GIANTS: GENERATIVE INSIGHT ANTICIPATION FROM SCIENTIFIC LITERATURE

Anonymous authors
Paper under double-blind review

ABSTRACT

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 human expert 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 33% 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.

1 INTRODUCTION

Language Models (LMs) are becoming useful tools for scientific discovery (Karpathy, 2026). Recent work has shown promising results, from virtual LM teams designing SARS-CoV-2 nanobody binders (Swanson et al., 2025) to models proposing NLP research directions that human experts judge to be novel (Si et al., 2024). 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 (Si et al., 2025), often due to the lack of diversity and feasibility (Wadhwa et al., 2026).

To bridge this gap, we propose the task of *insight anticipation*: given a small set of prior papers, can a model anticipate a 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 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 a plausible 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

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

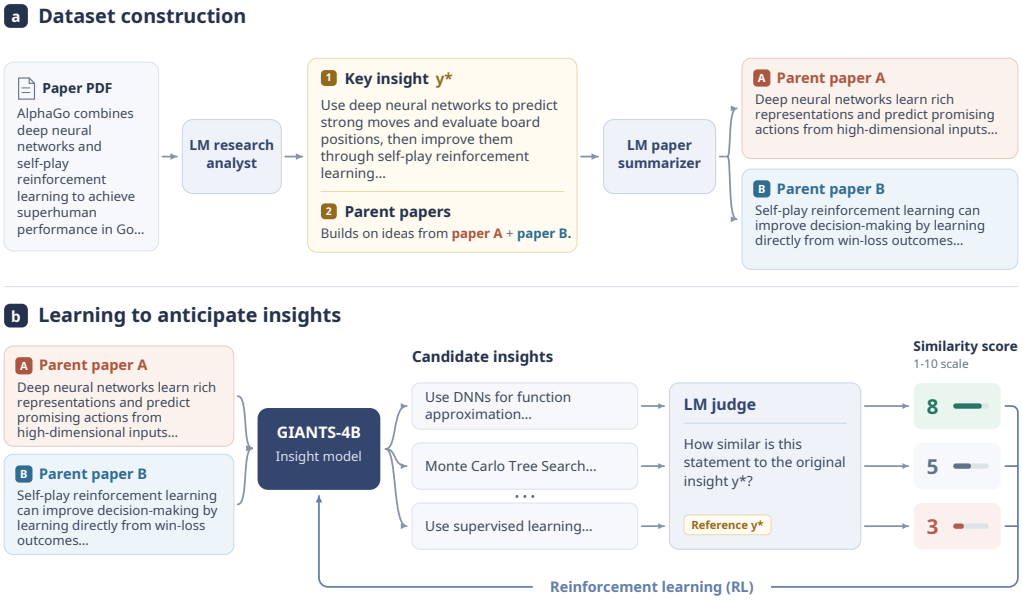


Figure 1: **Overview of GIANTSBENCH and `GIANTS-4B`.** *(top)* For dataset construction, we use an LM as a research analyst to process each paper PDF to identify two parent papers whose ideas are synergistically combined to produce the paper's key insight, which is extracted as the ground-truth target 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.

signal for scientific reasoning: learning to anticipate the downstream insights implied by a lineage of parent works provides a useful signal for improving how models combine and build on prior work.

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 (Asai et al., 2026), 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 $\rho = 0.645$, $p < 0.001$, $n = 180$).

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 (Shao et al., 2024) 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 (e.g., `gemini-3-pro`, `claude-opus-4-7`, and `Qwen3-4B`) on GIANTSBENCH. `Qwen3-4B` performs similarly to `gemini-3-pro` and `claude-opus-4-7`, 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 33% 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 (Tong et al., 2026), 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 proprietary and open LMs such as `gemini-3-pro`, improves conceptual clarity of insights, and generalizes zero-shot to unseen domains.

2 DEFINING A BENCHMARK FOR INSIGHT ANTICIPATION

Task Definition. We define an *insight* as a natural language description of a paper's primary methodological or empirical advance. 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^* and the model-generated insight as $\hat{y}$. The task challenges the model to reason about the joint influence of A and B and produce a $\hat{y}$ that is semantically similar to y^*. Figure 12 shows the insight generation prompt used for training and evaluation. To manage context constraints, we intentionally restrict the context to two parent papers.

We do not treat y^* as the *unique* valid insight derivable from (x_A, x_B): many scientifically sound insights may follow from the same parents, and y^* is one realized instance among them. We therefore view insight anticipation as conditional generation over a distribution of plausible downstream insights, with y^* serving as a grounded, verifiable sample from that distribution rather than a closed-form reconstruction target. To account for the one-to-many nature of this mapping, Appendix D evaluates each model against *all* realized downstream children of the same parent pair.

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. 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. 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 13). We then download the two parent papers, prompt `gemini-2.5-flash` to summarize the full paper (see prompt in Appendix H.3), and use these summaries as the input x. For y^*, we use the synergy explanation that is part of the output to the parent-identification prompt in Figure 13. We prompt `gemini-3-pro` to rewrite the insight as a standalone statement without reference to the downstream paper (see prompt in Appendix H.5). To verify the quality of our dataset, we conduct two human expert studies (see details in Appendix F) to validate that y^* faithfully reflects each downstream paper's stated contribution, and that the selected parent papers are the true conceptual foundations of the downstream paper.

