GUIDE: Towards Scalable Advising for Research Ideas

Anonymous ACL submission

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, 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., 2024; 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

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., 2024)	✗	✗	✓
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.

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.

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.

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.

- **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.

- **Rubric-guided alignment**: We demonstrate that integrating rubric-based instruction significantly enhances the reviewer expertise of LLMs and improves evaluation usefulness.

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). 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 in a manner 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., 2024; 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., 2024) simulate the research-review loop through reinforcement learning, their reviews are not explicitly aligned with conference specific rubrics and lack retrieval 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 a crucial component of AI for Science. Rather than performing a full paper scan, our task focuses on assessing the core research hypothesis using four summarized sections: abstract, claimed contributions, method description, and experimental setup. This approach is especially useful in the early stages of paper writing, when only the outline of research ideas and experimental designs is available.

3.2 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.

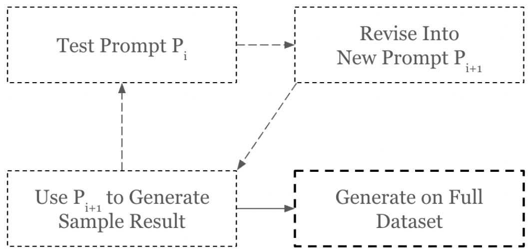

Figure 1: Contribution extraction with learned prompts.

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.

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 1. 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 $k_B T = 1$, 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.

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.

3.3 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 2.

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 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.

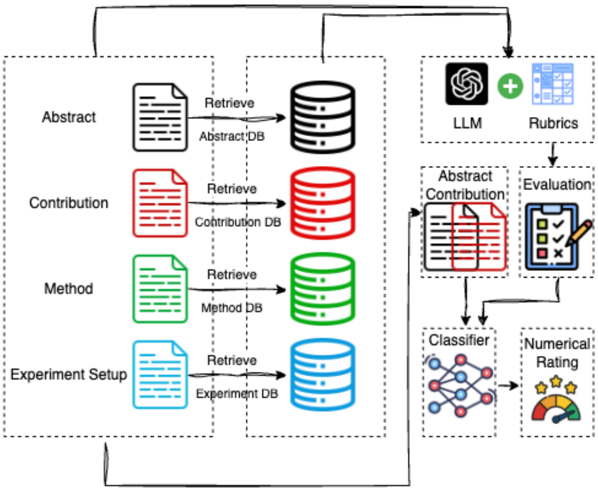

Figure 2: GUIDE: a RAG-based Advising System.

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

3.4 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 3.

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-

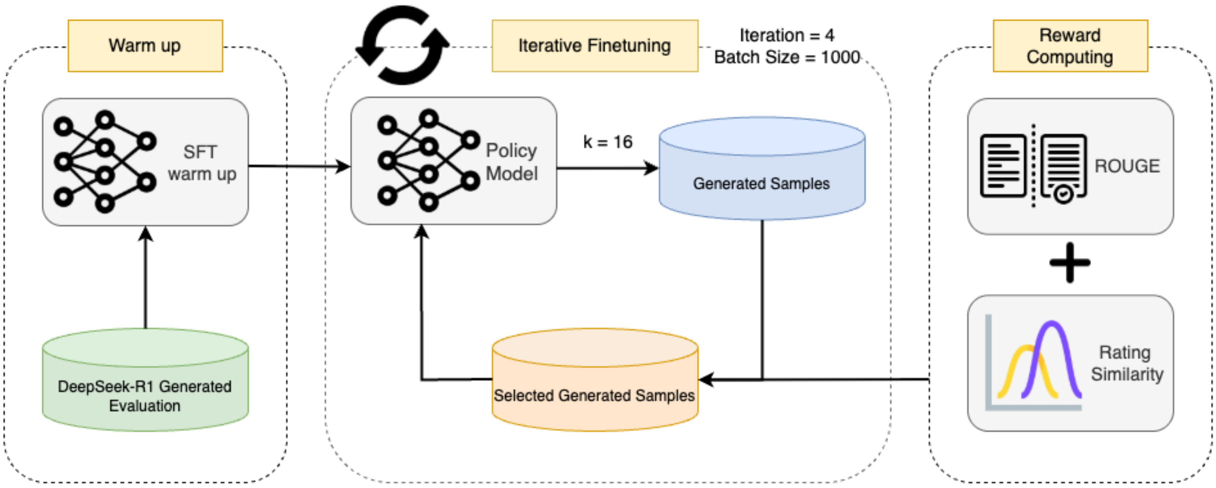

Figure 3: RAFT Pipeline

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 hypothesis with experimental setups.

