GIANTS: Generative Insight Anticipation from Scientific Literature

Joy He-Yueya[*1], **Anikait Singh**[*1], **Ge Gao**[1], **Michael Y. Li**[1], **Sherry Yang**[2], **Chelsea Finn**[1], **Emma Brunskill**[1], **Noah D. Goodman**[1]
[1]Stanford University, [2]New York University, [*]Equal contribution

Scientific breakthroughs often emerge from synthesizing prior ideas into novel contributions. While language models (LMs) show promise in scientific discovery, their ability to perform this targeted, literature-grounded synthesis remains underexplored. We introduce *insight anticipation*, a generation task in which a model predicts a downstream paper's core insight from its foundational parent papers. To evaluate this capability, we develop GIANTSBENCH, a benchmark of $17k$ examples across eight scientific domains, where each example consists of a set of parent papers paired with the core insight of a downstream paper. We evaluate models using an LM judge that scores similarity between generated and ground-truth insights, and show that these similarity scores correlate with expert human ratings. Finally, we present `GIANTS-4B`, an LM trained via reinforcement learning (RL) to optimize insight anticipation using these similarity scores as a proxy reward. Despite its smaller open-source architecture, `GIANTS-4B` outperforms proprietary baselines and generalizes to unseen domains, achieving a 34% relative improvement in similarity score over `gemini-3-pro`. Human evaluations further show that `GIANTS-4B` produces insights that are more conceptually clear than those of the base model. In addition, SciJudge-30B, a third-party model trained to compare research abstracts by likely citation impact, predicts that insights generated by `GIANTS-4B` are more likely to lead to higher citations, preferring them over the base model in 68% of pairwise comparisons. We release our code, benchmark, and model to support future research in automated scientific discovery.

> "If I have seen further [than others], it is by standing on the shoulders of giants."
> — Isaac Newton

1. Introduction

Language Models (LMs) are becoming useful tools for scientific discovery [7]. Recent work has shown promising results, from virtual LM teams designing SARS-CoV-2 nanobody binders [23] to models proposing NLP research directions that human experts judge to be novel [16]. However, many of these successes rely heavily on prompting frontier models pre-trained on massive text corpora. In contrast, human researchers often achieve breakthroughs through a far more data-efficient process: synthesizing profound insights from a small set of prior works. Existing LMs still struggle to reliably generate hypotheses or *insights* of true impact and value [17], often due to the lack of diversity and feasibility [25].

To bridge this gap, we propose the task of *insight anticipation*: given a small set of prior papers, can a model reconstruct the core insight of a downstream paper that builds on them? Unlike open-ended research ideation, this setting evaluates targeted synthesis grounded in specific scientific literature. This framing is motivated by the classical view of scientific progress as '*standing on the shoulders of giants,*' where new contributions emerge by building upon prior foundations. Under this view, the challenge of automated discovery decomposes into two subproblems: (1) *parent selection*, which involves identifying

Corresponding author(s): heyueya@cs.stanford.edu, anikait@stanford.edu.
Project Website: https://giants-insights.github.io/

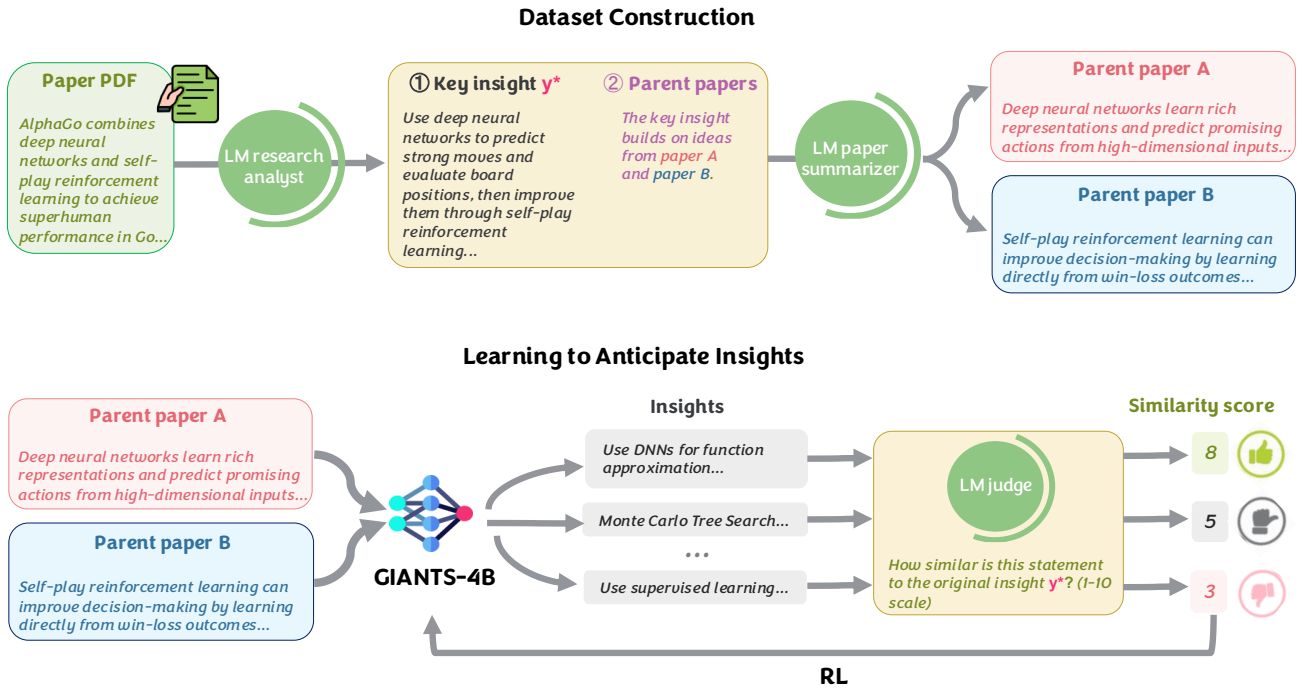

Figure 1: Overview of GIANTSBENCH and GIANTS-4B. *(top)* For dataset construction, we use an LM as a research analyst to process each paper PDF to identify two parent papers whose ideas are synergistically combined to produce the paper's key insight, which is extracted as the ground-truth target y^*. A second LM summarizes each parent paper. *(bottom)* Given the two parent summaries, the model generates candidate insights, which an LM judge scores by similarity to y^* on a 1–10 scale. These scores serve as a proxy reward for RL training, teaching the model to anticipate insights that more closely match real downstream papers.

the relevant prior works to build upon, and (2) *insight generation*, which involves synthesizing those works into a novel hypothesis.

In this work, we focus exclusively on the second problem. We assume that the parent papers are provided by an oracle literature selection criterion and ask a more targeted feasibility question: if the relevant lineage of prior work is known, can a model effectively predict the next conceptual leap? This controlled setup isolates the synthesis of research insights from the retrieval of prior work and allows us to directly test whether meaningful literature-grounded insight generation is possible at all. Framed this way, *insight anticipation* is not only a prediction task but also a potential training signal for scientific reasoning: learning to anticipate the downstream insights implied by a lineage of parent works may help models internalize patterns of progression, combination, and abstraction that underlie strong human insight synthesis.

To evaluate insight anticipation models, we develop GIANTSBENCH, a large-scale benchmark for testing the ability of models to synthesize two prior papers and derive the core insight of a downstream paper that builds on them (Figure 1, top). In contrast to a recent benchmark that evaluates whether models can answer literature synthesis questions by identifying relevant papers and generating long-form responses with citations [2], GIANTSBENCH assesses whether a model can combine two parent papers synergistically to derive insights that lead to a subsequent paper. GIANTSBENCH contains $17k$ examples drawn from Computer Science, Economics, Electrical Engineering, Mathematics, Physics, Quantitative Biology, Quantitative Finance, and Statistics. Each example consists of two parent-paper summaries

paired with the core insight of a downstream paper, where the insight is a concise description of the paper's primary contribution automatically constructed by an LM from the downstream paper PDF. We also introduce an automatic evaluation method that uses an LM judge to give a **similarity score** comparing the model-generated insight to the ground-truth insight. Expert evaluation shows that these LM-judge scores are positively correlated with human ratings (Spearman's $\rho = 0.761$, $p < 0.001$).

We then introduce GIANTS-4B, a 4B-parameter language model trained via reinforcement learning (RL) to generate the core insight of a downstream paper from its parent papers. By fine-tuning with GRPO [15] to maximize the similarity scores on the research insights, the model learns to reason about connections between the input papers and generate insights that are more similar to those from real downstream papers (Figure 1, bottom). This approach outperforms simply distilling the expert insight (even with reasoning traces).

We evaluate both proprietary and open LMs, including `gemini-2.5-pro`, `gemini-3-pro`, `Qwen3-4B`, on GIANTSBENCH. `Qwen3-4B` performs similarly to `gemini-2.5-pro` and `gemini-3-pro`, despite being a smaller open model. Moreover, the similar performance of `gemini-2.5-pro` and `gemini-3-pro` suggests that frontier LMs are not simply getting better at insight anticipation via scaling. In contrast, our GIANTS-4B outperforms these baselines and generalizes zero-shot to unseen domains, achieving a 34% improvement in similarity score over `gemini-3-pro`. GIANTS-4B also produces insights that human evaluators judge to be more conceptually clear than those of the base model. In addition, SciJudge-30B [24], a third-party judge trained to compare research abstracts by likely citation impact, prefers GIANTS-4B over the base model in 68% of pairwise comparisons.

We summarize our main contributions as follows:

- **Insight anticipation.** We introduce a new literature-grounded generation task that isolates the synthesis phase of scientific discovery by asking models to predict a downstream paper's core insight from its parent papers.
- **GIANTSBENCH and evaluation metric.** We construct a benchmark of 17k tuples of parent and downstream papers from arXiv across eight domains, together with an LM-based auto-evaluator for measuring the similarity between generated and ground-truth insights.
- **GIANTS-4B.** We train a model for insight anticipation via RL using similarity-based rewards and show that it outperforms a range of proprietary and open LMs such as `gemini-3-pro`, improves conceptual clarity of insights, and generalizes zero-shot to unseen domains.

