generation to code synthesis (Achiam et al., 2023). Recently, their application to *academic research assistance*—especially in providing feedback on scientific writing and research ideas—has garnered increasing attention. Systems such as Google's *Co-Scientist* (Gottweis et al., 2025) exemplify a broader shift toward *agentic LLMs* capable of collaborating with human researchers to improve scientific workflows, from hypothesis generation to peer review (Jin et al., 2024a; Tan et al., 2024). This emerging capability holds significant promise for accelerating scientific discovery and democratizing research access.

Among these agentic tasks, one particularly impactful yet underexplored area is the development of *LLM-based advising agents*, models designed to provide detailed, constructive, and *hallucination-free* suggestions for academic papers. The goal of these systems is to emulate human advisors by identifying strengths and weaknesses in submissions, suggesting actionable improvements, and assigning quantitative evaluations. However, existing LLMs often struggle with review fidelity: they may produce inflated ratings, fail to identify methodological flaws, or hallucinate evaluations not grounded in the text (Ye et al., 2024; Yu et al., 2024; Yan, 2024). These limitations stem from a lack of fine-grained supervision, domain-specific alignment, and proper advising-style training data.

To address these challenges, we propose a novel and scalable framework for generating reliable, constructive, and expert-aligned suggestions. Our system is built upon a compressed knowledge base of paper summaries and metadata, distilled from full-text scientific papers in the field of machine learning, which enables efficient and accurate retrieval through a retrieval-augmented generation (RAG) pipeline. Before hypothesis verification, the system retrieves dozens of relevant papers to provide rich external context. Furthermore, to ensure that our models produce high-quality feedback, we introduce a *rubric-guided alignment* strategy that instructs LLMs to follow and apply evaluation criteria akin to those used in major natural language processing (NLP) conferences.

However, even with clear rubrics and guidelines, LLMs still exhibit the tendency to produce overly favorable and superficial revision suggestions. To address this issue, Reward rAnked FineTuning (RAFT; Dong et al., 2023) is used to align an open-source LLM with expert review criteria and domain-specific literature. This alignment enables

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 42813–42834

System	Retrieval-Augmented	Modular Summarization	Rubric-Guided Alignment
MetaGen (Bhatia et al., 2020)	✗	✗	✓
MReD (Shen et al., 2021)	✗	✗	✗
ReviewRobot (Wang et al., 2020)	✓	✗	✗
Reviewer2 (Gao et al., 2024)	✗	✗	✗
CycleResearcher (Weng et al., 2025)	✗	✗	✓
GUIDE	✓	✓	✓

Table 1: **Comparison of LLM-based peer review systems**. While some systems (e.g., CycleResearcher) are a part of broader end-to-end scientific agents, this comparison focuses specifically on their review capabilities.

our model to generate detailed, rubric-grounded feedback, with a particular emphasis on methodological rigor and experimental soundness—areas often neglected by existing systems. The combination of aforementioned techniques gives rise to our advising system: **G**uidelines (Rubrics), **U**nderstanding (Summarized), **I**nformation Retrieval (RAG), **D**irection (Advising Improvement with RLHF), and **E**xplanation (LLM reasoning), or **GUIDE** in short. As shown in Table 1, compared to previous systems, GUIDE is the only one to incorporate all three components of retrieval augmentation, modular summarization, and rubric-guided alignment, highlighting its unique design and broader capability compared to prior approaches.

To evaluate GUIDE, we conduct a controlled experiment using the ICLR 2016–2024 paper submissions dataset, demonstrating the systematic improvements of our methods by predicting the acceptances of ICLR 2025 submissions in the test set. Specifically, we adopt the metrics of Top-5% Precision, Top-30% Precision, and Recall, which respectively evaluate the system's ability to identify high-quality papers, acceptable papers, and retrieve good papers. Additionally, a customized reward model based on rank classification is employed to accelerate the system's intermediate alignment.

Empirical results show that our system fine-tuned on the Qwen-2.5-7B-Instruct backbone model, outperforms large general-purpose language models in terms of rating alignment with actual acceptance. Moreover, rubric-guided prompting focusing on novelty and significance reduces hallucinated content and leads to more grounded, constructive feedback. Overall, our contributions are summarized as follows:

- **End-to-end hypothesis advisor**: We introduce an LLM-based system **GUIDE** that provides actionable suggestions for both *research ideas* and *experimental design*. Our advising system, GUIDE-7B, outperforms large general-purpose LLMs such as GPT-4o-mini (Achiam et al., 2023) and DeepSeek-R1 (Guo et al., 2025) in terms of Top-30% precision—a metric that measures the acceptance rate of the top 30% of papers, as rated based on the suggested strengths and weaknesses.
- **Bayesian Optimized Data Reweighting**: We propose a novel variant of Bayesian-optimized data reweighting algorithms that introduces an additional transformation $\mathcal{T}$ to adapt to the underlying real data weighting distribution. This allows faster empirical convergence and retains the same theoretical guarantees as the vanilla Bayesian Optimization algorithm.
- **Scalable advising with modular summarization**: We show that summarizing different sections of the literature separately effectively mitigates the limited context-length issue in idea advising scenarios, allowing more relevant content to be retrieved and compared for advising. In particular, the abstract and methodology sections are shown to be the most important for evaluating the quality of a paper.

2 Related Works

Hypothesis Discovery in Scientific Research. Recent progress in large language models has enabled their integration into early-stage scientific workflows, particularly in hypothesis generation and ideation (Zhou et al., 2022; Ruan et al., 2024; Singhal et al., 2025; Yu et al., 2025). While these models have demonstrated promise in producing plausible hypotheses (Yao et al., 2023; Tu et al., 2024; Ruan et al., 2024), significantly less attention has been devoted to *hypothesis verification*, the task of evaluating whether hypotheses are substantiated, methodologically sound, and experimentally grounded (Yang et al., 2022; Qiu et al., 2023). Our work advances this understudied area by proposing a rubric-guided framework that evaluates scientific claims that aligned with human peer reviewers.

Jones (2025); Ifargan et al. (2025); Swanson et al. (2024); Saab et al. (2024); Taylor et al. (2022) have proposed retrieval-augmented generation (RAG; Lewis et al., 2020) techniques to improve LLMs' access to external knowledge during hypothesis assessment. However, their methods do not adequately address the compression of retrieved content, leading to inefficiencies in multi-document settings. We introduce a prompt-learning-based compression approach that distills full texts into progressively shorter representations (e.g., summaries, abstracts, and titles) enabling more scalable and interpretable RAG pipelines.

End-to-End AI Scientist Agents. Recent systems such as *Co-Scientist* and *CycleResearcher* aim to operationalize the full scientific lifecycle via autonomous agents (Gottweis et al., 2025; Weng et al., 2025; Lu et al., 2024; Xu et al., 2024) from idea generation to paper drafting and reviewing. While promising in scope, these systems treat review and critique as peripheral components (Skarlinski et al., 2024). While systems such as CycleResearcher (Weng et al., 2025) simulate the research-review loop through reinforcement learning, their reviews are not explicitly aligned with conference specific rubrics and lack grounding. Our system focuses on producing high-quality critique using rubric supervision and retrieval from similar papers, offering more actionable feedback for scientific writing. This specialization allows us to outperform generalist agents in review-centric evaluations and better support iterative paper improvement.

Summary. Our work lies at the intersection of scientific hypothesis verification, automated peer review, and modular AI scientist systems. Departing from approaches that produce surface-level critiques or aim for full-lifecycle coverage, we present a focused, retrieval-augmented, rubric-aligned system that generates structured, high-fidelity scientific feedback.

3 Method

3.1 Problem Definition

Hypothesis evaluation is crucial in the research process. Rather than analyzing the full paper, we focus on assessing the core research hypothesis using four core sections: abstract, claimed contributions, method description, and experimental setup. This approach is especially useful in the early stages of paper writing, when only the research ideas and experimental designs are available.

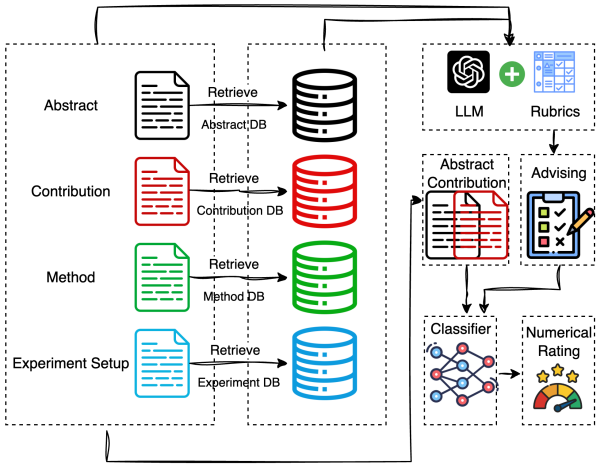

Figure 1: **Overview of GUIDE**, a RAG-based advising system. The GUIDE pipeline begins by receiving the target paper's abstract, contribution, method, and experiment setup. For each of these sections, it retrieves corresponding content from a database of prior works. These target–exemplar pairs are fed into an LLM, which applies predefined rubrics to generate structured advice. Finally, the idea's abstract and contribution, together with this structured advic,e are passed to a lightweight classifier that produces the final numerical rating.

3.2 Retrieval-Augmented Hypothesis Evaluator

Accurately evaluating a research hypothesis requires more than just reading its abstract, claimed contributions, method description and experimental setup. It demands an understanding of how that specific hypothesis fits into the broader scientific literature. To address this, we design a rubric-guided RAG system. An illustration of our RAG pipeline is provided in Figure 1.

Data Collection. To prepare a database for literature comparison, we collected data from ICLR conferences spanning 2016 to 2024, using the publicly available OpenReview platform. For each submission, we extracted the main contribution and section information using a prompt learning algorithm, with further details provided in Appendix C.1.