Step 2: Top-K Selection For each candidate advice a_i, we compute its rating distribution $\hat{d}_i = [\hat{p}_{i,1}, \hat{p}_{i,2}, ..., \hat{p}_{i,10}] \in \mathbb{R}^{10}$ 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 $\{1, \dots, 10\}$, the 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

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

Here $\tilde{p}_{i,j}$ is the smoothed probability of rating j. Given the model's predicted distribution $\hat{d}_i = [\hat{p}_{i,1}, \hat{p}_{i,2}, \dots, \hat{p}_{i,10}]$, 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 3.3. 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.

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.

4.2 GUIDE-7B v.s. Deepseek-R1

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. 3.3. 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.

Model	Top-5% Precision	Top-30% Precision	Accept Recall
GPT-4o-mini	68.0%	47.7%	44.8%
QwQ-32B	68.0%	48.3%	45.5%
DeepSeek-R1	64.0%	50.3%	47.3%
GUIDE-7B	**72.0%**	**51.7%**	**48.6%**

Table 2: Performance of advising systems with different backbones on the ICLR 2025 test set.

Results As shown in Table 2, GUIDE-7B attains the highest Top-30% Precision (51.7%), 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 Scalable Advising with Modular Summarization

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 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.

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 3: Ablation on retrieved contents: Top-30% Precision (%) across different backbone LLMs.

Results As shown in Table 3, 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.

4.4 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 4: 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 enhance hypothesis evaluation, with the full system benefiting most from Significance and Novelty instructions.

Case Study As shown in Table 4, we observe that the significance rubric yields the most pronounced improvement in evaluation quality. To demonstrate the effectiveness of this rubric-guided approach, a concrete example is presented to illustrate the model's assessment with and without significance rubrics applied to an ICLR2025 Oral hypothesis, as shown in Table 5.

Hypothesis: *Turning Up the Heat: Min-p Sampling for Creative and Coherent LLM Outputs (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 5: **Case Study:** Comparison of evaluation outcomes with and without significance rubrics for an ICLR2025 Oral hypothesis, demonstrating that rubrics guide the model to correctly identify high-impact contributions.

4.5 Uncertainty Analysis

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% of model-ranked hypotheses.

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

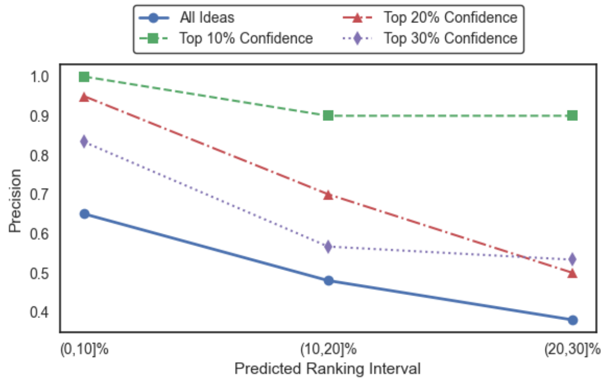

Figure 4: **Uncertainty Analysis**: Precision means the ratio of actually accepted papers over all papers that were within the specific confidence and predicted rating ranking interval. Predicted ranking interval means the set of papers sorted in descending order in terms of predicted rating.

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.

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.

For a strong and well-known baseline, we employ the review agent component of AI Scientist (Lu et al., 2024). It processes the entire paper through multiple rounds of reflection and aggregates outputs via an ensemble of model passes to produce its final evaluation. By comparing against this full-text, multi-stage baseline, we demonstrate the advantages of our hypotheses-centric advisor. 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.

Baselines	**Accuracy**	**F1**
Human (NeurIPS)	73.4%	48.4%
AI Scientist with DeepSeek-R1	40.7%	49.5%
AI Scientist with QwQ-32B	42.7%	43.3%
AI Scientist with GPT-4.1-nano	61.2%	20.8%
GUIDE-7B	**69.1%**	**50.1%**

Table 6: Performance of baselines on the ICLR 2025 test set.

Results Table 6 reveals that our idea-centric advisor achieves similar performance as the human baseline in terms of acceptance rate. Moreover, our system outperforms the AI Scientist's review agent, which exhibits strong decision biases: GPT-4.1-nano accepts 17.1% of papers, DeepSeek-R1 accepts 85.4% of papers, and QwQ-32B accepts 69.2%, all far away from the true 31.7% rate. These results underscore 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. These results suggest a promising path forward for building scalable, domain-aware advising systems that can meaningfully augment human creativity and decision-making in research.