2. Defining and Instantiating a Benchmark for Insight Anticipation

Task Definition. We define an *insight* as a concise, natural language description of a paper's primary methodological or empirical advance. Building on this, we introduce the task of *insight anticipation*: given an input context comprising the content of two parent papers, $x = (x_A, x_B)$, the goal is to generate the key insight of a downstream paper that builds on both parent papers (A and B). We denote this downstream insight as the ground-truth insight, y^*, which emerges from the synthesis of the two foundational papers. While y^* is just one of many possible subsequent ideas, it serves as a measurable proxy for the next conceptual leap. This setup allows us to probe a model's ability to reason about the joint influence of A and B, challenging it to generate a research insight, $\hat{y}$, that is semantically similar to y^*. Figure 10 shows the insight generation prompt used for training and evaluation of all models. To manage context constraints during training and inference, we intentionally restrict the context to two parent papers, establishing a

conservative lower bound on the broader problem. Furthermore, we bypass parent discovery to focus entirely on whether meaningful insight synthesis is possible given fixed inputs.

Dataset Construction. We create a dataset $\{((x_A, x_B)_i, y_i^*)\}_{i=1}^N$ as follows. We collect 17,839 papers from arXiv that are published between May 23rd, 2007, and January 23rd, 2026, according to their latest update date on arXiv. Since arXiv is not peer reviewed, the raw corpus may contain noisy or non-substantive documents. To mitigate this, we retain only papers with at least two citations.[1] We download these papers as a PDF. For each paper, we prompt `gemini-2.5-flash` to identify two prior papers that this paper explicitly cites and builds upon by combining their ideas in a synergistic way. We also ask `gemini-2.5-flash` to explain the synergy (see the full prompt in Figure 11). We then download the two parent papers as a PDF and use their content as input context x. Ideally, we would like to take the full paper content as input to the model and ask it to generate a downstream insight. However, given the context-length limitations of existing LMs as well as the high inference cost, we instead opt to prompt `gemini-2.5-flash` to summarize the paper. We then use paper summaries as the input. We use the following prompt to summarize each paper into key insights and contributions: "*Summarize the document, clearly describing the method used and highlighting the key insights or findings. Provide sufficient detail so that the approach and main contributions are fully understood.*" For y^*, we use the synergy explanation that is part of the output to the parent-identification prompt in Figure 11. We cannot use the raw explanation directly as the target because it refers to a future downstream paper and talks about the insight in the context of the two parent papers. Instead, we would like the model to generate a standalone insight statement directly from the two parent papers alone. Therefore, we prompt `gemini-3-pro` to rewrite the insight and explanation without reference to the downstream paper (see prompt in Appendix C.4). After constructing all (x, y^*) pairs, for each unique pair of parent papers $x = (x_A, x_B)$, we keep the insight y^* from the most cited downstream paper in order to prioritize generating more impactful insights.

Temporal Hold-Out Evaluation. To assess generalization, we conduct all evaluations on a *future held-out test set* consisting of downstream papers published after the training cutoff date. We split the dataset by the publication date of the downstream paper in each (x, y^*) pair. Papers published before July 1st, 2023 are used for training. To study the ability of models to generalize to new domains, we further restrict the training set to papers tagged with cs.CL (Computation and Language), yielding $N = 10{,}335$ training examples. For evaluation, we consider papers published after July 1, 2023 and randomly sample up to 600 papers from each domain (see the category taxonomy in Appendix A), resulting in a test set of 7,504 papers. Although we deduplicate the dataset by parent pair, some test examples may still share at most one parent paper with a training example. To address this, we also report results on a stricter

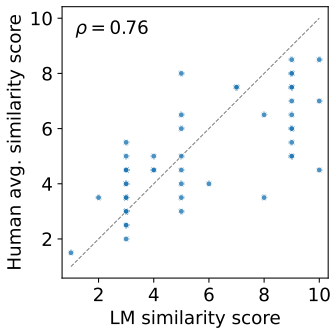

Figure 2: LM judgements of insight similarity correlate with human judgements. We ask both human annotators and the LM judge (`gemini-3-pro`) to score the similarity between a model-generated insight and a ground-truth downstream insight on a 1–10 scale. The LM's scores are positively correlated with average human ratings (Spearman $\rho = 0.761$, $p < 0.001$, $n = 60$).

[1] We use citation counts from Semantic Scholar as a proxy for paper quality.

subset of the test set, **Test-unseen-parents** ($N = 5{,}294$), which excludes any test example that shares a parent with the training set.

Evaluation Metric. For each generation, we prompt an LM judge to give a **similarity score** ranging from 1 to 10 (higher is better) comparing the model-generated insight $\hat{y}$ to the ground-truth insight y^* (see Figure 12 for the prompt). We use `gemini-3-pro` as the primary judge model. To assess the reliability of the LM-based evaluation, we conduct a human evaluation study on 30 pairs of insights generated by `Qwen3-4B` and `GIANTS-4B` ($n = 60$). Two human annotators, who are PhD students in Computer Science, independently rate the similarity between model-generated insights and ground-truth insights using the same rating scale as the LM judge. The LM's scores show a statistically significant positive correlation with the average scores across human annotators (Figure 2). In particular, we observe a Spearman rank correlation of $\rho = 0.761$ ($p < 0.001$). More details can be found in Appendix F.1.

Conceptual View. Conceptually, this framework can be viewed as an *auto-encoding task* [8] over the citation graph. A target paper is passed through a highly lossy channel, the summaries of its two parent papers, from which the model must successfully reconstruct the original paper's core insight. By linearizing the citation graph into input-target pairs, we challenge the model to recreate the conceptual leap required to bridge adjacent nodes.

Takeaways of Insight Anticipation Instantiation

We introduce the *insight anticipation* task, which challenges models to generate a future research insight by synthesizing the summaries of two parent papers. To evaluate this, we construct a dataset of over $17k$ examples from arXiv with LM-extracted ground truths, using a temporal and cross-domain hold-out split and a human-validated LM judge for scoring.

3. Training an Insight Anticipation Model

We explore two training paradigms for insight anticipation: (1) distillation via supervised fine-tuning (SFT), leveraging ground-truth insights and rationalization from target downstream papers, and (2) reinforcement learning (RL) via similarity optimization. Across all experiments, we use `Qwen3-4B` as the base model. Next, we will detail the specific formulations of these training methodologies.

3.1. Ground Truth Insight Distillation via Supervised Fine-Tuning (SFT)

Our first approach is to fine-tune the base model to generate a downstream paper's core insight from the summaries of its two parent papers. We investigate two distinct SFT strategies to achieve this. In our standard SFT approach, the model is directly fine-tuned to map the input context (summaries of parent papers), x, to the target ground-truth insight, y^* (from the downstream paper). This process optimizes the standard cross-entropy loss for autoregressive LMs over the paired examples (x, y^*).

To bridge the logical gap between the source papers and the final insight, we also evaluate an SFT strategy enhanced with chain-of-thought reasoning [28], which we denote as *SFT-think*. In this setup, we introduce an intermediate synthetic rationalization step, z. To generate a supervision target, we prompt a high-capacity teacher model (`gemini-3-pro`) to generate a detailed chain-of-thought that

logically deduces the ground-truth insight conditioned on the parent paper summaries (as detailed in Section 2). The base model is then trained on augmented tuples (x, z, y^*), learning to sequentially predict the rationale followed by the final insight. This approach explicitly encourages the model to internalize the step-by-step inferential process rather than merely memorizing the final output mapping. This follows distillation approaches in reasoning as seen in works such as OpenThoughts [6] and s1 [11].

3.2. Reinforcement Learning for Insight Anticipation via Similarity Optimization

So far, we have explored distilling insights via supervised fine-tuning (SFT). However, in complex tasks such as scientific discovery, the downstream paper's insights can be difficult to clone directly. This difficulty often arises when the downstream insight has low likelihood under the policy, or when the model lacks the capacity to adequately capture the distribution of the parent insight. Therefore, we explore an alternative approach: using semantic similarity to the downstream paper's insights y^* as a proxy reward. Formally, given a ground-truth insight from the downstream paper y^*, we define the proxy reward for a predicted insight $\hat{y}$ as:

$$r_{\text{sim}}(\hat{y}) = \textit{similarity}(\hat{y}, y^*) \tag{1}$$

where *similarity* measures the semantic equivalence between the generated insight and downstream paper's insight measured using an LM-as-a-Judge [5], matching the evaluation criterion in Section 2. This proxy reward is then optimized via RL to recover the behavior of the ground-truth insight of the downstream paper conditioned on the two prespecified parent papers.

We optimize this reward using Group Relative Policy Optimization (GRPO) [15]. In particular, for each input context x, we sample a group of $G = 8$ candidate insights from the current policy. An LM judge evaluates these candidates, and GRPO updates the policy relative to the sampled group (illustrated in Figure 1, bottom). GRPO is well-suited for this setup because it avoids the need to train and maintain a separate, memory-intensive value function model (see implementation details in Appendix D).

Crucially, to mitigate the risk of reward hacking and ensure a rigorous evaluation, we enforce a strict separation between the training and testing judges. We use `gemini-2.5-flash` as the active reward model during GRPO training, while reserving the independent `gemini-3-pro` model exclusively for the final evaluation phase. This decoupling guarantees a more objective, unbiased assessment of the model's true generalization capabilities. We additionally evaluate with other LM judges (e.g., `Qwen3-14B`) to showcase robustness across model families.

Takeaways of Training an Insight Anticipation Model

We optimize our insight anticipation model using reinforcement learning to maximize semantic similarity to downstream ground-truth insights. We strictly separate the LM judges used for training and final evaluation to mitigate the risk of reward hacking.