Retrieval Augmented System. We use OpenAI's `text-embedding-3-large` as our embedding model to build four separate databases, each storing abstract, claimed contributions, method description, and experimental setup respectively from 24,146 ICLR papers submitted between 2016-2024. During inference time, the system takes the target hypothesis's abstract, claimed contribution, method description, and experimental setup, and computes an embedding for each field. The system then queries each corresponding database to retrieve

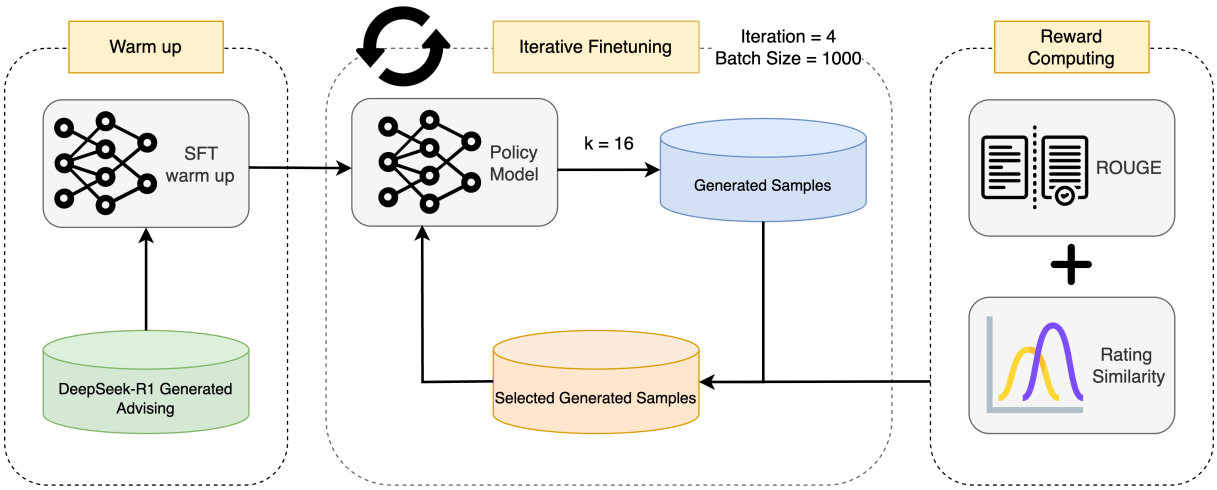

Figure 2: **Training Pipeline of GUIDE**: Overview of the two-stage training process for GUIDE-7B. **Stage 1 (Warming Up)** initializes GUIDE-7B with 4K idea-evaluation examples distilled from DeepSeek-R1, pairing each idea with rubric-based advice. **Stage 2 (RAFT)** further aligns GUIDE-7B with human evaluations by optimizing advice similarity (via ROUGE) and rating similarity (dot product of predicted and actual rating distributions). In RAFT, GUIDE-7B generates and selects the top-k candidate advice responses for additional fine-tuning. Low-rated ideas encourage identification of weaknesses, while high-rated ideas prompt more positive feedback.

the top-k most similar entries based on cosine similarity. The retrieved contexts are then appended to the original abstract, contribution, method, and experiment fields to form the full RAG context.

Rubrics Guided Prompting. During our experiments, we discovered that when evaluating hypothesis, existing LLMs tend to produce overly general feedback and fail to leverage the rich contextual information retrieved by our RAG pipeline. We address this by incorporating a set of evaluation rubrics into our system prompt that were extracted and distilled from the ICLR, ICML, and NeurIPS reviewing guidelines, with a focus on three core dimensions: novelty, significance, and soundness. We also instruct the LLM to partition its feedback into dedicated sections - one each for novelty, significance, and soundness. Each section contains focused commentary aligned with the corresponding rubric, and the final output conforms to our predefined JSON schema. For the full prompt and output specification, please refer to Appendix A

3.3 Reward Ranked Fine-Tuning

LLMs often suffer from implicit biases, leading to suboptimal or skewed outputs. In our experiments, we observed that off-the-shelf models tend to offer only positive and overly general praise and rarely provide neutral or critical feedback. This phenomenon highlights the need for better alignment with human evaluative standards. To improve the alignment of our models, we use Reward-rAnked Fine-Tuning (RAFT; Dong et al., 2023), an iterative fine-tuning algorithm with rejection sampling. Details of the pipeline are in illustrated in Figure 2.

Warming Up. To empower general-purpose small LLMs to learn advising-centered reasoning and output formats, we adopt a warm-up phase at the start of training. Rubrics-prompted DeepSeek-R1 (Guo et al., 2025) is employed to generate evaluations for a randomly sampled subset of ICLR 2024 papers, producing 4,000 high-quality idea–evaluation pairs. These examples are used to perform an initial round of supervised fine-tuning (SFT) of Qwen2.5-7B-Instruct.

Step 1: Generation. After warming up, RAFT (Dong et al., 2023) is applied to further align the model with human preferences, which iteratively optimizes the model via generation, top-k selection, and fine-tuning. At each iteration, the latest model generates $K = 16$ candidate advices for each of 1000 randomly selected ICLR 2024 hypotheses with experimental setups, where the top-1 candidate is selected for each hypothesis.

Step 2: Best-of-N Selection. For each candidate advice a_i, we compute its rating distribution $\hat{d}_i = [\hat{p}_{i,1}, \hat{p}_{i,2}, \ldots, \hat{p}_{i,10}] \in \mathbb{R}^{10}$ by concatenating the advice with the hypothesis's abstract and contributions and pass these contexts through our lightweight classifier, where $\hat{p}_{i,j}$ denotes the probability of assigning rating j to the i-th hypothesis. We construct the human reference distribution in two steps. First, given a set of observed human ratings $\{r_k\}_{k=1}^{K}$ taking values in $\{1, \ldots, 10\}$, the

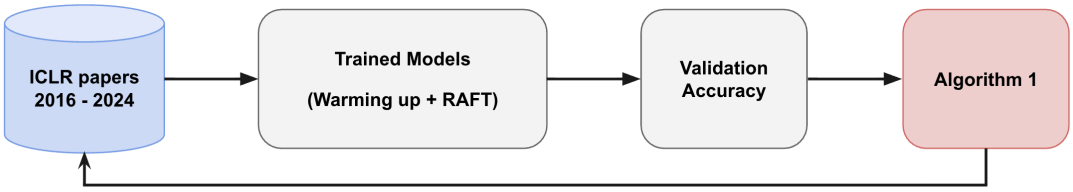

Figure 3: **Data Reweighting** by Bayesian Optimization based on the model's validation accuracy each cycle.

class counts are computed and and normalized into the distribution as follows:

$$c_i = \big|\{k : r_k = i\}\big|, \quad p_i = \frac{c_i}{\sum_{j=1}^{10} c_j},$$

so that $\sum_{i=1}^{10} p_i = 1$. Second, to avoid overly peaked distributions, we apply neighbor smoothing with coefficient $\alpha \in [0, 1]$, defining

$$\tilde{p}_j = \begin{cases} (1-\alpha)p_1 + \alpha\, p_2, & j = 1, \\ (1-\alpha)p_j + \frac{\alpha}{2}\big(p_{j-1} + p_{j+1}\big), & 2 \le j \le 9, \\ (1-\alpha)p_{10} + \alpha\, p_9, & j = 10. \end{cases}$$

We denote the smoothed human distribution for hypothesis i as

$$\tilde{d}_i = [\tilde{p}_{i,1}, \tilde{p}_{i,2}, \ldots, \tilde{p}_{i,10}].$$

Here $\tilde{p}_{i,j}$ is the smoothed probability of rating j. Given the model's predicted distribution $\hat{d}_i = [\hat{p}_{i,1}, \hat{p}_{i,2}, \ldots, \hat{p}_{i,10}]$, the reward is calculated as

$$R_i^{\text{rating}} = \hat{d}_i \cdot \tilde{d}_i = \sum_{j=1}^{10} \hat{p}_{i,j}\, \tilde{p}_{i,j},$$

where the form of weighted-sum avoids gradient vanishing issues in conventional softmax-based loss functions and encourages one-hot prediction.

To further reduce learning difficulty, we introduce an additional text-similarity reward R_i^{text} by measuring the ROUGE score (Lin, 2004) between the generated advice and the concatenation of all reference human reviews. The overall reward is then given by

$$R_i \;=\; \lambda\, R_i^{\text{rating}} \;+\; (1-\lambda)\, R_i^{\text{text}},$$

where $\lambda \in [0, 1]$ balances the two objectives. We select the candidate advice with the highest R_i at each iteration for supervised fine-tuning.

Step 3: Fine-Tuning. After computing the combined reward, the advice a_i^* with highest reward among K candidates are selected, which forms a supervised fine-tuning set $\mathcal{S} = \{(x_i, a_i^*)\}_i$. Here x_i denote the retrieval augmented input for hypothesis i, i.e. the concatenation of the paper's abstract, claimed contributions, method description, experimental setup, and the retrieved summaries from Sec 3.2. Qwen2.5-7B-Instruct is fine-tuned on $\mathcal{S}$. By iterating this generate–select–fine-tuning cycle, the model progressively learns to produce advices that maximize alignment with human judgments and textual fidelity.