Table 1: **LM judgments of insight similarity correlate with human judgments.** We ask 10 human annotators and the LM judge 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 $\rho = 0.645$, $p < 0.001$, $n = 180$). $^*p < 0.05$, $^{**}p < 0.001$.

Domain	Spearman ρ
EE & Sys. Sci.	0.789^*
Vision	0.702^{**}
Robotics	0.678^{**}
HCI	0.661^{**}
Mathematics	0.377^*
Overall	0.645^{**}

Temporal Hold-Out Evaluation. To assess generalization, we evaluate on a *future held-out test set* of downstream papers published after the training cutoff. We split each (x, y^*) pair by its downstream paper's publication date, using papers published before July 1, 2023 for training. To test cross-domain

generalization, we restrict training to cs.CL (Computation and Language), yielding $N = 10{,}335$ examples. For evaluation, we sample up to 600 papers per domain published after July 1, 2023 (see the taxonomy in Appendix B), yielding 7,504 test papers. We also report results on **Test-unseen-parents** ($N = 5{,}294$), which excludes test examples sharing 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 14 for the prompt). We use `gemini-3.1-pro` as the primary judge model. To assess the reliability of the LM-based evaluation, we conduct a human evaluation study in which 10 graduate students rate the similarity between model-generated insights and ground-truth insights using the same rating scale as the LM judge ($n = 180$). The annotators have expertise spanning five different domains. The LM's scores show a statistically significant positive correlation with the average scores across human annotators (Table 1). In particular, we observe a Spearman rank correlation of $\rho = 0.645$ ($p < 0.001$). More details can be found in Appendix K.1.

3 Training an Insight Anticipation Model

We explore two training paradigms for insight anticipation: (1) distillation via supervised fine-tuning (SFT), and (2) reinforcement learning (RL) via similarity optimization. Across all experiments, we use `Qwen3-4B` as the base model.

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 (Wei et al., 2023), 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 (see 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 follows distillation approaches in reasoning as seen in works such as OpenThoughts (Guha et al., 2025) and s1 (Muennighoff et al., 2025).

3.2 Reinforcement Learning for Insight Anticipation

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. Therefore, we explore an alternative approach: using semantic similarity to the downstream paper's insights y^* as a proxy reward. 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* denotes the semantic equivalence between the generated insight and downstream paper's insight measured using an LM-as-a-Judge (Gu et al., 2025), matching the evaluation criterion in Section 2. This proxy reward is then optimized via RL to produce insights that match the downstream paper conditioned on the two prespecified parent papers. We optimize this reward using Group Relative Policy Optimization (GRPO) (Shao et al., 2024). 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). Implementation details are in Appendix I.

Crucially, to mitigate the risk of reward hacking, 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-3.1-pro` model for evaluation.

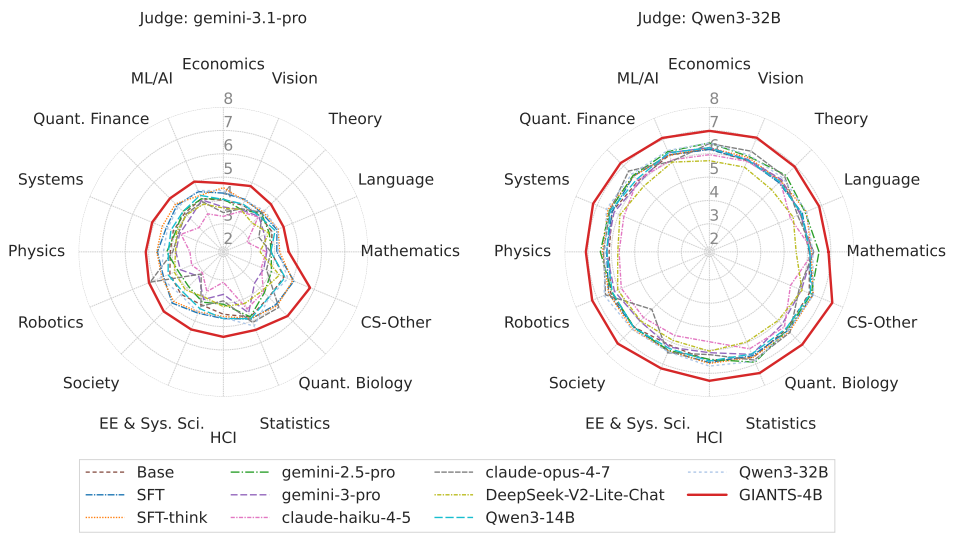


Figure 3: **Similarity scores on a subset of GIANTSBENCH containing only unseen parent papers.** `GIANTS-4B` achieves the highest similarity scores across domains and generalizes to new domains and distribution of parent papers, as judged by `gemini-3.1-pro` *(left)* and `Qwen3-32B` *(right)*.

4 EXPERIMENTAL EVALUATION 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 use the default or recommended sampling hyperparameters for every model (see details in Appendix I). 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.

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 8 (left), the base open model, `Qwen3-4B`, achieves an average similarity score of only 4.13 (see raw statistics corresponding to Figure 8 in Appendix C). Based on our evaluation rubric (Figure 14), 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. Notably, significantly larger proprietary models, including `gemini-3-pro` and `claude-opus-4-7`, perform similarly to the smaller `Qwen3-4B` baseline. This suggests that literature-grounded synthesis capabilities do not scale linearly with model size alone, reinforcing the necessity for specialized training paradigms like `GIANTS-4B`.