4. Experimental Evaluation on GiantsBench

Our experimental evaluation studies whether `GIANTS-4B` improves insight anticipation on GiantsBench. We compare `GIANTS-4B` against proprietary and open-weight language models, as well as supervised fine-tuning baselines built from the same base model. Unless otherwise noted, we evaluate all

Qwen-based models (Base, SFT, SFT-think, and GIANTS-4B) using the recommended Qwen thinking-mode decoding settings: temperature $= 0.6$, top-p $= 0.95$, top-k $= 20$, and min-p $= 0$. For the Gemini baselines, we use temperature $= 0$. We report results on both the full test set and the stricter **Test-unseen-parents** subset, which excludes any test example that shares a parent paper with the training set. Unless otherwise noted, our primary evaluation uses gemini-3-pro as the LM judge, and we additionally evaluate with other LM judges (e.g., Qwen3-14B, gemini-2.5-pro) to test whether model rankings are robust across judges.

Through these experiments, we address three questions: (1) how well current frontier and open-source language models perform on insight anticipation, (2) whether directly optimizing for insight similarity using RL (GIANTS-4B) improves insight anticipation, and (3) whether a model trained on one domain transfers to unseen domains.

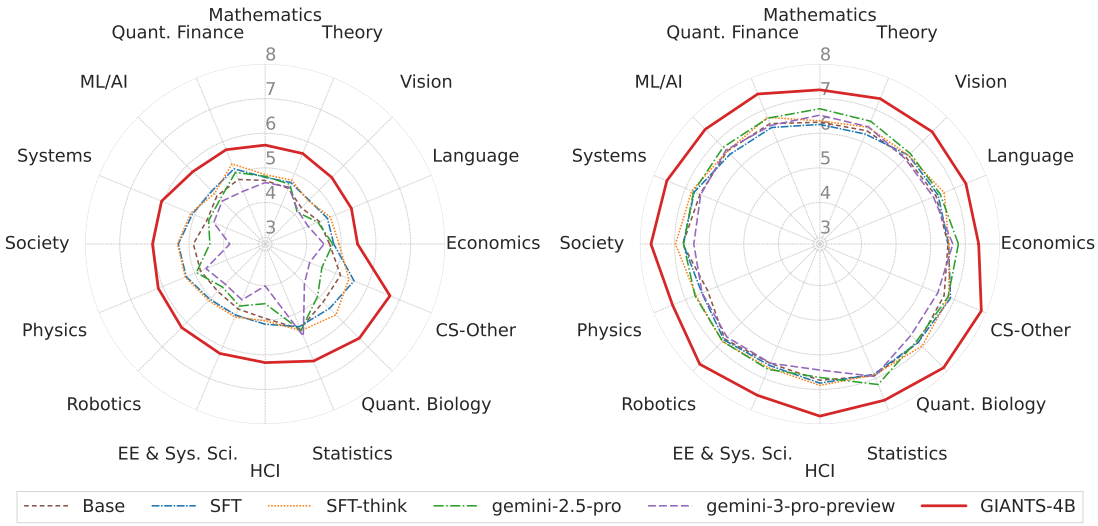

Figure 3: Similarity scores on GiantsBench (higher is better). GIANTS-4B consistently achieves the highest similarity scores across domains. *(left)* scores from our primary evaluation judge gemini-3-pro. *(right)* scores from Qwen3-14B.

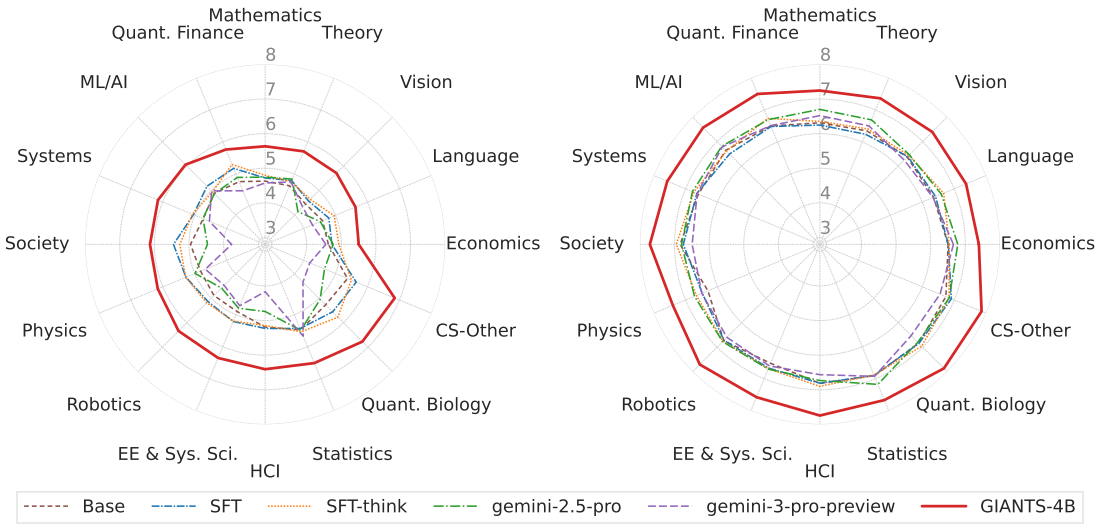

Figure 4: Similarity scores on a subset of GiantsBench containing only unseen parent papers. GIANTS-4B generalizes to new domains and distribution of parent papers, as judged by gemini-3-pro *(left)* and Qwen3-14B *(right)*.

1) Insight anticipation is a challenging task even for large proprietary models. Despite their success across diverse NLP benchmarks, current frontier models exhibit significant limitations in literature-grounded synthesis. As shown in Figure 3, the base open model, `Qwen3-4B`, achieves an average similarity score of only 4.75 (see raw statistics corresponding to Figure 3 in Appendix B). Based on our LM-based evaluation rubric (Figure 12), a score in this range indicates that while generated insights may align topically with the ground-truth insights, they fail to capture the core scientific contribution or technical nuance. Notably, significantly larger proprietary models, including `gemini-2.5-pro` and `gemini-3-pro`, perform similarly to the smaller `Qwen3-4B` baseline. This finding suggests that literature-grounded synthesis capabilities do not scale linearly with model size alone, reinforcing the necessity for specialized training paradigms like `GIANTS-4B`.

2) Directly optimizing similarity scores substantially improves insight anticipation. `GIANTS-4B` achieves the highest performance among all evaluated methods. Figure 3 shows that, on the full future held-out test set, `GIANTS-4B` consistently outperforms the base model, both supervised baselines, and the proprietary baselines under two different judge LMs. Relative to `gemini-3-pro`, `GIANTS-4B` achieves a 35% improvement in similarity score on the full test set. Figure 4 shows that this advantage persists on the stricter **Test-unseen-parents** split, where GIANTS-4B achieves a 34% improvement over `gemini-3-pro` (see raw statistics corresponding to Figure 3 and Figure 4 in Appendix B). In contrast, standard SFT and SFT-think both improve slightly over the base model. These results suggest that standard imitation learning provides a modest benefit for this synthesis task; however, RL with a similarity-based reward effectively aligns the model's capabilities with target human insights. This performance trend remains robust as we increase the number of samples to optimize the similarity score via inference-time scaling, as illustrated in Figure 5. Since best-of-k evaluation requires scoring many samples per example, we conduct the inference-time scaling evaluation in Figure 5 on a subset of 480 examples sampled from GiantsBench, using `gemini-3-flash` as the LM judge for cost considerations. We use temperature $= 0.6$ for all models in this evaluation.

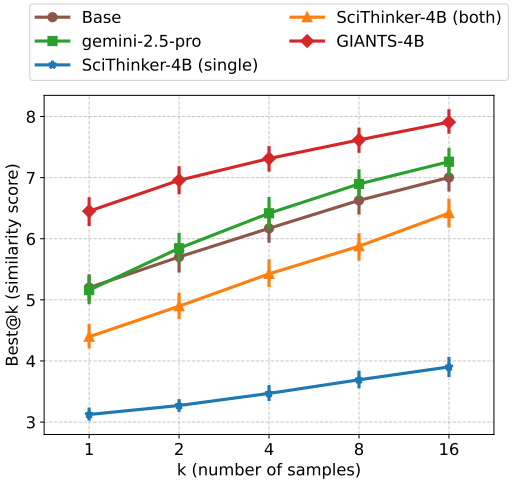

Figure 5: `GIANTS-4B` remains strongest under test-time scaling. As the number of samples per example increases, `GIANTS-4B` consistently outperforms the base model, `gemini-2.5-pro`, and SciThinker-4B [24], which is a scientific-ideation model trained using a citation-preference reward model. Error bars show 95% confidence intervals.

In Figure 5, we additionally compare against SciThinker-4B [24], a related scientific-ideation model rather than a direct same-task baseline. In the original setting of Tong et al. [24], SciThinker takes as input a single seed paper's title and abstract and outputs a follow-up research idea. Its training objective is to generate ideas with high potential impact under a citation-preference reward model. By contrast, our task requires synthesizing two parent papers to generate the core insight of a downstream paper. To test whether this kind of general scientific ideation training transfers to insight anticipation, we evaluate two variants: SciThinker-4B (single), which is given one parent paper, and SciThinker-4B (both), which is given both parent papers. Figure 5 shows that using both parents improves SciThinker-4B, but it still remains well below `GIANTS-4B` across all k. This suggests that general scientific ideation training may not transfer well to insight anticipation, and that strong performance on our task requires training aligned with literature-grounded multi-paper synthesis rather than open-ended follow-up ideation.