Algorithm 1 Bayesian Data Reweighting

Input: the category number of data d; evaluation $h : \Delta^d \to \mathbb{R}$ (in our case, this is validation accuracy), GP kernel $\mathcal{K}$, prior mean μ, Acquisition function, weight transformation $\mathcal{T} : \mathbb{R}^d \to \Delta^d$, iteration T, batch size B
Initialize the Gaussian Process (GP) surrogate model with $\mathcal{K}$ and μ
Initialize sample history $\mathcal{H}$
for $t \leftarrow 1, \ldots, T$ **do**
  for each sample in the batch **do**
    Optimize the batch acquisition function $g_t : (\Delta^d)^B \to \mathbb{R}^B$ to solutions $Z_t \in \mathbb{R}^{d \times B}$, which corresponds to a batch of size B of data weights:
    $$Z_t = \arg\max_{Z \in \mathbb{R}^{d \times B}} g_t(\mathcal{T}(Z)),$$
    Evaluate samples $(z_{t,i}, h(\mathcal{T}(z_{t,i})))$ with $z_{t,i}$ as the i-th column of Z_t ($i \in [B]$)
    Store results in the sample history $\mathcal{H}$
    Update posterior of GP based on $\mathcal{H}$
  end for
end for
Output: $\mathcal{T}(z^*)$ with $z^* = \arg\max_t h(z_t)$

3.4 End-to-End Data Reweighting

Uniform reweighting is normally not the optimal choice for the aforementioned training pipeline in Figure 2. To achieve reasonably good data reweighting in our setting, we introduce a Bayesian optimization (BO) approach as presented in Algorithm 1. Bayesian optimization for data reweighting is similarly to hyperparameter tuning, which

has been widely studied in the literature and proven to be effective (Snoek et al., 2012; Klein et al., 2017; Yen et al., 2025; Chen et al., 2025). Specifically, we consider the reweighting problem over d groups of data based on the publication year of the paper. For each category, we assign a weight $\pi \in [0,1]^d$, which is inside the probability simplex $\pi \in \Delta^d$ such that $\sum_{i=1}^d \pi_i = 1$, to change the sample importance in the training process.

To determine the data weight, Algorithm 1 starts from an initialization based on some prior knowledge about the model. In each iteration t, we choose an acquisition function that applies to the current Gaussian Process (GP) model to determine the weights $\pi_t = \mathcal{T}(z_t)$ that can lead to the largest expected performance improvement compared to the existing choices. Then, we evaluate these new weight samples through the function h (in our process, this is the final top-30% accuracy in the validation set), after which the sample points $(z_t, h(z_t))$ are further incorporated into the GP model. Finally, the weight with the best performance is chosen.

In practice, different grouping fashions can lead to different distributions of optimal weighting schemes. To better adapt to this implicitly unknown distribution, we propose a novel variant of vanilla Bayesian optimization with an additional transformation $\mathcal{T}$ inside the probability simplex Δ^d to capture the differences.

Compared to the standard approach that employs constrained optimization in each iteration to find the next samples π, this transformed method utilizes the geometry of the data reweighting problem and avoids possible numerical instability of constrained optimization. Also, we can prove that this algorithm still enjoys the same theoretical convergence guarantee as BO in (Srinivas et al., 2009), as shown in Appendix G.

Theorem 1. *For any transformation $\mathcal{T} : \mathbb{R}^d \to \Delta^d$, under the same assumptions as Srinivas et al. (2009), the output $h(\mathcal{T}(z^*))$ of Algorithm 1 converges to the optimum $h^* = \min_{\pi \in \Delta^d} h(\pi)$ as $T \to \infty$.*

4 Experiment

4.1 Experiment Setting

We build our retrieval databases using ICLR papers from 2016 to 2024, with a total of 24,146 valid papers. To prevent data leakage, we construct a test set of 1,000 papers randomly sampled from the set of ICLR 2025 submissions. Among these papers, 319 papers were accepted by the ICLR committee, which closely matches the conference acceptance rate of 31.7%. To measure the advising system's alignment with human experts, the following metrics are adopted,

1. **Top-5% Precision:** Among all the hypotheses with the top-5% highest predicted rating, the proportion that were actually accepted.

2. **Top-30% Precision:** Among all the hypotheses with the top-30% highest predicted scores, the proportion that were actually accepted.

3. **Accept Recall:** Among all the hypotheses that were accepted by ICLR 2025, the proportion that appear within the top 30% predictions.

Model	Top-5% Precision	Top-30% Precision	Accept Recall
GPT-4o-mini	70.0 ± 4.6%	47.7± 2.4%	44.8± 2.2%
QwQ-32B	66.7± 1.2%	48.6± 1.5%	45.8± 1.4%
DeepSeek-R1	69.3± 4.6%	50.2± 0.5%	47.2± 0.5%
GUIDE-7B	**72.0± 2.0%**	**51.3± 0.4%**	**48.3± 0.3%**

Table 2: **GUIDE-7B v.s. Deepseek-R1 (671B)**. Performance of advising systems with different backbones on the ICLR 2025 test set over three trials.

4.2 GUIDE-7B v.s. Deepseek-R1 (671B)

To validate the strong advising ability of GUIDE, we compare the predictiveness of its generated advice against that produced by general-purpose LLMs. **Setup**. We compare GUIDE with baselines using various large general-purpose LLMs, including GPT-4o-mini, QwQ-32B, and DeepSeek-R1, all equipped with retrieval-augmented generation, as described in Sec. 3.2. For all LLMs, the decoding hyperparameters are set to temperature $= 0.6$ and top-p with $p = 0.95$. Both GUIDE-7B and baselines will receive the input hypothesis's abstract, claimed contribution, method description, and experimental setup, along with the ten most relevant literature sections from our database.

Results. As shown in Table 2, GUIDE-7B attains the highest Top-30% Precision (51.3%), outperforming all other variants. It is especially intriguing that GUIDE-7B, warmed up using datasets distilled from DeepSeek-R1, can surpass the original DeepSeek-R1. This improvement is largely attributed to the iterative RAFT alignment process, where GUIDE further learns from human preferences and acquires the ability to produce expert-level advice.

4.3 High-Confidence Prediction Improves

Practical real-world applications normally demand high-confidence advising, which calls for further investigation into GUIDE's effectiveness under such conditions. **Setup**. The model's uncertainty is quantified via the Shannon entropy of the predicted rating distribution:

$$H(\hat{d}) = -\sum_{j=1}^{10} \hat{p}_j \log \hat{p}_j.$$

A high entropy indicates that the classifier's probability mass is spread across many rating classes, whereas a low entropy reflects a focused, high-confidence prediction.

To assess the impact of prediction confidence on evaluation accuracy, we rank all hypotheses by ascending entropy and select three subsets corresponding to the lowest-entropy: 10%, 20%, and 30% of papers. Within each subset, we recompute the Top-30% Precision metric, measuring the proportion of truly accepted papers among the top 30%

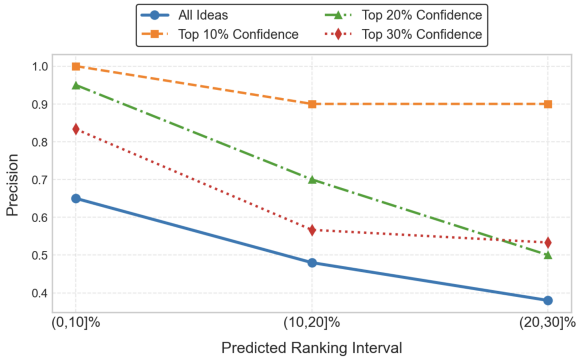

Figure 4: **Over 90% precision with top 10% confidence predictions**. Precision means the ratio of actually accepted papers over all papers that were within the specific confidence and predicted rating ranking interval. Predicted ranking interval means the set of papers sorted in descending order in terms of predicted rating.

Results. Fig. 4 reveals how model confidence modulates evaluation reliability: as the confidence and ranking threshold tightens, we generally observe higher precision, indicating that low-entropy predictions are more trustworthy indicators of acceptance. Notably, within the top 10% confidence subset, the Top-30% Precision reaches the high accuracy of 93.3%, suggesting that even for hypotheses not yet fully formalized into manuscripts, meeting both high-confidence and high-ranking criteria is *a strong indicator of the final acceptance*. In automated hypothesis generation pipelines, where large batches of hypotheses can be generated at low cost, leveraging uncertainty enables the automatic selection of high-quality hypotheses, underscoring the potential for accelerating research discovery.

Method	Top-5% Precision	Top-30% Precision	Accept Recall
Uniform weighting	60.7%	45.0%	42.3%
BO w/ log square norm	64.7%	47.7%	44.9%
BO w/ square norm	65.3%	47.9%	45.0%
BO w/ softmax	**66.0%**	**48.1%**	**45.2%**

Table 3: **Bayesian data reweighting improves**, especially with softmax transformation. Here, the performance is reported on the ICLR 2025 test set.

4.4 Data Reweighting Matters

Our final data reweighting is inspired by the Bayesian optimized weighting solution, which discovers a highly skewed distribution over paper groups.

Setup. Algorithm 1 is implemented based on Botorch (Balandat et al., 2020). We use the Matérn kernel with $\nu = 2.5$ as kernel, q-Noisy Expected Improvement (qNEI) as the acquisition function, and three transform methods: **(1) log-square:** for $\pi = \mathcal{T}(z)$, we have $[\pi]_i = \frac{\log(1+z_i^2)}{\sum_{j=1}^{d} \log(1+z_j^2)}$. **(2) Square:** for $\pi = \mathcal{T}(z)$, we have $[\pi]_i = \frac{z_i^2}{\sum_{j=1}^{d} z_j^2}$. **(3) Softmax:** for $\pi = \mathcal{T}(z)$, we have $[\pi]_i = \frac{\exp(z_i)}{\sum_{j=1}^{d} \exp(z_j)}$.

To accelerate the data reweighting computation, we adopt a low-budget setting, limiting the number of training epochs to 1, with a smaller number of samples 256 in each RAFT iteration. The full details are available in Appendix F.3.

Results. As shown in Table 3, Bayesian optimized data weighting demonstrates significant performance improvements compared to uniform weighting. On top of it, a non-trivial gap is observed between different choices of transformations. Further investigation in Appendix D.5 shows that this gap can become salient under different priors. This highlights the importance of properly chosen transformations, which adapt to the underlying implicit data reweighting and allow faster empirical convergence.