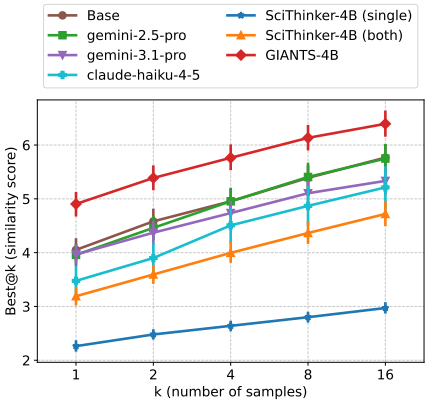


Figure 2: **`GIANTS-4B` remains strongest under test-time scaling.** As the number of generated samples per example increases, `GIANTS-4B` consistently outperforms the other baselines. SciThinker-4B (Tong et al., 2026) is a scientific-ideation model trained using a citation-preference reward model. Error bars show 95% confidence intervals.

2) Directly optimizing similarity scores substantially improves insight anticipation. `GIANTS-4B` achieves the highest performance among all evaluated methods. Figure 8 shows that, on the full future held-out test set, `GIANTS-4B` consistently outperforms the other baselines under two different judge LMs. Figure 3 shows that this advantage persists on the stricter **Test-unseen-parents** split, where `GIANTS-4B` achieves a 33% improvement over `gemini-3-pro`

(see raw statistics corresponding to Figure 8 and Figure 3 in Appendix C). 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 2. Since best-of-k evaluation requires scoring many samples per example, we conduct the inference-time scaling evaluation in Figure 2 on a subset of 480 examples sampled from GIANTSBENCH for cost considerations. We use temperature $= 0.6$ for all models in this evaluation.

In Figure 2, we additionally compare against SciThinker-4B (Tong et al., 2026), a related scientific-ideation model rather than a direct same-task baseline. In the original setting of Tong et al. (2026), 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 into a downstream insight. 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 2 shows that SciThinker-4B remains well below `GIANTS-4B` across all k, which suggests that general scientific ideation training may not transfer well to insight anticipation.

Relatedly, since a single parent pair can give rise to more than one valid downstream paper, we also evaluate each model against *all* realized downstream papers that build on the same parents and credit the best match across them. `GIANTS-4B` remains strongest under this multi-reference evaluation (Appendix D, Table 6). This confirms that `GIANTS-4B`'s advantage is not an artifact of comparing against a single deterministic target.

We also evaluate generations using SciJudge-30B, a third-party model from Tong et al. (2026) 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 4). 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 human studies. First, we ask 10 human annotators to rate the similarity between model-generated insights and their corresponding ground-truth insights for 90 head-to-head pairs comparing `GIANTS-4B` against the base model or comparing `GIANTS-4B` against `gemini-3-pro`, using the same rating scale as the LM judge (Figure 14). Each pair is independently rated by two to three annotators. We then compare the win rate of `GIANTS-4B` against each model in terms of alignment with the ground-truth insight (Figure 5). Against the base model, `GIANTS-4B` achieves a 75.0% win rate under the LM judge (`gemini-3.1-pro`) and an 89.7% win rate under human evaluation. Against `gemini-3-pro`, `GIANTS-4B` achieves an 81.0% win rate under the LM judge and a 70.4% win rate under human evaluation. Second, we assess the feasibility of generated insights along two axes: engineering/algorithmic complexity (rated on a scale of 1–10, where higher scores indicate easier implementation) and conceptual clarity/scope (rated on a scale of 1–10, where higher scores indicate clearer and better-scoped ideas), comparing `GIANTS-4B` against both the base model and `gemini-3-pro`. Figure 6 shows that all three models produce insights of comparable algorithmic complexity, while `GIANTS-4B` significantly improves conceptual clarity over both the base model and `gemini-3-pro`, making the generated ideas more interpretable and actionable. Details regarding human studies are in Appendix K.

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 8 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. This suggests that the model transfers to unseen domains and temporally held-out papers rather than memorizing domain-specific heuristics. In addition, Figure 3 shows model performance on a subset of test examples whose parent papers are entirely unseen during training. `GIANTS-4B` remains the

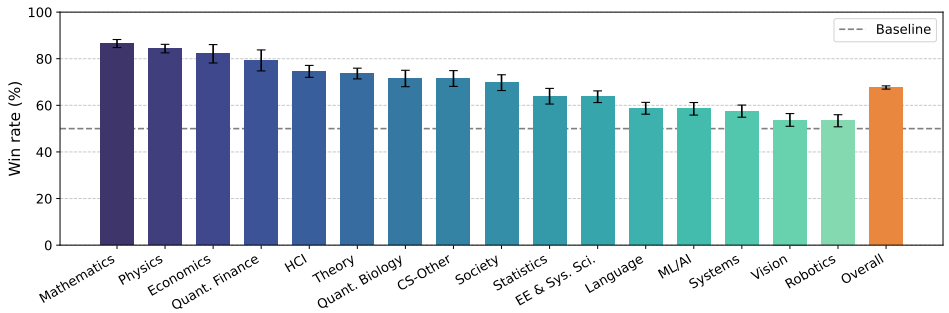


Figure 4: **`GIANTS-4B` achieves a 68% overall win rate against the base model under SciJudge-30B.** We evaluate win-rate using SciJudge-30B (Tong et al., 2026), a third-party judge trained to compare research abstracts by likely citation impact. This provides complementary evidence that optimizing for insight anticipation also improves performance under an independent, impact-oriented evaluation signal. Error bars show standard errors.