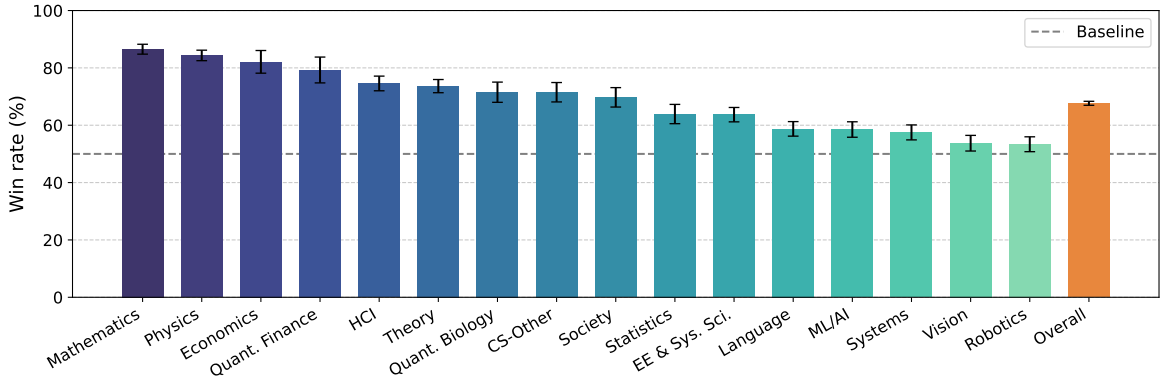

Figure 6: GIANTS-4B achieves a 68% overall win rate against the base model under SciJudge-30B. We evaluate pairwise preferences using SciJudge-30B [24], a third-party judge trained to compare research abstracts by likely citation impact. This provides complementary evidence that optimizing for insight anticipation also improves performance under an independent, impact-oriented evaluation signal. Error bars show standard errors.

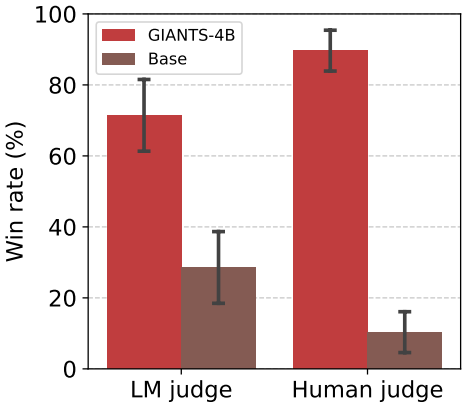

Figure 7: Win rates of GIANTS-4B against the base model for insight similarity. We ask both human annotators and an LM judge to score the similarity between a model-generated insight and a ground-truth downstream insight and find that GIANTS-4B better matches the ground-truth insight. Error bars show standard errors.

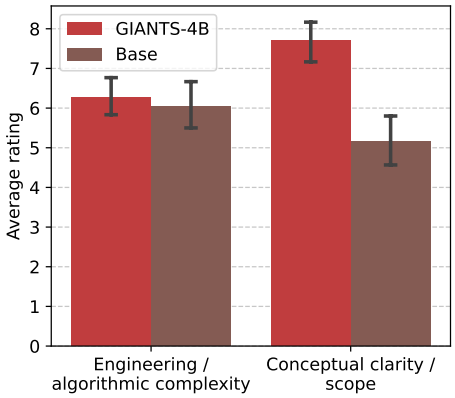

Figure 8: GIANTS-4B produces insights with similar perceived algorithmic complexity but substantially higher conceptual clarity compared to the base model. Human annotators assess the feasibility of generated insights along two axes: (1) engineering / algorithmic complexity and (2) conceptual clarity. Error bars show 95% confidence intervals.

To further validate these gains, we also evaluate generations using SciJudge-30B, a third-party model from Tong et al. [24] trained to compare research abstracts by likely citation impact. Under this external preference signal, GIANTS-4B achieves a 68% overall win rate against the base model (Figure 6), with variation across domains. While this metric is only a proxy for research impact, it provides complementary evidence that optimizing for similarity produces outputs that are also preferred by an independent quality-oriented judge. To reduce order effects, we evaluate each pair twice with reversed presentation order and filter inconsistent judgments.

We supplemented these automated evaluations with two independent preliminary human studies. First, we ask two human annotators to rate the similarity between model-generated insights and their corresponding ground-truth insights for 30 head-to-head pairs from the base model and GIANTS-4B, using the same rating scale as the LM judge (Figure 12). We then compare the win rate of GIANTS-4B against the base model in terms of alignment with the ground-truth insight (Figure 7). In this comparison,

GIANTS-4B achieves a 71.4% win rate under the LM judge (gemini-3-pro) and an 89.7% win rate under human evaluation. Second, we assess the feasibility of generated insights along two axes: algorithmic complexity and conceptual clarity. Figure 8 shows that while GIANTS-4B produces insights of similar algorithmic complexity to the base model, it significantly improves conceptual clarity, making the generated ideas more interpretable and actionable. Further details regarding the human study methodology are provided in Appendix F.

3) GIANTS-4B zero-shot generalizes to new domains and distribution of parent papers.

Crucially, the synthesis capabilities acquired by GIANTS-4B are not restricted to its training data. Although trained exclusively on papers from the Language domain (CS.CL), Figure 3 shows consistent gains across all evaluated domains. Moreover, these evaluations are conducted on papers published after the training cutoff, so the gains reflect performance on temporally held-out downstream literature rather than memorization of the training set. This suggests that the model learns a generalizable mechanism for combining disparate ideas rather than memorizing domain-specific heuristics. In addition, Figure 4 shows model performance on a subset of test examples whose parent papers are entirely unseen during training. GIANTS-4B remains the top-performing method on this subset, suggesting that its gains are not driven by partial overlap in parent-paper lineage.

Evaluation reliability check. To ensure our findings are robust and that similarity scores are not biased toward the specific LM judge that was used during training, we conducted a cross-model validation using an independent LM judge, Qwen3-14B (Figure 3, right). This secondary evaluation corroborated our primary findings, ranking GIANTS-4B as the top-performing model across all metrics, albeit with a slightly more compressed performance margin. We also report results under three additional judge models in Appendix E, with consistent findings across model sizes.

Qualitative Comparison of Models. We present qualitative examples to demonstrate that our generated insights meaningfully integrate concepts from prior literature. We use NeurIPS 2025 award-winning papers as representative high-quality parent papers and compare the insights produced by GIANTS-4B against those generated by the base model given these parent papers.

The first example (Figure 9, left) shows insights derived from Wang et al. [26] and Liu et al. [10]. While the base model merely summarizes the parent papers without synthesizing a novel perspective, GIANTS-4B proposes a concrete mechanistic connection between the two studies. The second example (Figure 9, right) shows insights based on Yue et al. [30] and Qiu et al. [13]. In this example, the base model generates an overly ambitious claim, suggesting that gating mechanisms expand reasoning boundaries in ways that RLVR cannot, which is unsupported by the source texts. In contrast, GIANTS-4B remains grounded in the actual findings while still identifying a non-trivial connection: that attention gating may dictate how effectively reinforcement learning concentrates probability mass across useful reasoning trajectories. This comparison underscores a critical distinction between boldness and genuine insightfulness. While the base model's novelty relies on unsupported extrapolation, GIANTS-4B maintains a narrower scope to identify a highly plausible interaction between the foundational works.

These qualitative examples are consistent with the quantitative results and human evaluations. GIANTS-4B more often produces insights that connect the parent papers in a concrete, interpretable, and grounded way, rather than simply restating their contents or making broader but weakly supported claims.

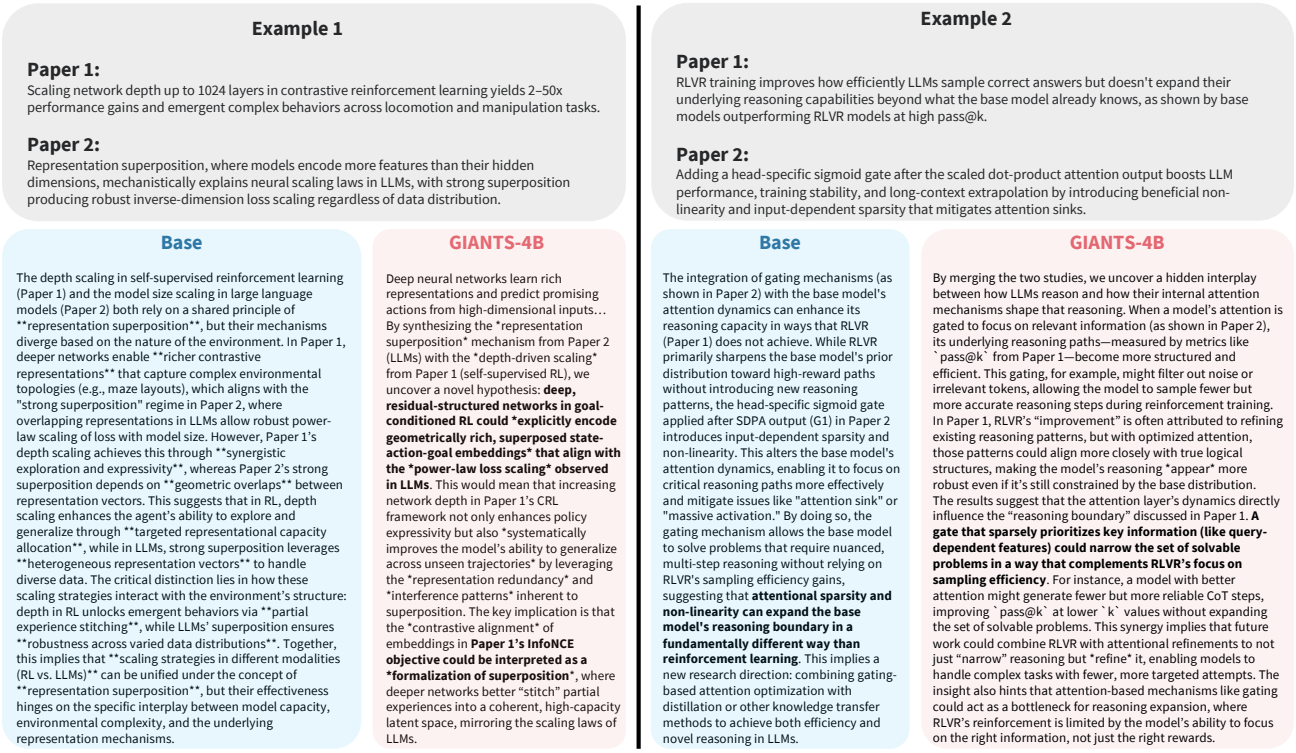

Figure 9: Qualitative comparison of insights derived from NeurIPS 2025 award-winning papers. *(left)* GIANTS-4B identifies a more concrete cross-paper mechanism than the base model (Qwen3-4B). *(right)* GIANTS-4B produces a grounded, more plausible interaction, while the base model makes a broader conjecture that is less directly grounded in the parent papers. We show a one-sentence summary of each parent paper for readability. These are illustrative abbreviations only and are not the full parent-paper summaries used as model input.