4.5 Database Compression Matter

The compressed database also non-trivially contributes to the scalability of the system, as the retrieved content is summarized in shorter lengths to allow more literature to fit within the limited context window. To empirically verify this claim, we

Retrieved Content	GPT-4o-mini	QwQ-32B	DeepSeek-R1
Full paper	45.0%	44.3%	46.7%
Abstract only	47.0%	45.7%	49.0%
+ Contribution	46.7%	46.0%	48.7%
+ Method	**47.7%**	**48.7%**	49.7%
+ Experiment	**47.7%**	48.3%	**50.3%**

Table 4: **Database Compression Matters**. Ablation on retrieved contents, where top-30% Precision (%) is reported across different backbone LLMs.

conduct ablation studies to compare the system's performance across different types of datasets.

Setup. For all comparisons, the input hypothesis is still formed by abstract, claimed contributions, method description, and experimental setup. The only difference lies in the different retrieval content. All ablation runs use the same three backbone LLMs introduced in Section 4.2: GPT-4o-mini, QwQ-32B, and DeepSeek-R1. We evaluate performance solely via the Top-30% Precision metric on the held-out ICLR 2025 test set.

Results. As shown in Table 4, summarization improves performance by allowing more relevant literature to be retrieved. The abstract and methodology turn out to be the two most conducive sources of information for advising, as the abstract naturally presents the main contribution of the paper, and the methodology contains objective information related to novelty and significance. One surprising observation is that experimental setups sometimes do not help. This is attributed to misaligned settings across the literature, where different works may use different setups to support their claims, while LLMs tend to prefer consistent setups.

5 Discussion

Although our model and system are specifically tailored for academic hypotheses advising rather than full-text academic paper review, in direct comparisons they outperform existing general-purpose LLM-based baselines and achieve precision levels that closely approximate those of human reviewers.

Setup As a benchmark for human agreement, we consider the NeurIPS 2021 consistency experiment (Beygelzimer et al., 2023), which assigns 10 percent of conference submissions to two independent review committees. It is important to note that the human baseline is drawn from the NeurIPS 2021 consistency experiment, which operated at a 24.5 % acceptance rate—substantially lower than the 31.7 % rate of ICLR 2025. As a result, GUIDE's performance should be even better, given that Top-24.5% precision is strictly higher than Top-30%. In addition, the human accuracy, precision, and recall reported here should be interpreted as a rough reference rather than a directly comparable benchmark.

Baselines	Accuracy	F1
Human (NeurIPS)	73.4%	48.4%
AI Scientist w/ DeepSeek-R1	40.7%	49.5%
AgentReview w/ GPT-4.1-nano	41.0%	45.5%
AI Scientist w/ QwQ-32B	42.7%	43.3%
AgentReview w/ DeepSeek-R1	44.2%	46.1%
AgentReview w/ QwQ-32B	49.3%	43.0%
CycleReviewer-8B	57.3%	48.4%
NAIPv1	59.8%	38.2%
AI Scientist w/ GPT-4.1-nano	61.2%	20.8%
GUIDE-7B	**69.1%**	**50.1%**

Table 5: **Reaching near human-level performance** based on results of ICLR 2025 test set. Here **Accuracy** (or Precision) $= \frac{\#\text{Correct Predicted Acceptances}}{\#\text{Predicted Acceptances}}$ represents the proportion of correct accept decisions over all accept decisions made by the system, and **F1** is the harmonic mean of Precision and Recall, with the Recall $= \frac{\#\text{Correct Predicted Acceptances}}{\#\text{Actual Acceptances}}$ being the proportion of correct accept decisions made by the system over all the actual accepted papers.

For baselines, we employ the review agent component of AI Scientist (Lu et al., 2024), CylceResearcher (Weng et al., 2025), AgentReview (Jin et al., 2024b), and NAIPv1 (Zhao et al., 2025). Since GPT-4o-mini tends to reject all papers under AI Scientist's default settings, we replaced it with GPT-4.1-nano to ensure stable results.

Results Table 5 reveals that our idea-centric advisor achieves similar performance as the human baseline in terms of acceptance rate. Moreover, our system outperforms all the aforementioned review agent, underscoring the advantage of using a ranking-based evaluation standard, which more faithfully reflects selective thresholds and yields more balanced, reliable assessments.

6 Conclusion

Our study demonstrates that effective advising in hypothesis generation and experimental design does not necessarily require massive language models. By leveraging a compact model integrated with a compressed literature corpus and structured reasoning mechanisms, we achieve superior performance compared to larger, general-purpose models. The system's high acceptance rates, particularly on high-confidence predictions, highlight its practical utility in supporting scientific inquiry.

Limitations

This system currently focuses exclusively on the machine learning literature, allowing for more targeted retrieval, reasoning, and evaluation within a well-defined domain. By concentrating on a specific field, we are able to better optimize the summarization, advising, and alignment processes. However, extending the system to cover a broader range of scientific disciplines—such as biology, physics, or social sciences—remains an important direction for future work. Such expansion would require addressing additional challenges related to domain-specific terminology, varied writing styles, and diverse evaluation criteria.

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*.

Tuo An, Yunjiao Zhou, Han Zou, and Jianfei Yang. 2024. Iot-llm: Enhancing real-world iot task reasoning with large language models. *arXiv preprint arXiv:2410.02429*.

Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Ben Letham, Andrew G Wilson, and Eytan Bakshy. 2020. Botorch: A framework for efficient monte-carlo bayesian optimization. *Advances in neural information processing systems*, 33:21524–21538.

Alina Beygelzimer, Yann N Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2023. Has the machine learning review process become more arbitrary as the field has grown? the neurips 2021 consistency experiment. *arXiv preprint arXiv:2306.03262*.

Chaitanya Bhatia, Tribikram Pradhan, and Sukomal Pal. 2020. Metagen: An academic meta-review generation system. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1653–1656.

Shengzhuang Chen, Xu Ouyang, Michael Arthur Leopold Pearce, Thomas Hartvigsen, and Jonathan Richard Schwarz. 2025. Admire-bayesopt: Accelerated data mixture re-weighting for language models with bayesian optimization. *arXiv preprint arXiv:2508.11551*.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. 2024. Reviewer2: Optimizing review generation through prompt generation. *arXiv preprint arXiv:2402.10886*.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, and 1 others. 2025. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Tal Ifargan, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony. 2025. Autonomous llm-driven research—from data to human-verifiable research papers. *NEJM AI*, 2(1):AIoa2400555.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024a. Agentreview: Exploring peer review dynamics with llm agents. *arXiv preprint arXiv:2406.12708*.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024b. AgentReview: Exploring peer review dynamics with LLM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Nicola Jones. 2025. Openai's' deep research'tool: is it useful for scientists? *Nature*.

Aaron Klein, Stefan Falkner, Simon Bartels, Philipp Hennig, and Frank Hutter. 2017. Fast bayesian optimization of machine learning hyperparameters on large datasets. In *Artificial intelligence and statistics*, pages 528–536. PMLR.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024. Reducing hallucinations in vision-language models via latent space steering. *arXiv preprint arXiv:2410.15778*.

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*.

Minh Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. 2024. Turning up the heat: Min-p sampling for creative and coherent llm outputs. *arXiv preprint arXiv:2407.01082*.

Rui Pan, Shuo Xing, Shizhe Diao, Wenhe Sun, Xiang Liu, Kashun Shum, Renjie Pi, Jipeng Zhang, and Tong Zhang. 2024. Plum: Prompt learning using metaheuristic. *Preprint*, arXiv:2311.08364.

Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, and 1 others. 2023. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. *arXiv preprint arXiv:2310.08559*.

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. 2024. Liveideabench: Evaluating llms' scientific creativity and idea generation with minimal context. *arXiv preprint arXiv:2412.17596*.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, and 1 others. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2021. Mred: A meta-review dataset for structure-controllable text generation. *arXiv preprint arXiv:2110.07474*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. 2009. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. 2024. The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation. *bioRxiv*, pages 2024–11.

Cheng Tan, Dongxin Lyu, Siyuan Li, Zhangyang Gao, Jingxuan Wei, Siqi Ma, Zicheng Liu, and Stan Z Li. 2024. Peer review as a multi-turn and long-context dialogue with role-based interactions. *arXiv preprint arXiv:2406.05688*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, and 1 others. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, and 1 others. 2024. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. ReviewRobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 384–397, Dublin, Ireland. Association for Computational Linguistics.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. Cycleresearcher: Improving automated research via automated review. In *The Thirteenth International Conference on Learning Representations*.

Yi Xu, Bo Xue, Shuqian Sheng, Cheng Deng, Jiaxin Ding, Zanwei Shen, Luoyi Fu, Xinbing Wang, and Chenghu Zhou. 2024. Good idea or not, representation of llm could tell. *arXiv preprint arXiv:2409.13712*.

Ziyou Yan. 2024. Evaluating the effectiveness of llm-evaluators (aka llm-as-judge). eugeneyan. com.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. 2022. Language models as inductive reasoners. *arXiv preprint arXiv:2212.10923*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Rui Ye, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhenfei Yin, Zhen Xiang, Xiaowen Dong, Jing Shao, and Siheng Chen. 2024. Are we there yet? revealing the risks of utilizing large language models in scholarly peer review. *arXiv preprint arXiv:2412.01708*.

Thomson Yen, Andrew Wei Tung Siah, Haozhe Chen, Tianyi Peng, Daniel Guetta, and Hongseok Namkoong. 2025. Data mixture optimization: A multi-fidelity multi-scale bayesian framework. *arXiv preprint arXiv:2503.21023*.