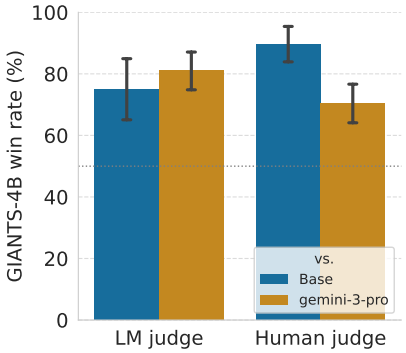


Figure 5: **Win rates of `GIANTS-4B` for insight similarity against the base model and `gemini-3-pro`.** We ask both human annotators and an LM judge to score the similarity between a model-generated insight and its corresponding ground-truth downstream insight, and find that `GIANTS-4B` better matches the ground-truth insight than the other baselines. Error bars show standard errors.

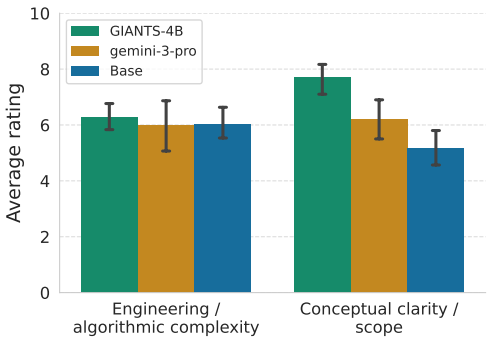


Figure 6: **`GIANTS-4B` produces insights with similar perceived algorithmic complexity but substantially higher conceptual clarity than both the base model and `gemini-3-pro`.** Human annotators assess the feasibility of generated insights along two axes: (1) engineering / algorithmic complexity and (2) conceptual clarity. Error bars show 95% confidence intervals.

top-performing method on this subset, suggesting that its gains are not driven by partial overlap in parent-paper lineage. Additionally, we find that on the subset of GIANTSBENCH restricted to papers published after the release of `Qwen3-4B` (April 28, 2025; Table 5), and therefore after its pretraining data cutoff, `GIANTS-4B` retains its advantage over all baselines. Therefore, the gains cannot stem from the base model having memorized these papers.

The improvements are not explained by metric gaming or judge style. Four lines of evidence indicate that `GIANTS-4B` improves genuine synthesis rather than exploiting the similarity judge. **(i) Train/test judge separation.** The RL reward uses `gemini-2.5-flash`, whereas our primary evaluation uses the independent `gemini-3.1-pro` judge, and `GIANTS-4B` is ranked first across five judges spanning three model families (Appendix J). **(ii) Style and length control.** Using `gemini-3.1-pro` to rewrite every model's output into a standardized surface form leaves the ranking and `GIANTS-4B`'s margin essentially unchanged (Table 7), so the metric rewards content rather than phrasing. **(iii) Agreement with an independent, non-similarity signal.** SciJudge-30B, trained to predict citation impact, prefers `GIANTS-4B` over the base model in 68% of comparisons (Figure 4), and human experts independently rate its insights as both more similar to the ground truth and more conceptually clear (Figures 5 and 6). **(iv) Synthesis.** Performance collapses when the two parents are replaced by a single parent, shuffled with a random paper, or removed entirely, but is preserved when the parent summaries are reordered (Table 8), showing that `GIANTS-4B` genuinely synthesizes both inputs rather than exploiting topical cues in a single summary, positional bias, or the prior over target insights.

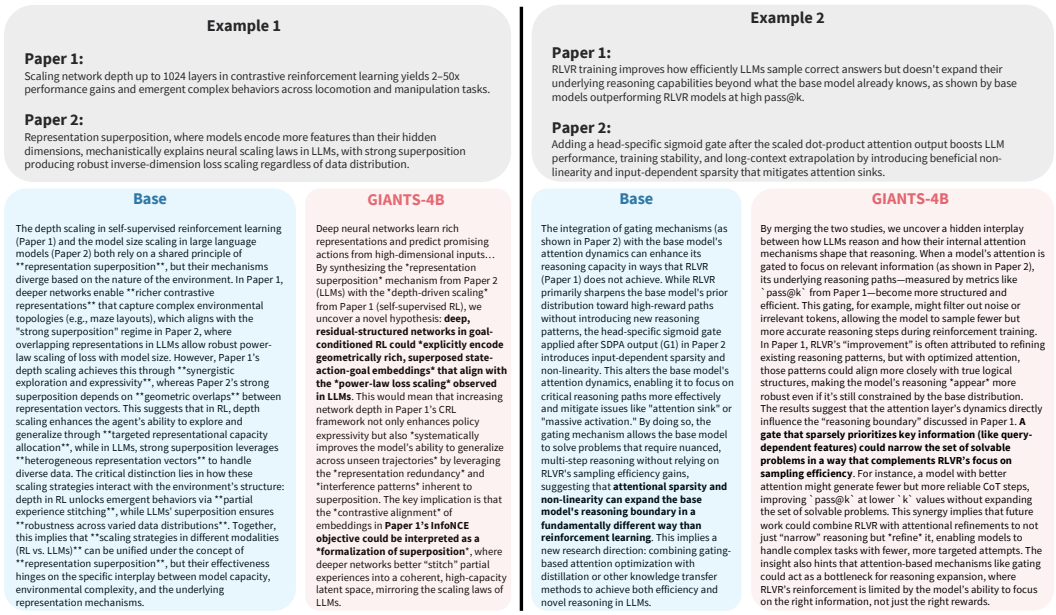


Figure 7: **Qualitative comparison of insights derived from NeurIPS 2025 award-winning papers.** *(left)* `GIANTS-4B` identifies a more concrete cross-paper mechanism than the base model (`Qwen3-4B`). *(right)* `GIANTS-4B` produces a grounded, more plausible interaction, while the base model makes a 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 of the full parent-paper summaries used as model input.