Takeaways of Experimental Evaluation

GIANTS-4B significantly outperforms both frontier models and SFT models on our insight anticipation task. GIANTS-4B produces insights that are rated by human experts as more conceptually clear than those of the base model. A third-party citation-preference judge also prefers GIANTS-4B's outputs over the base model. Furthermore, despite training exclusively on a single domain, the model successfully zero-shot generalizes its synthesis capabilities across diverse, unseen scientific disciplines and temporally held-out literature.

5. Related Work

AI for Research. Many works have studied how to use LMs for different components of the scientific research pipeline, including literature search [12, 32], idea generation [9, 16, 24, 27, 31], idea execution [7, 17, 18], and paper review [29]. Asai et al. [2] evaluate the ability of models to answer literature synthesis questions by identifying relevant papers and generating long-form responses with citations. Many existing methods for idea generation either directly prompt LMs to perform open-ended brainstorming or expose LMs to a set of prior works without requiring cross-paper synthesis [9, 16]. SciMON [27] studies literature-grounded scientific idea generation, but its setting is different from ours: SciMON takes background problem contexts as input, retrieves literature inspirations, and generates

open-ended ideas optimized for novelty relative to prior work. By contrast, we focus on synthesizing insights from prior work and do not optimize for open-ended novelty. A concurrent work trains models to judge research impact from citation signals, then uses the learned judge as a reward model for idea generation [24]. In contrast, we focus on the problem of synthesizing insights from two parent papers, using similarity to ground-truth downstream insights as the training signal. A related line of work studies scientific progress as a forecasting problem. PRESCIENCE decomposes the research process into collaborator prediction, prior work selection, contribution generation, and impact prediction [1]. In its contribution generation task, models generate a future paper's title and abstract conditioned on prior work and other historical context, and are evaluated using an LM-based similarity metric. Compared to this broader scientific forecasting setup, we isolate the idea-synthesis problem and test whether models can generate the core insight of a downstream paper from a given pair of parent papers.

Literature-based Discovery. Literature-Based Discovery (LBD) studies how computational methods can uncover hidden links between seemingly unrelated bodies of research to infer novel and potentially useful knowledge [20, 22]. A classic example is Swanson's hypothesis that fish oil could treat Raynaud's syndrome: one set of papers suggested that fish oil reduces blood viscosity, while another linked high blood viscosity to Raynaud's syndrome [21]. However, existing LBD approaches often struggle to scale to the volume of modern scientific literature, rely on domain-specific knowledge sources, or output ranked candidate connections rather than clear, standalone hypotheses that researchers can directly act on [3, 4, 14, 19]. Our work is complementary but more focused: rather than identifying latent cross-literature links, we study whether a model can synthesize two parent papers into the core insight of a real downstream paper.

6. Discussion, Limitations, Future Work

In this work, we introduced *insight anticipation* as a measurable paradigm for automated scientific discovery. By isolating the insight generation phase, we demonstrated that models can effectively synthesize the core insight of a downstream paper when provided with its foundational parent papers. Our results with GIANTS-4B indicate that the trajectory of scientific intuition is partially predictable, and that optimizing language models via reinforcement learning with similarity-based rewards is a highly effective training strategy for this task. However, this foundational formulation leaves several important limitations and open questions.

One limitation of the work is that we assume that downstream contributions derive from two parent papers due to context constraints, despite research ideas being heavily shaped by broader intellectual contexts. Furthermore, parent identification is imperfect. Citations do not always reflect true conceptual influence, and influential ideas may remain uncited. Finally, by assuming an oracle literature selection criterion, we explicitly decoupled insight generation from parent selection, which may not be feasible for some scientific tasks.

Future research can address the parent selection problem directly or integrate automated retrieval systems to unify parent selection and synthesis within a single end-to-end framework. Additionally, extending the task to accommodate multi-source lineage and developing evaluation metrics that prioritize conceptual novelty over textual similarity could be interesting. Ultimately, testing these models in active, human-in-the-loop research settings will determine their true potential as catalysts for scientific discovery.

7. Ethics Statement

This research involves training the GIANTS-4B model on a dataset of 17,839 publicly available arXiv preprints to synthesize scientific insights. While automating scientific ideation offers significant potential for discovery, it raises important ethical considerations regarding the proper attribution of foundational ideas, especially since academic citation graphs can be noisy and may not always reflect true conceptual influence. Additionally, there is an inherent risk of language models generating plausible but unverified scientific claims; our framework mitigates this by using reinforcement learning to directly align model outputs with grounded, human-derived insights. Finally, we acknowledge the environmental and computational costs associated with training models on high-performance GPUs, though utilizing a more efficient 4B parameter architecture helps limit this footprint compared to massive proprietary models.

8. Reproducibility Statement

To ensure the reproducibility of our results, we provide a comprehensive account of our methodology, code, and data. The source code for our models and experiments is available at the following repository: https://github.com/joyheyueya/giants. Our benchmark and model weights can be found in the following HuggingFace repository: https://huggingface.co/giants2026. Our implementation is built upon the verl framework (https://verl.readthedocs.io/en/latest/). All experimental details, including hyperparameter settings, are documented in Appendix D. The computational experiments were conducted on a machine with NVIDIA A100 GPUs, and the required software dependencies are listed in the requirements.txt file within our code repository.

9. Acknowledgments

We thank Shirley Wu, Chenglei Si, Ryan Louie, Rui Li, Yoonho Lee, Jack Bai, Aviral Kumar, Andreas Stuhlmüller and others in the Stanford CoCoLab, Brunskill lab, and Iris Lab for discussions and feedback. This work was supported by the Junglee Corporation Stanford Graduate Fellowship. AS gratefully acknowledges the support of the NSF Graduate Research Fellowship Program, Modal Academic Research Program, and the Toyota Research Institute. CF was supported by Schmidt Sciences.

References

[1] Anirudh Ajith, Amanpreet Singh, Jay DeYoung, Nadav Kunievsky, Austin C Kozlowski, Oyvind Tafjord, James Evans, Daniel S Weld, Tom Hope, and Doug Downey. Prescience: A benchmark for forecasting scientific contributions. *arXiv preprint arXiv:2602.20459*, 2026.

[2] Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'Arcy, et al. Synthesizing scientific literature with retrieval-augmented language models. *Nature*, pages 1–7, 2026.

[3] Tanja Bekhuis. Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomedical digital libraries*, 3(1):2, 2006.

[4] Murat C Ganiz, William M Pottenger, and Christopher D Janneck. Recent advances in literature based discovery. *Journal of the American Society for Information Science and Technology, JASIST (Submitted)*, 2005.

[5] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

[6] Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models, 2025. URL https://arxiv.org/abs/2506.04178.

[7] Andrej Karpathy. autoresearch: Ai agents running research on single-gpu nanochat training automatically. https://github.com/karpathy/autoresearch, 2026.

[8] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

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

[10] Yizhou Liu, Ziming Liu, and Jeff Gore. Superposition yields robust neural scaling. *arXiv preprint arXiv:2505.10465*, 2025.

[11] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling, 2025. URL https://arxiv.org/abs/2501.19393.

[12] OpenAI. Introducing deep research. https://openai.com/index/introducing-deep-research/, 2025. Accessed: 2025-02-02.

[13] Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *arXiv preprint arXiv:2505.06708*, 2025.

[14] Yakub Sebastian, Eu-Gene Siew, and Sylvester O Orimaye. Emerging approaches in literature-based discovery: techniques and performance review. *The Knowledge Engineering Review*, 32:e12, 2017.

[15] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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

[17] Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. The ideation-execution gap: Execution outcomes of llm-generated versus human research ideas. *arXiv preprint arXiv:2506.20803*, 2025.

[18] Chenglei Si, Zitong Yang, Yejin Choi, Emmanuel Candès, Diyi Yang, and Tatsunori Hashimoto. Towards execution-grounded automated ai research. *arXiv preprint arXiv:2601.14525*, 2026.

[19] Neil R Smalheiser. Literature-based discovery: Beyond the abcs. *Journal of the American Society for Information Science and Technology*, 63(2):218–224, 2012.

[20] Don R Swanson. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18, 1986.

[21] Don R Swanson. Undiscovered public knowledge. *The Library Quarterly*, 56(2):103–118, 1986.

[22] DR Swanson. Literature-based discovery? the very idea. In *Literature-based discovery*, pages 3–11. Springer, 2008.

[23] Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, pages 1–3, 2025.

[24] Jingqi Tong, Mingzhe Li, Hangcheng Li, Yongzhuo Yang, Yurong Mou, Weijie Ma, Zhiheng Xi, Hongji Chen, Xiaoran Liu, Qinyuan Cheng, et al. Ai can learn scientific taste. *arXiv preprint arXiv:2603.14473*, 2026.

[25] Manya Wadhwa, Tiasa Singha Roy, Harvey Lederman, Junyi Jessy Li, and Greg Durrett. Create: Testing llms for associative creativity, 2026. URL https://arxiv.org/abs/2603.09970.

[26] Kevin Wang, Ishaan Javali, MichaŁ Bortkiewicz, Benjamin Eysenbach, et al. 1000 layer networks for self-supervised rl: Scaling depth can enable new goal-reaching capabilities. *arXiv preprint arXiv:2503.14858*, 2025.

[27] Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific inspiration machines optimized for novelty. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 279–299, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.18. URL https://aclanthology.org/2024.acl-long.18/.

[28] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

[29] Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cycleresearcher: Improving automated research via automated review. *arXiv preprint arXiv:2411.00816*, 2024.

[30] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

[31] Xinran Zhao, Boyuan Zheng, Chenglei Si, Haofei Yu, Ken Liu, Runlong Zhou, Ruochen Li, Tong Chen, Xiang Li, Yiming Zhang, et al. The ramon llull's thinking machine for automated ideation. *arXiv preprint arXiv:2508.19200*, 2025.