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*.

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.

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. 2024. Agentreview: Exploring peer review dynamics with llm agents. *arXiv preprint arXiv:2406.12708*.

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

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.

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*.

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*.

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. 2024. Cycleresearcher: Improving automated research via automated review. *arXiv preprint arXiv:2411.00816*.

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*.

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*.

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

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 7: Prompt details and JSON output format

B 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 5 shows the annual publication counts.

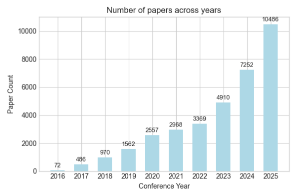

Figure 5

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 8 reports the average token counts.

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

Table 8: Average token lengths per section (GPT-4 tokenizer).

C Training Details

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 6 shows how the average reward and best reward evolved over these iterations.

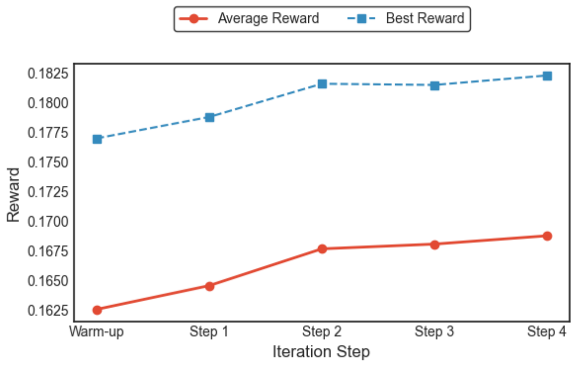

Figure 6

D Case Study

Setup	Evaluation
Our Model Rubrics Guided	**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 9: Evaluation & Suggestion generated by GUIDE-7B

Setup	Evaluation
DeepSeek-R1 Rubrics Guided	**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 10: Evaluation & Suggestion generated by DeepSeek-R1 with full rubrics-guided prompt.

Setup	Evaluation
DeepSeek-R1 No Rubrics	**Summary**: Overall, the paper introduces the concept of cross capabilities in LLMs, defines seven individual and paired cross capabilities, constructs the CrossEval benchmark with human-annotated prompts and expert evaluations, and identifies the 'Law of the Weakest Link' effect where cross-capability performance is constrained by the weakest individual capability. **Comparison with previous works**: Compared to previous works like SKILL-MIX and CMMLU, which focus on skill combinations or language-specific evaluations, this paper uniquely systematizes cross capabilities via hierarchical taxonomies and human-annotated benchmarks. Unlike L-Eval or SCALE, which emphasize long-context or domain-specific tasks, it explores interdependencies between core capabilities. While Corex studies multi-model collaboration for reasoning, this work focuses on intrinsic capability limitations within single models. **Strengths**: The strengths of the paper are its systematic taxonomy of cross capabilities, rigorous human annotation process (8,400 expert ratings), and empirical validation of the 'Law of the Weakest Link' across 17 models. The integration of LLM-based evaluators aligned with human judgments adds methodological robustness. The focus on capability interdependencies addresses a critical gap in LLM evaluation literature. **Weaknesses**: The weaknesses of the paper are the limited scope of seven predefined capabilities, which may not cover all real-world cross-domain tasks. The manual annotation process, while thorough, lacks scalability compared to automated methods like those in SKILL-MIX. The study does not explore how model scaling (as in GPT-Fathom) or training strategies (as in How Abilities in Large Language Models...) might mitigate the weakest-link effect. **Evaluation**: In conclusion, the paper makes a valuable contribution by formalizing cross capabilities and demonstrating systemic performance bottlenecks in LLMs. The findings align with trends observed in Corex and Critique Ability of Large Language Models but offer novel insights into capability interdependencies. **Suggestion**: To improve the paper, the authors could expand capability taxonomies using automated methods from SKILL-MIX, analyze scaling effects on cross capabilities as in GPT-Fathom, and explore mitigation strategies like multi-model collaboration (Corex) or in-context learning (Beyond task performance...). Including non-English cross capabilities beyond Spanish would enhance generalizability.

Table 11: Evaluation & Suggestion generated by DeepSeek-R1 without rubrics-guided prompt.

E 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.

F Human Annotation for Contribution Extraction

The 100 annotated contributions in Sec. 3.2 were manually generated by a PhD student, who is one of the authors of this paper.

G AI Usage

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