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 7, left) shows insights derived from Wang et al. (2025) and Liu et al. (2025). 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 7, right) shows insights based on Yue et al. (2025) and Qiu et al. (2025). 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 RL concentrates probability mass across useful reasoning trajectories. 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. Beyond the base model, Appendix G extends this comparison to include `gemini-3-pro` (Figure 11) and illustrates a characteristic failure mode of frontier models: although their outputs are fluent and internally coherent, they tend to drift from the realized contribution, either narrowing it to a different mechanism or over-reaching toward a more ambitious or speculative research project. These qualitative examples are consistent with human evaluations. `GIANTS-4B` more often produces insights that connect the parent papers in a concrete, interpretable, and grounded way.

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 (OpenAI, 2025; Zheng et al., 2025), idea generation (Si et al., 2024; Wang et al., 2024; Li et al., 2024; Zhao et al., 2025; Tong et al., 2026), idea execution (Si et al., 2025; 2026; Karpathy, 2026), and paper review (Weng et al., 2024). Asai et al. (2026) 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 (Si et al., 2024; Li et al., 2024). SCIMON (Wang et al., 2024) 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 (Tong et al., 2026). Goel et al. (2025) train AI co-scientists with rubric-based rewards to generate research plans given a concrete research goal; in contrast, we target the upstream insight-generation phase, producing a paper's core insight from a pair of parent papers, which could in turn feed such plan-generation systems. PRESCIENCE asks a model to generate a future paper's title and abstract conditioned on prior work and other historical context (Ajith et al., 2026). 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 (Swanson, 1986a; 2008; 1986b). 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 (Sebastian et al., 2017; Ganiz et al., 2005; Bekhuis, 2006; Smalheiser, 2012). 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 task for literature-grounded hypothesis generation, one component of the broader scientific-discovery pipeline. 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 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. While our human expert study confirms that the selected parents are rated as strong conceptual foundations for the downstream papers, we emphasize that conceptual attribution is fundamentally difficult even for human experts, since such ratings are external judgments that may not capture the actual creative process that inspired the original authors. Automated conceptual attribution thus remains an open problem, and unifying it with insight synthesis (building on recent efforts such as MIR (Garikaparthi et al., 2025)) is an important direction for future work. By assuming an oracle literature selection criterion, we explicitly decoupled insight generation from parent selection, which may not be feasible for some scientific tasks. Our primary metric also rewards similarity to realized historical insights and therefore does not credit equally valid insights that diverge from the historical record. Our multi-reference evaluation, independent citation-impact judge, and human studies partially mitigate this, but developing metrics that reward scientific novelty and validity independently of the realized target remains important future work. Finally, we rely on LMs throughout dataset construction; although the target insight is anchored to real downstream papers and validated by humans, a fully model-independent construction pipeline (e.g., varying the summarizer and parent-identifier across model families) would further isolate the effect of the construction LMs and is left to future work.

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 is a natural next step. Ultimately, testing these models in human-in-the-loop research settings will help clarify their practical value as assistants for literature-grounded hypothesis generation.

7 AI USE STATEMENT

Language models are a core component of our research methodology, and every such use is described in the paper. To construct GIANTSBENCH, we used `gemini-2.5-flash` to identify parent papers and summarize papers, and `gemini-3-pro` to rewrite downstream insights into standalone target statements (Section 2; prompts in Appendix H.2 and H.5). We used LMs as similarity judges: `gemini-2.5-flash` as the reward model during RL training, and `gemini-3.1-pro`, `Qwen3-32B`, and other models for evaluation. Proprietary and open-weight LMs also serve as baselines. We validated these LM-based components with human studies: experts rated the quality of the dataset and the parent-paper lineage (Appendix F), and we measured agreement between the LM judge and human similarity ratings (Table 1). Beyond these methodological uses, we used generative AI tools to help write and edit code used to implement our methods and experiments; the authors reviewed and tested all such code. We did not use generative AI tools to propose or refine research hypotheses, to design or give feedback on the research methodology or experiments, to support qualitative data analysis, or to interpret results. Developing theoretical models, formulating mathematical claims, writing proofs, and translation are not applicable to this work. Additionally, we used generative AI tools to help polish the paper's writing and to help create and refine figures. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

8 ETHICS STATEMENT

This research involves training the `GIANTS-4B` model on a dataset of 10,335 publicly available arXiv preprints to synthesize scientific insights. While insight anticipation 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.

9 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://anonymous.4open.science/r/giants-391C/README.md. We will publicly release our benchmark and model weights. Our implementation is built upon the verl framework (https://verl.readthedocs.io/en/latest/). All experimental details, including hyperparameter settings, are documented in Appendix I. 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.

REFERENCES

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.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

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*, pp. 1–7, 2026.

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

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

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.