Haofei Yu, Zhaochen Hong, Zirui Cheng, Kunlun Zhu, Keyang Xuan, Jinwei Yao, Tao Feng, and Jiaxuan You. 2025. Researchtown: Simulator of human research community. *ICML*.

Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, RenJing Cui, Chengcheng Han, Qiushi Sun, and 1 others. 2024. Automated peer reviewing in paper sea: Standardization, evaluation, and analysis. *arXiv preprint arXiv:2407.12857*.

Penghai Zhao, Qinghua Xing, Kairan Dou, Jinyu Tian, Ying Tai, Jian Yang, Ming-Ming Cheng, and Xiang Li. 2025. From words to worth: Newborn article impact prediction with llm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1183–1191.

Ming Zhong, Aston Zhang, Xuewei Wang, Rui Hou, Wenhan Xiong, Chenguang Zhu, Zhengxing Chen, Liang Tan, Chloe Bi, Mike Lewis, and 1 others. 2024. Law of the weakest link: Cross capabilities of large language models. *arXiv preprint arXiv:2409.19951*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and 1 others. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

A Prompts and Output Format

A.1 Inference

In this section, we provide the detailed prompt and also the structured output JSON format.

Prompt Details	SYSTEM: "You are a professional hypothesis evaluator with expertise in machine learning.Your task is to evaluate a given target academic hypothesis step by step, with a focus on novelty, contribution and soundness. You will be given: 1. The hypothesis's title, abstract, claimed contribution, method description, and experimental setup. 2. A set of relevant prior works, each with abstract, claimed contribution, method descriptions and experimental setups. **Review Guidelines** Read the given idea's content: It's important to carefully read through the given content, and to look up any related work and citations that will help you comprehensively evaluate it. Be sure to give yourself sufficient time for this step.**Evaluation Criteria** 1. Motivation / Objective: What is the goal of the paper? Is it to better address a known application or problem, draw attention to a new application or problem, or to introduce and/or explain a new theoretical finding? A combination of these? Different objectives will require different considerations as to potential value and impact. Is the approach well motivated, including being well-placed in the literature? 2. Novelty & Originality: Are the tasks or methods new? Is the work a novel combination of well-known techniques? (This can be valuable!) Is it clear how this work differs from previous contributions? 3. Significance & Contribution: Are the questions being asked important? Does the submission address a difficult task in a better way than previous work? Would researchers or practitioners likely adopt or build on these ideas? 4. Soundness: Can the proposed method and experimental setup properly substantiate the claimed contributions? Will the claims be well supported under the proposed experimental setup? Are the methods used appropriate? Is this a complete piece of work or work in progress?**Related-Works Usage** 1. **Abstract & Contribution**: frame the problem, scope, and high-level "what" and "why." Used for evaluating significance and novelty. 2. **Method**: describe "how" (algorithms, architectures and theoretical derivations). Used for checking whether the proposed method is novel or internally consistent, well-justified, and mathematically rigorous. 3. **Experimental setup**: specify experiment design, datasets, baselines, metrics. Used to evaluate whether this work's experiment is appropriately designed and whether the experiment is comprehensive enough to support the claims. This content may also contain expected results.**Criticality** Noting the idea will become a paper submitted to top conferences with acceptance rate of 30%, you should be more critical. Feel free to give negative evaluations if the idea's quality is poor. For empirical works, there is no need to contain theoretical analysis. For theoretical works, there is no need to contain experimental. Do not give negative evaluations for the two cases.**Output Format** Provide a structured evaluation **strictly in valid JSON format**. Include both an overall evaluation and constructive suggestions for improvement. Do not add explanations, extra text, or Markdown formatting. When mentioning a related work, please use the title of the related work." USER: "**Title**: . . . **Abstract**: . . . **Claimed Contribution**: . . . **Method Description**: . . . **Experimental Setup**: . . . Below are the abstracts of key related works, outlining each study's scope and main findings. Use this section to evaluate the hypothesis's novelty and contributions: . . . Below are the key contributions of selected prior works, highlighting their novel ideas and advancements. Use this section to benchmark the new hypothesis's originality and impact: . . . Below are the methods of key related works, summarizing their technical approaches and algorithms. Use this section to assess the hypothesis's technical novelty, contribution, and soundness: . . . Below are the experimental setups of key prior works, detailing their protocols and evaluation metrics. Use this section to judge whether the proposed experiments are sufficiently sound to support the hypothesis's claims: . . ."
Output Format	{"summary": ". . . ", "comparison with previous works": ". . . ", "novelty": ". . . ", "significance": ". . . ", "soundness": ". . . ", "strengths": ". . . ","weaknesses": ". . . ", "evaluation": ". . . ", "suggestion": ". . . "}

Table 6: Prompt details and JSON output format

A.2 Data Generation

In this section, we provide examples of evolving prompts for contribution extraction, with the following notable trends observed along the iterations of evolution:

- Rubrics and guidelines emerge during the prompt evolution, which provide more specific instructions for LLMs
- More specific and detailed instructions emerge during the optimization, such as where to look for certain statements
- Prompt becomes more concise during the evolution.

Final Prompt @ Iteration 28	You are an expert at extracting contribution statements from academic papers. Follow these instructions carefully: Task: Extract the contribution statements from the introduction section of the provided academic paper. Do not extract from the abstract. Guidelines: - Exclude Abstract: Only consider the introduction section for extraction. - Marked Contributions: If contributions are explicitly marked (e.g., bullet points, numbered lists), copy them exactly. - Include All Points: Ensure every numbered or bulleted point is included. - Unmarked Contributions: If not clearly marked, extract full contribution paragraph(s) near the end of the introduction. - Headers and Phrases: Include headers or introductory contribution phrases. - Single Source: Extract from one continuous block only. - No Inference: Copy text verbatim unless inference is required for 0-label. - Conciseness: Keep instructions clear and direct. - Output Format: {"contribution_label": <0 or 1>, "contribution_text": "<text>"} If contributions are found, set "contribution_label": 1. If not, set "contribution_label": 0 and provide a concise (at most 3 sentence) inferred summary.
Initial Prompt @ Iteration 0	Extract the contribution statements from the introduction section of the given academic paper (exclude the abstract). - Copy Only Continuous Text: - Start at the beginning of the relevant paragraph. - Include the entire paragraph and its final sentence. - Include any headers or introductory phrases. - Output: - If contributions are found, set "contribution_label": 1 and include the extracted text. - If not, set "contribution_label": 0 and provide a concise (at most 3 sentence) inferred summary. - Do Not: - Search beyond the introduction. - Paraphrase or reword. - Hallucinate information. Return the result in JSON: {"contribution_label": <0 or 1>, "contribution_text": "<text>"}

Table 7: Prompt evolution across iterations with structured output format

B End-to-End Example

Idea to evaluate: *Law of the weakest link: Cross capabilities of large language models* (Zhong et al., 2024) (ICLR2025 Poster)

Setup	Evaluation
GUIDE-7B (with rubrics)	**Summary**: Overall, the paper introduces CrossEval, a benchmark to evaluate cross-capabilities of LLMs, and identifies the 'Law of the Weakest Link,' showing that cross-capability performance is constrained by the weakest individual capability. It systematically defines cross-capabilities via taxonomies and conducts extensive evaluations on 17 LLMs. **Comparison with previous works**: Compared to previous works like CMMLU and SKILL-MIX, which focus on individual or combinatorial skills, this work uniquely addresses cross-capabilities. Unlike Corex or AgentBench, which emphasize multi-model collaboration or agent behavior, this paper emphasizes static, task-complexity analysis. While How Abilities in Large Language Models are Affected by Supervised Fine-tuning Data Composition studies capability scaling, this work highlights the importance of addressing weaker individual capabilities. **Novelty**: The paper introduces the 'Law of the Weakest Link' and a systematic framework for defining cross-capabilities via taxonomies. While prior works like GPT-Fathom evaluate LLM capabilities, this work focuses on cross-capability interactions, a novel angle. **Significance**: The findings highlight the critical role of addressing weaker individual capabilities to improve cross-capability performance, influencing model development priorities. The benchmark provides a comprehensive evaluation framework for future research. **Soundness**: The methodology uses expert annotations, multiple model responses, and LLM-based evaluators to ensure reliability. The experiments cover 17 models and 14 capabilities, with correlation analyses supporting claims. However, reliance on human judgments and limited model diversity (no open-source models) may introduce biases. **Strengths**: The strengths of the paper are the comprehensive taxonomy, expert-validated benchmark, systematic evaluation of cross-capability performance, and clear demonstration of the 'Law of the Weakest Link.' The inclusion of both static and dynamic (principle-based prompting) analyses strengthens the claims. **Weaknesses**: The weaknesses of the paper are the lack of diversity in evaluated models (all commercial), potential bias in taxonomy validation, and insufficient comparison with real-world tasks. The cross-capability definitions are abstract, and the experiments do not test cross-capability transfer to real-world scenarios. **Evaluation**: In conclusion, the paper makes a valuable contribution by systematically studying cross-capabilities and empirically validating the 'Law of the Weakest Link.' While the methodology is rigorous, broader model diversity and real-world validation would strengthen its impact. **Suggestion**: To improve the paper, the authors could expand model diversity to include open-source models, validate the taxonomy with real-world task datasets, and compare with benchmarks like AgentBench to demonstrate practical relevance.

Table 8: Evaluation and suggestion generated by GUIDE-7B.