[32] Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. DeepResearcher: Scaling deep research via reinforcement learning in real-world environments. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 414–431, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.22. URL https://aclanthology.org/2025.emnlp-main.22/.

A. arXiv Paper Domain Classification

We use the arXiv category taxonomy. Since Computer Science has many subfields, we further group the arXiv category labels into broader macro-domains.

Domain	arXiv Labels
Language	cs.CL
ML/AI	cs.LG, cs.AI, cs.NE, cs.MA
Robotics	cs.RO
Vision	cs.CV, cs.GR, cs.MM, cs.SD
Theory	cs.CC, cs.DS, cs.FL, cs.LO, cs.DM, cs.CG, cs.GT, cs.CR, cs.IT
Systems	cs.AR, cs.OS, cs.DC, cs.NI, cs.PF, cs.SY, cs.PL, cs.SE, cs.DB, cs.IR, cs.SI
Society	cs.CY
HCI	cs.HC
CS-Other	cs.ET, cs.GL, cs.OH, cs.DL, cs.NA, cs.MS, cs.CE, cs.SC
Economics	econ
Electrical Engineering & Systems Science (EE & Sys. Sci.)	eess
Mathematics	math
Physics	astro-ph, cond-mat, gr-qc, hep-ex, hep-lat, hep-ph, hep-th, math-ph, nlin, nucl-ex, nucl-th, physics, quant-ph
Quantitative Biology (Quant. Bio.)	q-bio
Quantitative Finance (Quant. Fin.)	q-fin
Statistics	stat

Table 1: Mapping from domains to arXiv category labels.

B. Further Quantitative Analysis

In this section, we provide a more granular, domain-by-domain breakdown of our quantitative results to better understand the performance characteristics of each model across different scientific disciplines.

Table 2 presents the average similarity scores across the entirety of GIANTSBENCH. The results demonstrate that GIANTS-4B achieves consistent, robust improvements over the other methods. Notably, this performance gap holds steady across highly diverse fields, ranging from largely theoretical domains like Mathematics and Theory to applied disciplines such as Vision, Robotics, and HCI.

To further validate the generalization capabilities of our approach and ensure the model is not merely memorizing training distributions, we separately evaluated performance on a highly constrained subset of GIANTSBENCH: **Test-unseen-parents**. This subset exclusively contains test examples where the parent papers were never encountered during the training phase.

The results for this holdout set are detailed in Table 3. Encouragingly, GIANTS-4B maintains its substantial performance gap over the baselines even on unseen literature. This indicates that our model has successfully learned generalized mechanisms for cross-paper insight synthesis that translate effectively to novel scientific texts.

Domain	Base	SFT	SFT-think	gemini-2.5-pro	gemini-3-pro	GIANTS-4B
Economics	$4.66_{\pm 0.21}$	$4.78_{\pm 0.20}$	$4.94_{\pm 0.21}$	$4.75_{\pm 0.22}$	$4.50_{\pm 0.20}$	$5.47_{\pm 0.22}$
Language	$4.50_{\pm 0.10}$	$4.75_{\pm 0.10}$	$4.84_{\pm 0.11}$	$4.45_{\pm 0.11}$	$4.14_{\pm 0.10}$	$5.50_{\pm 0.11}$
Vision	$4.28_{\pm 0.10}$	$4.59_{\pm 0.10}$	$4.57_{\pm 0.10}$	$4.11_{\pm 0.10}$	$4.14_{\pm 0.10}$	$5.52_{\pm 0.12}$
Theory	$4.54_{\pm 0.10}$	$4.72_{\pm 0.10}$	$4.80_{\pm 0.11}$	$4.67_{\pm 0.11}$	$4.58_{\pm 0.11}$	$5.63_{\pm 0.11}$
Mathematics	$4.64_{\pm 0.11}$	$4.73_{\pm 0.10}$	$4.81_{\pm 0.10}$	$4.76_{\pm 0.11}$	$4.58_{\pm 0.10}$	$5.66_{\pm 0.11}$
Quant. Fin.	$4.82_{\pm 0.25}$	$5.15_{\pm 0.23}$	$5.31_{\pm 0.25}$	$5.04_{\pm 0.26}$	$4.45_{\pm 0.24}$	$5.75_{\pm 0.26}$
ML/AI	$4.77_{\pm 0.11}$	$4.97_{\pm 0.11}$	$4.87_{\pm 0.11}$	$4.64_{\pm 0.11}$	$4.56_{\pm 0.11}$	$5.76_{\pm 0.12}$
Systems	$4.62_{\pm 0.11}$	$5.07_{\pm 0.11}$	$5.13_{\pm 0.11}$	$4.57_{\pm 0.11}$	$4.41_{\pm 0.11}$	$6.04_{\pm 0.12}$
Society	$4.87_{\pm 0.14}$	$5.31_{\pm 0.14}$	$5.33_{\pm 0.15}$	$4.40_{\pm 0.13}$	$3.82_{\pm 0.11}$	$6.05_{\pm 0.15}$
Physics	$4.82_{\pm 0.11}$	$5.27_{\pm 0.11}$	$5.30_{\pm 0.11}$	$4.95_{\pm 0.11}$	$4.65_{\pm 0.11}$	$6.14_{\pm 0.12}$
Robotics	$4.74_{\pm 0.11}$	$5.06_{\pm 0.11}$	$5.11_{\pm 0.11}$	$4.55_{\pm 0.11}$	$4.43_{\pm 0.11}$	$6.21_{\pm 0.11}$
EE & Sys.	$4.83_{\pm 0.11}$	$5.02_{\pm 0.11}$	$5.08_{\pm 0.11}$	$4.74_{\pm 0.11}$	$4.55_{\pm 0.11}$	$6.22_{\pm 0.12}$
HCI	$4.95_{\pm 0.12}$	$5.12_{\pm 0.11}$	$5.02_{\pm 0.11}$	$4.52_{\pm 0.11}$	$4.01_{\pm 0.10}$	$6.23_{\pm 0.12}$
Statistics	$5.45_{\pm 0.16}$	$5.38_{\pm 0.15}$	$5.51_{\pm 0.15}$	$5.50_{\pm 0.16}$	$5.68_{\pm 0.17}$	$6.46_{\pm 0.16}$
Quant. Bio.	$5.17_{\pm 0.17}$	$5.43_{\pm 0.17}$	$5.69_{\pm 0.17}$	$4.94_{\pm 0.17}$	$4.41_{\pm 0.15}$	$6.65_{\pm 0.17}$
CS-Other	$5.16_{\pm 0.16}$	$5.58_{\pm 0.16}$	$5.42_{\pm 0.16}$	$4.57_{\pm 0.15}$	$4.19_{\pm 0.14}$	$6.70_{\pm 0.16}$
Overall	$4.75_{\pm 0.03}$	$5.01_{\pm 0.03}$	$5.05_{\pm 0.03}$	$4.65_{\pm 0.03}$	$4.43_{\pm 0.03}$	$5.97_{\pm 0.03}$

Table 2: Average similarity scores (mean $\pm$ standard error) per domain on GiantsBench. The judge LM is gemini-3-pro.

Domain	Base	SFT	SFT-think	gemini-2.5-pro	gemini-3-pro	GIANTS-4B
Economics	$4.63_{\pm 0.22}$	$4.73_{\pm 0.21}$	$4.93_{\pm 0.21}$	$4.76_{\pm 0.23}$	$4.55_{\pm 0.21}$	$5.50_{\pm 0.23}$
Language	$4.59_{\pm 0.19}$	$4.79_{\pm 0.19}$	$4.96_{\pm 0.20}$	$4.52_{\pm 0.19}$	$4.28_{\pm 0.18}$	$5.62_{\pm 0.20}$
Vision	$4.45_{\pm 0.14}$	$4.57_{\pm 0.15}$	$4.64_{\pm 0.14}$	$4.14_{\pm 0.14}$	$4.33_{\pm 0.15}$	$5.71_{\pm 0.17}$
Theory	$4.63_{\pm 0.12}$	$4.80_{\pm 0.12}$	$4.76_{\pm 0.12}$	$4.84_{\pm 0.13}$	$4.75_{\pm 0.13}$	$5.71_{\pm 0.13}$
Mathematics	$4.63_{\pm 0.11}$	$4.72_{\pm 0.10}$	$4.80_{\pm 0.10}$	$4.74_{\pm 0.11}$	$4.57_{\pm 0.10}$	$5.63_{\pm 0.11}$
Quant. Fin.	$4.76_{\pm 0.27}$	$5.18_{\pm 0.25}$	$5.30_{\pm 0.26}$	$4.89_{\pm 0.27}$	$4.48_{\pm 0.25}$	$5.77_{\pm 0.27}$
ML/AI	$4.88_{\pm 0.15}$	$5.18_{\pm 0.16}$	$4.95_{\pm 0.15}$	$4.87_{\pm 0.16}$	$4.99_{\pm 0.16}$	$6.06_{\pm 0.16}$
Systems	$4.74_{\pm 0.15}$	$5.04_{\pm 0.15}$	$5.07_{\pm 0.15}$	$4.74_{\pm 0.16}$	$4.54_{\pm 0.15}$	$6.15_{\pm 0.16}$
Society	$4.96_{\pm 0.19}$	$5.45_{\pm 0.19}$	$5.26_{\pm 0.19}$	$4.47_{\pm 0.17}$	$3.77_{\pm 0.14}$	$6.13_{\pm 0.20}$
Physics	$4.82_{\pm 0.11}$	$5.28_{\pm 0.11}$	$5.29_{\pm 0.11}$	$4.99_{\pm 0.12}$	$4.67_{\pm 0.11}$	$6.16_{\pm 0.12}$
Robotics	$4.89_{\pm 0.13}$	$5.14_{\pm 0.12}$	$5.19_{\pm 0.13}$	$4.58_{\pm 0.13}$	$4.48_{\pm 0.13}$	$6.34_{\pm 0.13}$
EE & Sys.	$4.90_{\pm 0.13}$	$5.21_{\pm 0.13}$	$5.20_{\pm 0.12}$	$4.80_{\pm 0.13}$	$4.72_{\pm 0.13}$	$6.35_{\pm 0.14}$
HCI	$5.19_{\pm 0.15}$	$5.23_{\pm 0.14}$	$5.15_{\pm 0.14}$	$4.74_{\pm 0.14}$	$4.17_{\pm 0.12}$	$6.41_{\pm 0.15}$
Statistics	$5.47_{\pm 0.17}$	$5.43_{\pm 0.15}$	$5.53_{\pm 0.16}$	$5.49_{\pm 0.17}$	$5.67_{\pm 0.18}$	$6.51_{\pm 0.16}$
Quant. Bio.	$5.27_{\pm 0.18}$	$5.56_{\pm 0.18}$	$5.77_{\pm 0.18}$	$5.05_{\pm 0.18}$	$4.35_{\pm 0.16}$	$6.77_{\pm 0.18}$
CS-Other	$5.38_{\pm 0.19}$	$5.65_{\pm 0.18}$	$5.50_{\pm 0.18}$	$4.66_{\pm 0.17}$	$4.20_{\pm 0.16}$	$6.85_{\pm 0.18}$
Overall	$4.87_{\pm 0.04}$	$5.10_{\pm 0.04}$	$5.11_{\pm 0.04}$	$4.78_{\pm 0.04}$	$4.57_{\pm 0.04}$	$6.11_{\pm 0.04}$

Table 3: Average similarity scores (mean $\pm$ standard error) per domain on **Test-unseen-parents**, which indicates the subset of GiantsBench test set with only unseen parent papers. The judge LM is gemini-3-pro.