Aniketh Garikaparthi, Manasi Patwardhan, Aditya Sanjiv Kanade, Aman Hassan, Lovekesh Vig, and Arman Cohan. Mir: Methodology inspiration retrieval for scientific research problems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28614–28659, 2025.

Shashwat Goel, Rishi Hazra, Dulhan Jayalath, Timon Willi, Parag Jain, William F Shen, Ilias Leontiadis, Francesco Barbieri, Yoram Bachrach, Jonas Geiping, et al. Training ai co-scientists using rubric rewards. *arXiv preprint arXiv:2512.23707*, 2025.

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.

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.

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

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.

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

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.

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

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.

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.

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.

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.

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.

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.

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

Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020.

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

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

DR Swanson. Literature-based discovery? the very idea. In *Literature-based discovery*, pp. 3–11. Springer, 2008.

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*, pp. 1–3, 2025.

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.

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.

Kevin Wang, Ishaan Javali, MichaĹ 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.

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 (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 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/.

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.

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.

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.

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.

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 (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 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 Additional Results

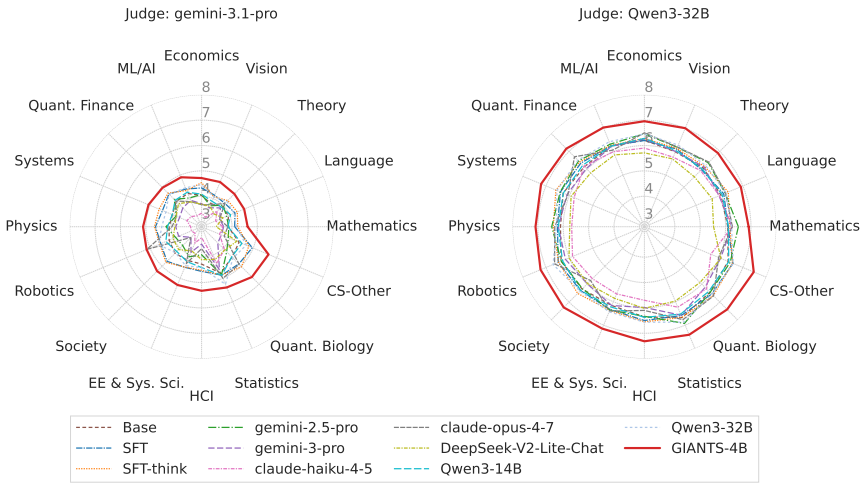


Figure 8: **Similarity scores on GIANTSBENCH (higher is better).** `GIANTS-4B` consistently achieves the highest similarity scores across domains. *(left)* scores from our primary evaluation judge `gemini-3.1-pro`. *(right)* scores from `Qwen3-32B`.

B 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 2: Mapping from domains to arXiv category labels.

C 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 3 presents the average similarity scores across 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. `claude-haiku-4-5` and `claude-opus-4-7` were evaluated on a subset of 480 examples sampled from GIANTSBENCH for cost considerations.

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 4. 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.

To further rule out the possibility of data contamination, we also evaluate performance on a subset of GIANTSBENCH containing only papers published after the release date of `Qwen3-4B`: April 28, 2025. This subset contains $N = 1{,}596$ examples. As shown in Table 5, the same overall trends hold for `GIANTS-4B` and the related baselines under the similarity metric. We exclude the Quantitative Finance domain from this analysis because it contains fewer than 30 samples in this subset.