Setup	Evaluation
DeepSeek-R1 (with rubrics)	**Summary**: Overall, the paper proposes the concept of cross capabilities in LLMs, introduces the CrossEval benchmark to evaluate them, and identifies the 'Law of the Weakest Link' phenomenon where cross-capability performance is constrained by the weakest individual capability. The methodology includes taxonomy construction, human-annotated benchmarks, and experiments across 17 models. **Comparison with previous works**: Compared to previous works like SKILL-MIX and L-Eval, which focus on skill combinations and long-context evaluation, this paper uniquely defines cross capabilities through structured taxonomies and emphasizes the bottleneck effect of weak individual skills. While Corex explores multi-model collaboration for complex tasks, this work focuses on intrinsic capability interactions within single models. SCALE and CMMLU address multilingual or domain-specific evaluations but do not systematically study cross-capability dependencies. **Novelty**: The paper introduces a novel framework for cross capabilities, combining predefined individual skills into paired tasks, and provides empirical evidence for the 'Law of the Weakest Link.' While skill combination is explored in SKILL-MIX, the structured taxonomy and weakest-link analysis are new contributions. **Significance**: The work highlights a critical limitation in LLMs' real-world applicability, offering a benchmark and actionable insights for improving weak capabilities. The findings could influence model development and evaluation practices, particularly in multi-dimensional tasks. **Soundness**: The methodology is rigorous, with 1,400 expert-annotated prompts, 8,400 human ratings, and evaluations across 17 models. Experiments include ablation studies and system-prompting interventions to validate claims. However, cross capabilities are limited to pairwise combinations, and the selection of individual capabilities (e.g., Spanish) lacks justification compared to core skills like reasoning. **Strengths**: The strengths of the paper are the systematic taxonomy of cross capabilities, the human-annotated CrossEval benchmark, the identification of the weakest-link phenomenon, and comprehensive experiments validating the findings across diverse models and evaluators. **Weaknesses**: The weaknesses of the paper are the limited scope of cross capabilities (only pairwise combinations), the inclusion of Spanish as a standalone capability without clear rationale, and insufficient comparison to SKILL-MIX's k-skill combination approach. The experiments do not explore mitigation strategies beyond system prompting. **Evaluation**: In conclusion, the paper makes a valuable contribution by formalizing cross capabilities and identifying a critical performance bottleneck in LLMs. While the benchmark and findings are impactful, the novelty is partially limited by prior work on skill combinations. **Suggestion**: To improve the paper, the authors could expand cross capabilities to include >2 skills, justify capability selection (e.g., Spanish vs. multilingual reasoning), compare with SKILL-MIX's combinatorial approach, and explore methods to mitigate the weakest-link effect beyond prompting.

Table 9: Evaluation and suggestion generated by DeepSeek-R1.

C Additional System Details

C.1 Data Collection & Generation

To prepare a database for literature comparison, we collect data from ICLR conferences spanning 2016 to 2024, from the publicly available OpenReview platform. For each submission, we obtained paper metadata (e.g., title, authors, abstract), full-text PDFs, official reviewer comments, and author rebuttals. Using a custom-built data cleaning pipeline, we processed these raw inputs into a structured database suitable for downstream use in both our RAG framework and in RAFT post-training (Dong et al., 2023). To convert full paper PDFs into markdown-formatted text, we utilized the open-source tool MinerU (Wang et al., 2024), which enables reliable text extraction and structural segmentation.

An essential step in our preprocessing pipeline is content compression through summary generation. To this end, we used `gpt-4.1-nano`, a cost-effective yet high-performing model from OpenAI (Achiam et al., 2023), to generate structured section-wise summaries for each paper (e.g., Introduction, Related Work, Methodology, Experiments). This summarization reduced the input length of each paper by approximately $16\times$, allowing us to incorporate substantially more context within the RAG input window, mitigating the token limit bottleneck and improving retrieval efficiency in downstream tasks.

A crucial component of our data generation pipeline is *contribution statement extraction* since contribution is considered the essence of a paper's strengths. This task is particularly challenging for two reasons: (1) not all papers explicitly state their contributions, and (2) such statements are rarely identifiable via simple rule-based or string-matching methods. To address this, we formalize the task as a sentence-level extraction problem over the full text of a paper in markdown format, where the objective is to identify explicit statements of the paper's key contributions. We assign a label of 1 if the contribution statement is explicitly present and correctly extracted, and 0 if the language model must infer it due to its implicit or absent formulation.

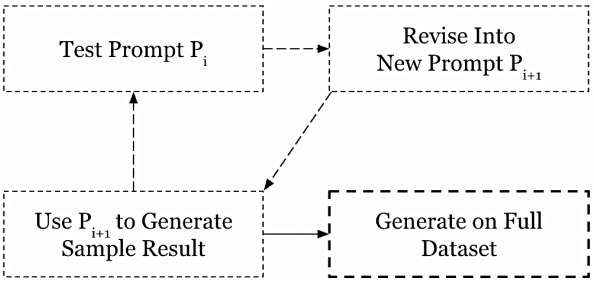

Figure 5: Contribution extraction with learned prompts.

We employ OpenAI's cost-efficient model `gpt-4o-mini` to perform this extraction, leveraging self-consistency decoding and prompt optimization strategies as outlined in Pan et al. (2024) and highlighted in Figure 5. To systematically optimize our prompts, we construct a validation set of 100 human-annotated papers containing gold-standard contribution statements. We then apply a genetic algorithm to evolve prompts over successive generations. In each iteration, candidate prompts are scored based on the similarity between the extracted and ground-truth contribution statements, measured via metrics such as longest common subsequence (LCS) or Levenshtein distance. The top-k prompts are selected and probabilistically propagated using Boltzmann-weighted sampling with $k_BT = 1$, guiding the evolutionary process toward higher-performing prompts across more than 100 trials. The resultant prompt achieves 94% accuracy with an 80% match based on Levenshtein distance or a 30% match based on LCS. Examples of prompt evolution at both the initial and final iterations are provided in Appendix A.2. A more detailed description of the genetic algorithm used for prompt engineering is also included.

Once the learned prompt is obtained, we employ it across all ICLR papers in our dataset. If a contribution statement is explicitly presented in the paper, it is labeled as 1; otherwise, it is labeled as 0. This automated step significantly improves scalability and accuracy over naive prompting or manual annotation, enabling high-quality contribution extraction at scale.

D Additional Experimental Results

D.1 The Effect of Different Databases

To further evaluate the robustness of our system, we conduct additional experiments using the arXiv paper database for retrieval. Specifically, we construct a retrieval corpus comprising all papers in the cs.LG (machine learning) category up to December 2024. For each target paper, retrieval is restricted to papers published prior to the target, determined by the earlier of its ICLR submission date or the earliest arXiv posting. To prevent retrieval-level contamination, we also apply a ROUGE-L similarity check to ensure that the target paper itself is never retrieved as supporting context.

The evaluation results are shown in Table 10. GUIDE-7B, without any additional fine-tuning on the arXiv database, continues to outperform large-scale general-purpose language models, demonstrating its robustness to distribution shifts in the retrieval corpus.

Model	Top-5% Precision	Top-30% Precision	Accept Recall
Qwen2.5-7B-Instruct	60.0%	43.0%	40.4%
GPT-4o-mini	70.0%	48.0%	45.1%
QwQ-32B	68.0%	49.0%	46.1%
DeepSeek-R1	70.0%	51.0%	48.0%
GUIDE-7B	**74.0%**	**52.3%**	**49.2%**

Table 10: Performance of different models when retrieving from the arXiv cs.LG category.

It is worth noticing that GUIDE-7B provides nearly 10% improvement compared to its base Qwen2.5-7B-Instruct model in almost all metrics. This demonstrates the significance of our training paradigm.

D.2 The Quality Loss of Summarization

To quantify the impact of modular summarization on advising performance, we conduct ablation experiments in which the number of retrieved papers is limited to 1–2, in order to control the total context length to 15k tokens, which is the same as in previous experiments. Table 11 presents the Top-30% Precision for each model when using either full papers or summaries as retrieved context.

As shown in Table 11, the performance loss from applying summarization is negligible across all models. This suggests that modular summarization successfully preserves the key aspects of contribution and novelty, while significantly reducing context length.

Model	GPT-4o-mini	QwQ-32B	DeepSeek-R1
Full Paper	45.0%	44.3%	46.7%
Summary	44.7%	44.0%	46.7%

Table 11: Top-30% Precision with full paper vs. summary as retrieved context. The context length for full papers is limited to 15K tokens, and each summary corresponds to its full paper version during retrieval.

D.3 Rating Distribution

We also analyze the distribution of predicted ratings to assess whether LLM-based systems exhibit optimism bias compared to human reviewers.

Evaluator	Mean	Variance
GPT-4o-mini	6.40	0.10
QwQ-32B	6.17	0.26
DeepSeek-R1	6.04	0.19
GUIDE-7B	5.81	0.22
Human	5.13	1.46

Table 12: Mean and variance of ratings assigned by human reviewers and different LLMs. GUIDE-7B significantly mitigates the optimism bias compared to existing general-purpose LLMs.

Table 12 shows that human reviewers give the lowest average score (5.13) but with the greatest spread in their judgments (variance = 1.46). All LLM-based evaluators assign higher mean ratings, yet their variances (0.10–0.26) are much lower, indicating they rate more consistently but also more leniently. Among them, GUIDE-7B comes closest to human behavior, with a mean of 5.81 and a variance of 0.22.

To offer deeper insights into the behavior of different systems, we also provide a more granular visualization of the full rating distributions. By plotting the proportion of scores falling into each interval for both human reviewers and the model evaluators, the distinct rating behaviors and biases become immediately apparent and easier to interpret.