C. Prompts

This section details the exact prompt templates utilized throughout our pipeline for both generation and evaluation tasks.

C.1. Insight Anticipation Prompt

This is the prompt used for the insight anticipation task. The model is provided with the textual summaries of two parent papers, which are highlighted in red in the following figure. The prompt instructs the model to analyze these summaries, synthesize their core methodologies or findings, and generate a scientific insight or research direction that conceptually bridges and builds upon both foundational works.

```
You are given summaries of two research papers.
<papers>
Paper 1:
{Paper 1 summary}

Paper 2:
{Paper 2 summary}
</papers>

Your task is to generate a novel and non-obvious insight that emerges only when both papers are considered together.

Quality requirements:
1) SPECIFIC & PRECISE: Ground the insight in fine-grained details (methods, mechanisms, results, limitations) from both papers, not just general themes.
2) Avoid simple combination or summary of contributions: Go further to find at least one of:
  - Chain reasoning patterns (e.g., A→B from one work and B→C from the other suggests A→C).
  - New research directions that neither paper alone proposes (methods, architectures, experimental setups).
  - Surprising analogies, implicit principles, or shared structures not made explicit in either paper.
3) SELF-CONTAINED: Write the insight so a reader who hasn't read the papers can understand and act on it. Do not refer to Paper 1 or Paper 2.
Output format:
<think>Explain your reasoning for the insight. Step by step, show how specific aspects of Paper 1 and Paper 2 connect and why they lead to this insight.</think>

<insight>
A clear and self-contained statement of the insight (3-10 sentences).
</insight>
```

Figure 10: Prompt for insight anticipation. The summaries of two parent papers are in red.

C.2. Parents Identification Prompt

We present the prompt designed for the parents identification task. The primary objective of this prompt is to instruct the model to determine the most influential parent papers for a given downstream target paper. By providing the model with the downstream paper PDF, the prompt asks the model to trace the scientific lineage backward and identify which prior works from a candidate pool served as the conceptual ancestors of an insight drawn from the downstream paper.

```
You are an expert research analyst. Given this research paper, your task is to:
1. Identify two prior papers that this paper explicitly cites and builds upon by combining their ideas in a synergistic way.
  - The synergy should come from an insight that only emerges when information from both papers is synthesized.
  - It does not need to be the main insight or methodology of the paper. It could be a technique, approach, or design choice inspired by the two works — something that neither paper alone could fully support.
2. Explain your reasoning for why these two papers together provide the foundation for the insight.
3. Output the titles of the two papers in the following format:
<paper1>The first paper title...</paper1>
<paper2>The second paper title...</paper2>

### Expected Output
- A brief explanation of the synergy (why combining these two papers matters).
- The titles of the two papers inside <paper1> and <paper2> tags.
```

Figure 11: Prompt for identifying two parent papers whose ideas are synergistically combined to produce the given paper's key insight, which is extracted as the ground-truth target y^*.

C.3. Similarity Judge Prompt

Below, we present the prompt used for automated evaluation and reward scoring via an LM-as-a-judge framework. To quantitatively assess the quality and relevance of our generated insights, this prompt provides the evaluator model with both the ground-truth insight and the corresponding model-generated insight. The model is then instructed to critically compare the two texts and output a similarity assessment, evaluating them on semantic alignment.

Below is a research insight:
<research_insight>
{ground-truth insight}
</research_insight>
Below is a statement you need to evaluate:
<statement>
{model-generated insight}
</statement>
Task: Rate how similar the statement is to the research insight (1–10).

STRICT RULES:
- Similarity requires matching the SAME core idea.
- 'Inspired by', 'motivated by', or 'reasonable extension' ≠ same idea.
- Shared topic or keywords alone ≠ similarity.

Compare explicitly:
1) Key mechanism/method
2) Causal logic/workflow
3) Primary contribution/novelty

Downgrade if the statement:
- Omits the central mechanism
- Generalizes/abstracts the insight
- Proposes a new framework/direction

Scale:
1–2: Unrelated.
3–4: Shares topic but not the actual insight.
5–6: Partial conceptual overlap; misses at least one core mechanism or misaligns assumptions.
7–8: Strong match with only minor differences in mechanisms or assumptions.
9: Near-identical conceptual + causal + motivational mapping; only minor, non-substantive deviations.
10: Perfect: same ideas, same mechanism and roles, same objective/assumptions.
Output Format
Format your response as follows:
<think>
Explain your reasoning for the rating you chose.
</think>
<rating>a number between 1 and 10</rating>

Figure 12: Prompt for assessing the similarity between the ground-truth insight y^* (in red) and model-generated insight $\hat{y}$ (in orange). The scale is **bolded**.

C.4. Rewriting Insight Prompt

To ensure the generated insights could be evaluated consistently, we employed a rewriting step to standardize their format. Figure 13 displays the prompt used to instruct the model to rewrite the raw insight into a clear, standalone statement without losing its original semantic meaning.

```
You are an expert research analyst.
You are given detailed summaries of two research papers:
<papers>
Paper 1:
{Paper 1 summary}

Paper 2:
{Paper 2 summary}
</papers>

You are also given an ORACLE RESPONSE that explains how a follow-up paper integrates the ideas from the two papers above.

Your task:
- Reconstruct the reasoning process that could lead to the *same final insight* as in the ORACLE RESPONSE, but using only the content of the two papers above.
- Do NOT mention, reference, or imply the existence of the follow-up paper, the oracle, or any external work.
- Your response must be self-contained and logically derived only from the two papers.
- Your response should preserve the full level of *conceptual and technical specificity* implied in the oracle response.

<oracle_response>
{Raw synergy explanation}
</oracle_response>

OUTPUT FORMAT (return ONLY these three sections, nothing else):
<think>
Internal reasoning and planning for how to reconstruct the synergy between the two papers.
Explain how you will approach integrating their ideas and aligning the final insight with the oracle's conceptual structure (without mentioning the oracle explicitly in the output itself).
</think>

<insight>
The final rewritten insight stated clearly and self-contained.
Assume the reader has not seen Paper 1 or Paper 2.
Do not refer to Paper 1 or Paper 2 by name or imply their existence.
The insight should capture the same underlying conceptual synergy and technical details as the oracle response.
</insight>

<reasoning>
Step-by-step explanation of how this insight is logically derived from the two papers.
Explicitly identify which core ideas or mechanisms from each paper combine to produce the final integrated conclusion.
Be explicit and precise about the conceptual dependencies, theoretical links, and reasoning chain that lead to the stated insight.
</reasoning>
```

Figure 13: Prompt for rewriting the insight as a standalone statement. The summaries of two parent papers are in red. The raw synergy explanation is in orange.

D. Hyperparameters for GIANTS-4B and Baselines

We detail the optimization and training configurations used for training GIANTS-4B and the SFT baselines below. The hyperparameter choices were guided by standard practices for aligning large language models, with slight adjustments made to accommodate our specific sequence lengths and batch sizes.

RL Hyperparameters. Table 4 outlines the configuration used during the reinforcement learning phase, specifically utilizing the GRPO algorithm. We applied a conservative KL penalty to prevent the model from drifting too far from the reference policy.

Hyperparameter	Value
Algorithm	GRPO [15]
Training steps	400
Train batch size	64
Group size	8
Max prompt length	3000
Max response length	1296
Learning rate	1×10^{-6}
Entropy coefficient	0.002
KL loss coefficient	0.001
KL loss type	`low_var_kl`
Sampling temperature (train / val)	0.6 / 0.6
Max batched tokens	32768
Uneven Clipping (low / high)	0.2 / 0.5

Table 4: Key reinforcement learning hyperparameters used in our experiments.

Supervised Learning Hyperparameters. Table 5 details the setup for our supervised fine-tuning (SFT) baselines.

Hyperparameter	Value
Algorithm	SFT
Epochs	10
Train batch size	64
Max length (input + output)	8192
Learning rate	1×10^{-6}
Gradient checkpointing	True

Table 5: Key supervised fine-tuning hyperparameters used in our experiments.