Domain	Base	SFT	SFT-think	gemini-2.5-pro	gemini-3-pro	claude-haiku-4-5	claude-opus-4-7	DeepSeek-V2-Lite-Chat	Qwen3-14B	Qwen3-32B	GIANTS-4B
Mathematics	$3.85_{\pm 0.07}$	$4.11_{\pm 0.08}$	$4.22_{\pm 0.08}$	$3.87_{\pm 0.08}$	$3.67_{\pm 0.07}$	$3.63_{\pm 0.25}$	$3.83_{\pm 0.35}$	$3.38_{\pm 0.05}$	$3.84_{\pm 0.07}$	$4.17_{\pm 0.08}$	$4.62_{\pm 0.08}$
Language	$3.94_{\pm 0.08}$	$4.22_{\pm 0.08}$	$4.32_{\pm 0.09}$	$3.86_{\pm 0.08}$	$3.58_{\pm 0.07}$	$3.40_{\pm 0.28}$	$3.57_{\pm 0.27}$	$3.71_{\pm 0.07}$	$4.04_{\pm 0.08}$	$4.18_{\pm 0.09}$	$4.62_{\pm 0.09}$
Theory	$3.94_{\pm 0.08}$	$4.10_{\pm 0.08}$	$4.26_{\pm 0.09}$	$4.00_{\pm 0.08}$	$3.85_{\pm 0.08}$	$3.67_{\pm 0.26}$	$4.10_{\pm 0.31}$	$3.50_{\pm 0.06}$	$3.90_{\pm 0.08}$	$4.21_{\pm 0.08}$	$4.63_{\pm 0.09}$
Vision	$3.78_{\pm 0.07}$	$4.14_{\pm 0.08}$	$4.08_{\pm 0.08}$	$3.69_{\pm 0.08}$	$3.61_{\pm 0.08}$	$3.20_{\pm 0.22}$	$3.73_{\pm 0.30}$	$3.67_{\pm 0.06}$	$3.88_{\pm 0.08}$	$4.07_{\pm 0.08}$	$4.71_{\pm 0.09}$
Economics	$4.04_{\pm 0.16}$	$4.33_{\pm 0.17}$	$4.54_{\pm 0.17}$	$4.07_{\pm 0.16}$	$3.71_{\pm 0.14}$	$3.33_{\pm 0.26}$	$3.67_{\pm 0.27}$	$3.69_{\pm 0.13}$	$4.08_{\pm 0.16}$	$4.46_{\pm 0.18}$	$4.72_{\pm 0.18}$
ML/AI	$4.22_{\pm 0.09}$	$4.41_{\pm 0.09}$	$4.34_{\pm 0.09}$	$4.03_{\pm 0.09}$	$3.93_{\pm 0.08}$	$3.27_{\pm 0.31}$	$3.87_{\pm 0.37}$	$3.86_{\pm 0.07}$	$4.27_{\pm 0.09}$	$4.46_{\pm 0.09}$	$4.91_{\pm 0.09}$
Quant. Fin.	$4.21_{\pm 0.20}$	$4.62_{\pm 0.19}$	$4.67_{\pm 0.19}$	$4.32_{\pm 0.20}$	$3.85_{\pm 0.19}$	$3.37_{\pm 0.20}$	$4.23_{\pm 0.35}$	$4.09_{\pm 0.17}$	$4.27_{\pm 0.20}$	$4.61_{\pm 0.20}$	$4.98_{\pm 0.20}$
Systems	$4.05_{\pm 0.08}$	$4.47_{\pm 0.09}$	$4.61_{\pm 0.09}$	$3.89_{\pm 0.08}$	$3.74_{\pm 0.08}$	$3.43_{\pm 0.24}$	$3.87_{\pm 0.38}$	$3.89_{\pm 0.07}$	$4.05_{\pm 0.08}$	$4.33_{\pm 0.09}$	$5.07_{\pm 0.10}$
Physics	$4.08_{\pm 0.08}$	$4.62_{\pm 0.09}$	$4.66_{\pm 0.09}$	$4.21_{\pm 0.09}$	$3.82_{\pm 0.08}$	$3.23_{\pm 0.24}$	$3.97_{\pm 0.32}$	$3.85_{\pm 0.06}$	$4.14_{\pm 0.08}$	$4.37_{\pm 0.09}$	$5.11_{\pm 0.10}$
Robotics	$4.18_{\pm 0.09}$	$4.54_{\pm 0.09}$	$4.58_{\pm 0.09}$	$3.91_{\pm 0.08}$	$3.78_{\pm 0.08}$	$3.23_{\pm 0.28}$	$5.13_{\pm 0.48}$	$4.02_{\pm 0.07}$	$4.33_{\pm 0.09}$	$4.64_{\pm 0.09}$	$5.15_{\pm 0.09}$
Society	$4.28_{\pm 0.11}$	$4.73_{\pm 0.12}$	$4.79_{\pm 0.12}$	$3.85_{\pm 0.10}$	$3.39_{\pm 0.08}$	$3.17_{\pm 0.18}$	$3.40_{\pm 0.23}$	$4.08_{\pm 0.10}$	$4.18_{\pm 0.11}$	$4.25_{\pm 0.12}$	$5.28_{\pm 0.13}$
EE & Sys. Sci.	$4.24_{\pm 0.08}$	$4.53_{\pm 0.09}$	$4.52_{\pm 0.09}$	$4.09_{\pm 0.09}$	$3.89_{\pm 0.08}$	$3.50_{\pm 0.36}$	$4.37_{\pm 0.43}$	$3.83_{\pm 0.07}$	$4.29_{\pm 0.09}$	$4.46_{\pm 0.09}$	$5.29_{\pm 0.10}$
HCI	$4.27_{\pm 0.10}$	$4.56_{\pm 0.10}$	$4.52_{\pm 0.10}$	$3.91_{\pm 0.08}$	$3.49_{\pm 0.07}$	$3.23_{\pm 0.27}$	$3.67_{\pm 0.23}$	$4.10_{\pm 0.08}$	$4.41_{\pm 0.10}$	$4.42_{\pm 0.10}$	$5.33_{\pm 0.10}$
Statistics	$4.78_{\pm 0.14}$	$4.87_{\pm 0.13}$	$4.86_{\pm 0.13}$	$4.85_{\pm 0.14}$	$4.64_{\pm 0.14}$	$4.37_{\pm 0.41}$	$5.00_{\pm 0.46}$	$4.19_{\pm 0.10}$	$4.88_{\pm 0.14}$	$5.26_{\pm 0.14}$	$5.39_{\pm 0.14}$
Quant. Bio.	$4.50_{\pm 0.13}$	$4.88_{\pm 0.14}$	$5.03_{\pm 0.15}$	$4.20_{\pm 0.13}$	$3.70_{\pm 0.11}$	$4.07_{\pm 0.29}$	$4.90_{\pm 0.48}$	$4.10_{\pm 0.11}$	$4.48_{\pm 0.14}$	$4.55_{\pm 0.13}$	$5.60_{\pm 0.15}$
CS-Other	$4.40_{\pm 0.13}$	$4.96_{\pm 0.14}$	$4.81_{\pm 0.13}$	$3.94_{\pm 0.11}$	$3.64_{\pm 0.10}$	$3.57_{\pm 0.25}$	$3.77_{\pm 0.33}$	$4.30_{\pm 0.11}$	$4.52_{\pm 0.13}$	$4.70_{\pm 0.14}$	$5.66_{\pm 0.14}$
Overall	$4.13_{\pm 0.02}$	$4.45_{\pm 0.03}$	$4.50_{\pm 0.03}$	$4.00_{\pm 0.02}$	$3.76_{\pm 0.02}$	$3.48_{\pm 0.07}$	$4.06_{\pm 0.09}$	$3.85_{\pm 0.02}$	$4.18_{\pm 0.02}$	$4.40_{\pm 0.03}$	$5.02_{\pm 0.03}$

Table 3: Average similarity scores (mean $\pm$ standard error) per domain on GIANTSBENCH ($N = 7{,}504$). The judge LM is `gemini-3.1-pro`.