As illustrated in Table 6, two key observations emerge. First, LLM-based systems exhibit a tendency to assign ratings that are concentrated near the mean, in contrast to human ratings, which are more evenly distributed across different intervals. Second, general-purpose LLMs such as GPT-4o-mini and QwQ-32B demonstrate a pronounced optimism bias, frequently assigning higher ratings in the [6.5, 7) range. In contrast, GUIDE-7B sig-

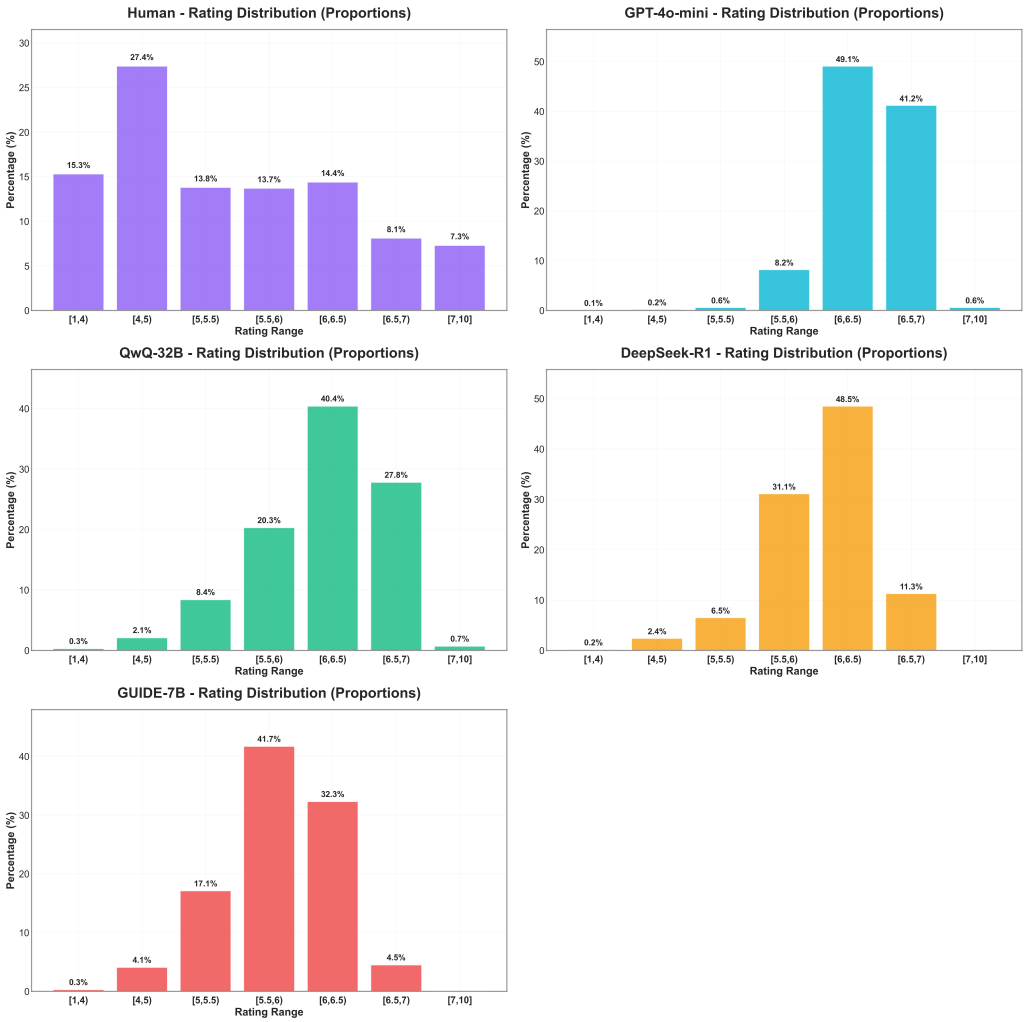

Figure 6: Distribution of ratings for human reviewers and various LLMs. General-purpose LLMs show optimism bias with ratings skewed toward higher intervals, while GUIDE-7B yields a more balanced, human-like rating distribution.

nificantly mitigates this optimism bias, producing more balanced, borderline ratings in the [5.5, 6) interval for most papers. These findings highlight both the differences in rating behaviors between LLM-based systems and human evaluators, as well as the effectiveness of GUIDE-7B in reducing optimism bias.

D.4 Rubrics-Guided Prompting

To quantify the effectiveness of rubric prompting and analyze the source of these potential improvements, we compare different rubrics under various backbones.

Setup. In this experiment, we fix the retrieved contents to be the same, all with 10 related abstracts, 10 contributions, 10 method summaries, and 10 experimental setups. The only difference across variants is the system prompt, which directs the LLM to emphasize a specific rubric (e.g., significance, novelty, soundness) and output the corresponding aspect-focused evaluation. Since QwQ-32B exhibits relatively weaker instruction-following capabilities under prompting, it is replaced with another widely adopted Gemini-flash-2.0 to ensure robust adherence to our system prompts.

Prompts	GPT-4o-mini	Gemini-flash-2.0	DeepSeek-R1
No rubrics	45.3%	47.0%	48.0%
+ Soundness only	44.7%	43.3%	47.3%
+ Novelty only	47.3%	48.3%	49.3%
+ Significance only	47.7%	48.3%	49.3%
+ All	**47.7%**	**49.7%**	**50.3%**

Table 13: Top-30% Precision (%) with rubric prompts.

Results. Significance and Novelty rubrics yield non-trivial gains in Top-30% Precision, while Soundness guidance hurts performance. This phenomenon indicates that the general-purpose LLMs are still lacking the ability to assess experimental rigor. Overall, rubric prompts demonstrably en-

hance hypothesis evaluation, with the full system benefiting most from Significance and Novelty instructions.

Case Study. As shown in Table 13, we observe that the significance rubric yields the most pronounced improvement in evaluation quality. To demonstrate the effectiveness of this rubric-guided approach, concrete examples are presented to illustrate the model's assessment with and without significance, novelty, or soundness rubrics applied to an ICLR2025 hypothesis, as shown in Table 14, 15, and 16 individually.

Hypothesis: *Turning Up the Heat: Min-p Sampling for Creative and Coherent LLM Outputs* (Nguyen et al., 2024) *(ICLR2025 Oral)*

	Final Evaluation	Predicted Rating
With Significance Rubrics	. . . It addresses a well-known problem in LLM decoding and offers a simple yet effective improvement over existing truncation methods, likely to be adopted widely.	6.74
No Rubrics Given	. . . While its empirical validation is thorough, the lack of theoretical grounding limits its conceptual novelty.	6.02

Table 14: **Case Study: Significance**. Comparison of evaluation outcomes with and without significance rubrics, showing that rubrics guide the model to correctly identify high-impact contributions.

Title: *IoT-LLM: Enhancing Real-World IoT Task Reasoning with Large Language Models* (An et al., 2024)

	Final Evaluation	Predicted Rating
Human	". . . I do not see substantial technical novelty in this study. . . the paper's novelty appears limited in this regard, . . . "	3.5
Without Novelty Rubrics	" . . . The strengths of the paper are: 1) Novel integration of IoT data preprocessing with LLMs, addressing a critical gap in physical world reasoning. . . "	5.8
With Novelty Rubrics	" . . . In conclusion, the paper presents a moderately novel application-focused framework with practical value in IoT-LLM integration . . . "	4.9

Table 15: **Case Study: Novelty**. Novelty rubrics help prevent overrating papers with incremental contributions.

Title: *Reducing Hallucinations in Large Vision-Language Models via Latent Space Steering* (Liu et al., 2024)

	Final Evaluation	Predicted Rating
Human	". . . The analysis in the paper is thorough . . . "	7.5
Without Soundness Rubrics	". . . In conclusion, the paper presents a conceptually novel and empirically validated approach to LVLM hallucination reduction . . . "	6.3
With Soundness Rubrics	". . . The idea is promising but insufficiently validated. While the approach is novel, the lack of comparison to state-of-the-art methods. . . "	5.2

Table 16: **Case Study: Soundness**. Soundness rubrics can negatively impact the system's performance by imposing overly strict criteria on experimental design.

D.5 Bayesian Optimized Data Weights

As shown in Figure 7, Bayesian optimized data weights demonstrate a common pattern of preferring papers in later years for training the model. This is not surprising, since many topics in ICLR only emerged recently, such as large language models and LLM agents. Only by being trained on these types of samples can the advising model know how to provide proper advice on those topics.

In addition, the specific choice of transformation plays a crucial role. As illustrated in Figure 8, when the prior distribution is selected to be uniform, the behaviors of different transformations diverge significantly. In particular, both the square norm and the logarithmic square norm transformations show very limited capacity to capture meaningful patterns from the data. In contrast, the softmax transformation demonstrates a notably faster convergence and learns more efficiently under the same conditions, achieving substantially better reweighting schemes.

This comparison clearly reinforces the critical importance of the additional transformation $\mathcal{T}$ within our framework. It is not merely an auxiliary operation but a key component that governs how the reweighting mechanism interprets and adjusts data representations. The consistent improvement observed with the softmax transformation underlines the sensitivity of our approach to the choice of $\mathcal{T}$, thereby emphasizing the novelty and effectiveness of our proposed reweighting algorithm in leveraging this transformation to achieve superior performance.

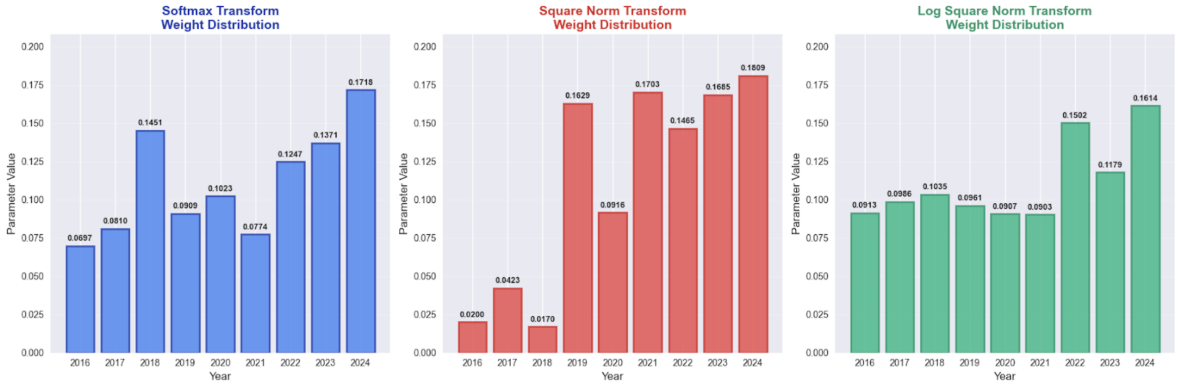

Figure 7: Learned data weights with different transformations in Bayesian Optimization.