E. Ablations with Diverse LM Judges

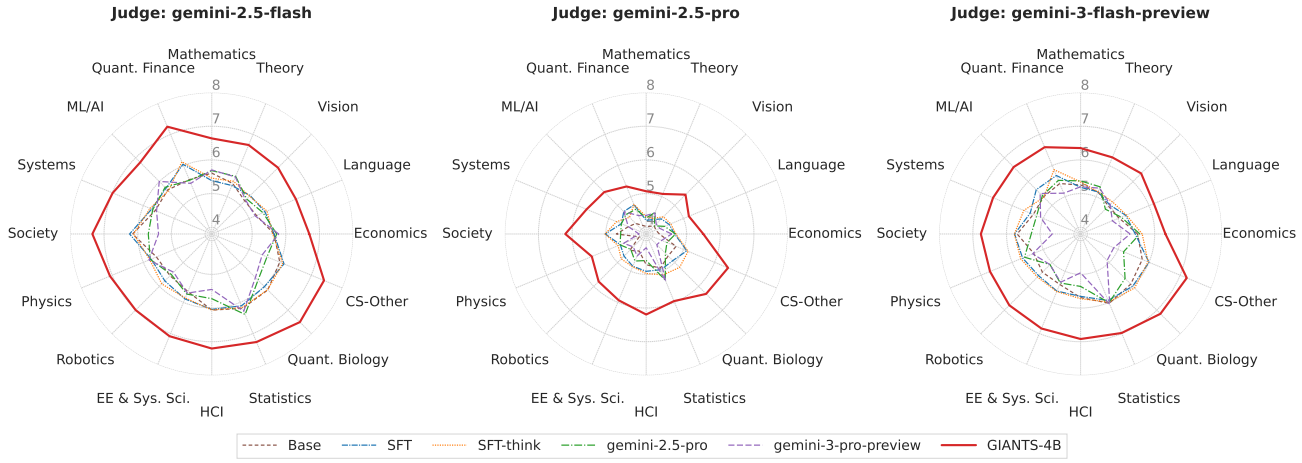

Figure 14: **Similarity scores on *Test-unseen-parents***. This is the subset of GIANTSBENCH test set with only unseen parent papers. Across all three LM judges, GIANTS-4B is ranked as the top-performing model.

F. Human Evaluation Details

We conduct two preliminary human studies, each involving two PhD students in Computer Science as annotators.

F.1. Human Evaluation of Insight Similarity

To validate the reliability of our automated metrics, we conducted a preliminary human evaluation study focusing on insight similarity. Two annotators were tasked with scoring the semantic alignment between model-generated insights and ground-truth insights. To facilitate this process and ensure an unbiased environment, we developed a custom web interface, shown in Figure 15. Importantly, the human annotators were provided with the exact same scoring rubric and guidelines that were given to the LM judge, allowing us to accurately measure the consistency between human judgment and our automated evaluation pipeline, with the exact context provided to the LM judge.

Both annotators rated the same 30 pairs of insights generated by the base model and GIANTS-4B against their corresponding ground-truth insights. For each insight, we compute the average similarity score across the two annotators. To compute the win rate under human judge, we compare the averaged scores of the base model and GIANTS-4B for each pair and ignore ties.

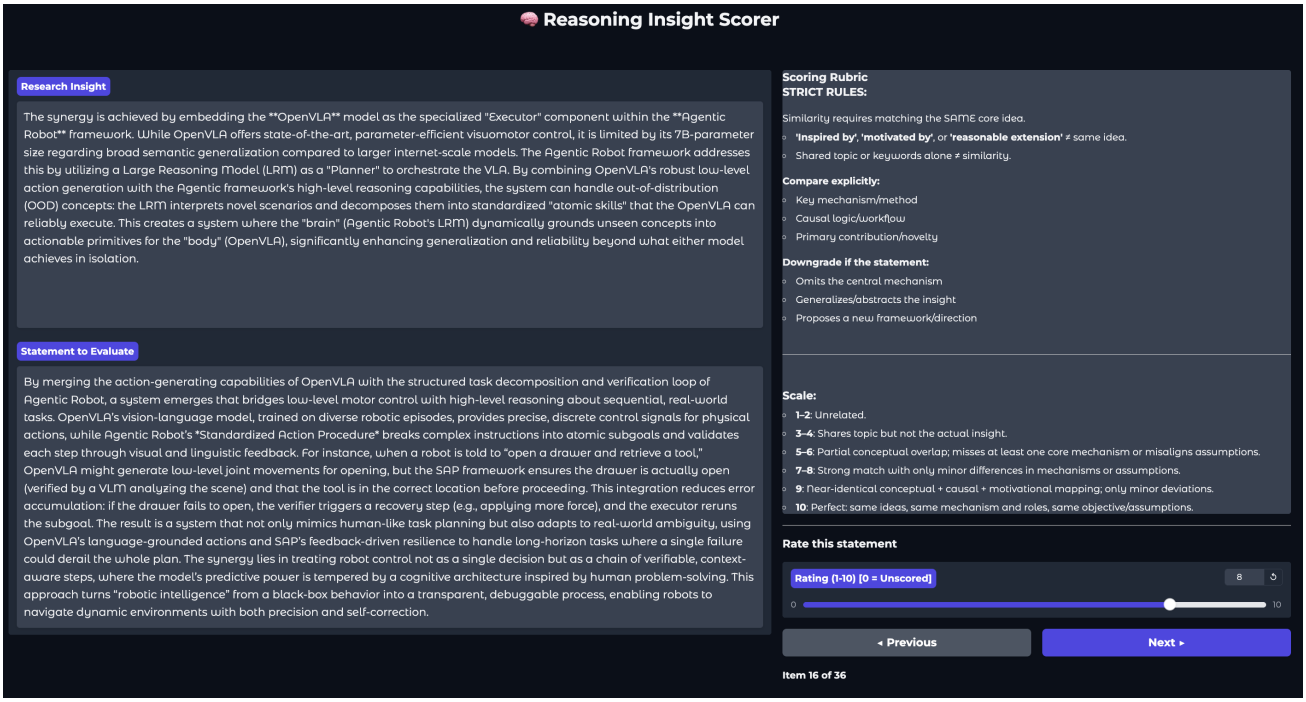

Figure 15: **Human eval labeling interface for assessing similarity.** This is the Gradio interface that was used by human annotators to measure the consistency of LM similarity scores with human ratings. The rubric provided to the humans exactly matched that provided to the LM judge for consistency (with markdown formatting for human interpretability).

F.2. Human Evaluation of Insight Feasibility

In the study on insight feasibility, both annotators rated the same 15 pairs of insights generated by the base model and GIANTS-4B along two dimensions: algorithmic complexity and conceptual clarity. For each insight and each dimension, we report the average rating across the two annotators. Figure 16 shows the rubric for the two dimensions.

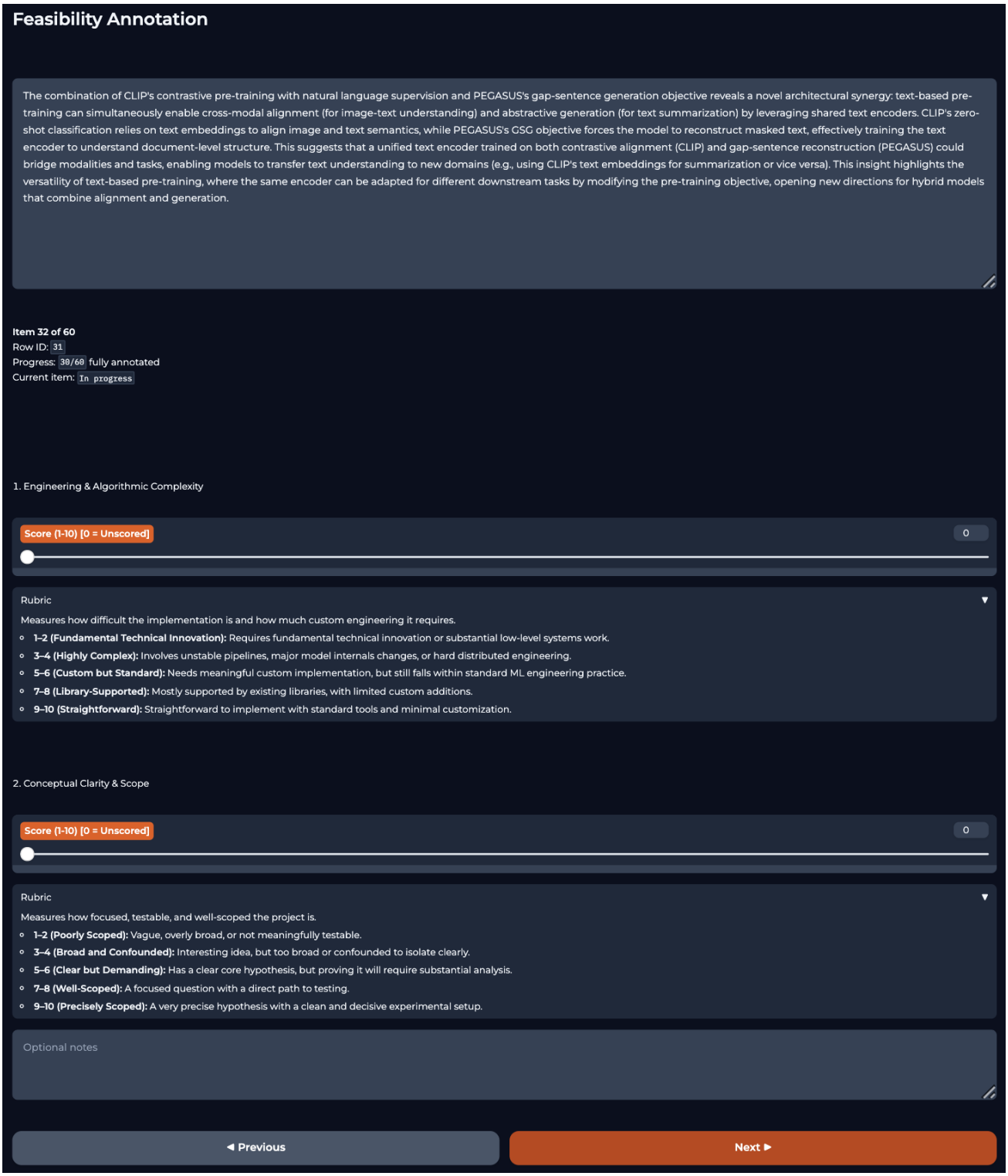

Figure 16: **Human eval labeling app for assessing feasibility.** This is the Gradio interface that was used by human annotators to measure the feasibility of a research insight along two dimensions: engineering/algorithmic complexity and conceptual clarity.