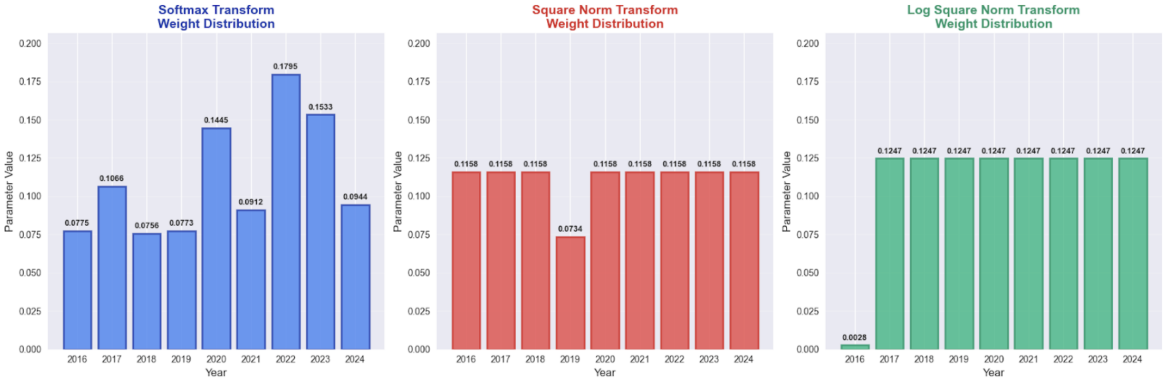

Figure 8: Learned data weights with different transformations in Bayesian Optimization, when being initialized with a prior of uniform distribution.

E Dataset Analysis

Dataset Size. Our dataset comprises all available ICLR papers from 2016 through 2025 (34,632 papers in total). Of these, we adopt the 2016–2024 papers (24,146 papers) as our retrieval database. Figure 9 shows the annual publication counts. **Database Information**. We measured the token lengths of four sections: *Abstract*, *Claimed Contribution*, *Method Description*, and *Experimental Setup* across our retrieval database (24,146 papers) using the GPT-4 tokenizer (via the `tiktoken` library). Table 17 reports the average token counts.

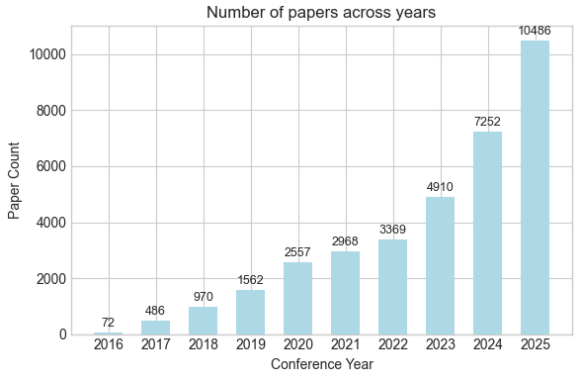

Figure 9: Annual ICLR paper counts from 2016 to 2025, illustrating rapid growth in recent years. Papers from 2016–2024 serve as our retrieval database.

Section	Avg. # tokens
Abstract	229.0
Claimed Contribution	212.8
Method Description	201.9
Experimental Setup	193.0

Table 17: Average token lengths per section (GPT-4 tokenizer).

F Experimental Details

F.1 Training

We initialize GUIDE-7B from Qwen2.5-7B-Instruct and perform a two-stage training procedure.

Warm-up. For the warm-up stage, we use our RAG system with DeepSeek-R1 as the backbone to synthesize a high-quality dataset of $4,000$ samples.

Training is carried out on 4x NVIDIA A100 40 GB GPUs with DeepSpeed's ZeRO 3 optimizations and CPU offload enabled. We train for 3 epochs using a batch size of 16, an initial learning rate of 1e-6 with a cosine decay schedule over a 15K context window.

RAFT. Our RAFT pipeline consists of three phases:

- **Generation:** For each iteration, we use vLLM to sample 1,000 ICLR 2024 papers and generate $K = 16$ candidate evaluations per hypothesis with temperature 0.7, top-$p = 0.8$, and repetition penalty $= 1.05$.
- **Reward Computation:** We smooth human rating distributions with neighbor coefficient $\alpha = 0.4$ and compute the combined reward

  $$R_i = \lambda\, R_i^{\text{rating}} + (1-\lambda)\, R_i^{\text{text}} \quad \text{with} \quad \lambda = 0.7.$$

  The text-similarity reward R^{text} is the sum of ROUGE-1, ROUGE-2, and ROUGE-L, each weighted by 0.1 (total weight 0.3).
- **Supervised Fine-Tuning:** We keep fine-tuning our warmed-up model via LoRA with rank $r = 64$ and alpha $= 64$, using learning rate 1×10^{-5}, batch size 16, for 2 epochs, and a cosine learning-rate schedule. We set the context window equal to 15k tokens. Training runs on 2 NVIDIA A100 40 GB GPUs with DeepSpeed's ZeRO 3 optimizations and CPU offload enabled.

Training Results:. We ran 4 RAFT iterations. Figure 10 shows how the average reward and best reward evolved over these iterations.

Figure 10: Average and best reward over successive RAFT iterations. Both metrics improve as training progresses, indicating the effectiveness of the RAFT pipeline for optimizing model performance.

F.2 Data Generation

Details of the genetic algorithm's hyperparameter settings are:

- **Population:** All previous prompts form the population.
- **Fitness Function:** Accuracy based on 100 human-annotated gold labels (contribution detection correctness).
- **Selection:** Top-K (5) prompts are selected using weighted sampling.
- **Crossover (Recombination):** Performed by o1-mini to revise and combine prompts.
- **Mutation:** N/A — no random or absurd changes were introduced.
- **Iteration:** 28 iterations were performed before stopping.

All prompts are formulated as discrete, human-readable strings.

F.3 Bayesian Optimization

We apply Bayesian optimization to adaptively search for the optimal data reweighting coefficients across $d = 9$ categories. The detailed hyperparameters and implementation settings are summarized in Table 18.

Table 18: Hyperparameters used in Bayesian Data Reweighting.

Name / Symbol	Value
Initial samples N_{init}	5
Iterations T	5
Batch size B	3
Mean function	`MonotonicMean`
Kernel	`MaternKernel(`ν`=2.5, ARD)`
Likelihood noise range	[1e-6, 1e-1]
Target normalization	`standardize(train_y)`
Acquisition Function	`qNoisyExpectedImprovement`
Baseline	`X_baseline = train_x`

To accelerate the training, we adopt a setting with lower budget, as illustrated in Table 19, while keeping other hyperparameters the same as in Appendix F.1.

Table 19: Hyperparameters used for warm-up SFT and RAFT in the Bayesian Data Reweighting setting.

Stage	Name / Symbol	Value
Stage 1: Warm-up SFT	Number of Epochs	1
Stage 2: RAFT	Number of Epochs	1
	Number of Samples	256
	Number of Iterations	3
	Best-of-K	8

G Theory of BO Convergence

Proof of Theorem 1. For any transformation $\mathcal{T} : \mathbb{R}^d \to \Delta^d$ and target evaluation function $h : \Delta^d \to \mathbb{R}^d$, there exists a function $\phi : \mathbb{R}^d \to \mathbb{R}$ such that for all $z \in \mathbb{R}^d$, it holds that

$$\phi(z) = h(\mathcal{T}(z)).$$

Then we consider ϕ to be the real black-box function we are trying to model through BO. Under the same settings as presented in Srinivas et al. (2009), we assume that ϕ is drawn from a GP whose kernel has bounded variance σ, and we choose the acquisition function to be

$$g_t(z) = \mu_{t-1}(z) + \sqrt{\beta_t}\sigma_{t-1}(z),$$

where we take $z \in [0, c]^d$, and μ_{t-1} to be the GP mean, σ_{t-1} to be the GP variance, and β_t to be a constant at iteration t, and c to be a constant. Then, based on Theorem 2 presented in Srinivas et al. (2009), we have that

$$\phi^* - \max_{t\in[T]} \phi(z_t) \leq \mathcal{O}\left(\sqrt{\frac{d\gamma_T}{T}}\right),$$

where γ_T is the maximum information gain after T iterations, determined by the kernel choice of the BO process. Specifically, for the Matérn kernel with the smooth controlling parameter $\nu > 1$ we use in our experiments, it holds that

$$\gamma_T = \mathcal{O}\left(T^{\frac{d(d+1)}{2\nu+d(d+1)}} \log T\right),$$

which results in the convergence guarantee of

$$\phi^* - \max_{t\in[T]} \phi(z_t) \leq \mathcal{O}\left(\sqrt{dT^{-\frac{2\nu}{2\nu+d(d+1)}}}\right),$$

which converges to 0 as $T \to \infty$.

Therefore, we show the theoretical correctness of our transformation applied to the Bayesian optimization. □

H Broader Impacts

AI advisors can accelerate research progress by offering automated guidance, which supports researcher education and speeds up the development of new ideas. On the downside, the scoring capabilities of AI advisors could be misused in conference review processes. To prevent such misuse, we will release the scoring system only under appropriate regulatory frameworks.

I Human Annotation for Contribution Extraction

The 100 annotated contributions in Appendix C.1 were manually generated by a PhD student, who is one of the authors of this paper.

J AI Usage

ChatGPT is used to correct grammatical errors and polish the paper writing.