Domain	Base	SFT	SFT-think	gemini-2.5-pro	gemini-3-pro	claude-haiku-4-5	claude-opus-4-7	DeepSeek-V2-Lite-Chat	Qwen3-14B	Qwen3-32B	GIANTS-4B
Mathematics	$3.85_{\pm 0.07}$	$4.09_{\pm 0.08}$	$4.21_{\pm 0.08}$	$3.85_{\pm 0.08}$	$3.66_{\pm 0.07}$	$3.63_{\pm 0.25}$	$3.83_{\pm 0.35}$	$3.37_{\pm 0.05}$	$3.83_{\pm 0.07}$	$4.17_{\pm 0.08}$	$4.60_{\pm 0.08}$
Language	$4.06_{\pm 0.14}$	$4.28_{\pm 0.15}$	$4.34_{\pm 0.16}$	$3.93_{\pm 0.15}$	$3.71_{\pm 0.13}$	$2.89_{\pm 0.35}$	$3.44_{\pm 0.38}$	$3.91_{\pm 0.13}$	$4.20_{\pm 0.15}$	$4.31_{\pm 0.16}$	$4.59_{\pm 0.15}$
Theory	$4.00_{\pm 0.09}$	$4.13_{\pm 0.09}$	$4.22_{\pm 0.10}$	$4.11_{\pm 0.10}$	$3.95_{\pm 0.10}$	$3.91_{\pm 0.33}$	$4.18_{\pm 0.40}$	$3.54_{\pm 0.07}$	$3.97_{\pm 0.09}$	$4.30_{\pm 0.10}$	$4.68_{\pm 0.10}$
Vision	$3.96_{\pm 0.11}$	$4.23_{\pm 0.12}$	$4.18_{\pm 0.12}$	$3.76_{\pm 0.11}$	$3.76_{\pm 0.11}$	$3.67_{\pm 0.36}$	$3.80_{\pm 0.43}$	$3.76_{\pm 0.10}$	$4.01_{\pm 0.12}$	$4.18_{\pm 0.12}$	$4.85_{\pm 0.13}$
Economics	$4.01_{\pm 0.16}$	$4.32_{\pm 0.18}$	$4.55_{\pm 0.18}$	$4.07_{\pm 0.17}$	$3.72_{\pm 0.15}$	$3.32_{\pm 0.30}$	$3.48_{\pm 0.28}$	$3.68_{\pm 0.14}$	$4.05_{\pm 0.17}$	$4.43_{\pm 0.18}$	$4.74_{\pm 0.18}$
ML/AI	$4.26_{\pm 0.12}$	$4.59_{\pm 0.13}$	$4.62_{\pm 0.13}$	$4.26_{\pm 0.13}$	$4.21_{\pm 0.12}$	$3.56_{\pm 0.52}$	$4.12_{\pm 0.57}$	$4.02_{\pm 0.10}$	$4.42_{\pm 0.13}$	$4.64_{\pm 0.13}$	$5.06_{\pm 0.13}$
Quant. Fin.	$4.14_{\pm 0.20}$	$4.62_{\pm 0.20}$	$4.70_{\pm 0.21}$	$4.25_{\pm 0.21}$	$3.82_{\pm 0.20}$	$3.27_{\pm 0.18}$	$4.08_{\pm 0.34}$	$4.04_{\pm 0.18}$	$4.24_{\pm 0.22}$	$4.53_{\pm 0.21}$	$5.04_{\pm 0.21}$
Systems	$4.12_{\pm 0.11}$	$4.45_{\pm 0.12}$	$4.64_{\pm 0.13}$	$4.02_{\pm 0.11}$	$3.83_{\pm 0.11}$	$3.76_{\pm 0.41}$	$3.88_{\pm 0.51}$	$3.97_{\pm 0.11}$	$4.23_{\pm 0.12}$	$4.39_{\pm 0.12}$	$5.10_{\pm 0.13}$
Physics	$4.08_{\pm 0.08}$	$4.62_{\pm 0.09}$	$4.65_{\pm 0.09}$	$4.22_{\pm 0.09}$	$3.84_{\pm 0.08}$	$3.23_{\pm 0.24}$	$3.97_{\pm 0.32}$	$3.85_{\pm 0.07}$	$4.16_{\pm 0.08}$	$4.37_{\pm 0.09}$	$5.12_{\pm 0.10}$
Robotics	$4.28_{\pm 0.10}$	$4.58_{\pm 0.11}$	$4.65_{\pm 0.11}$	$3.91_{\pm 0.09}$	$3.79_{\pm 0.09}$	$3.27_{\pm 0.32}$	$5.19_{\pm 0.53}$	$4.07_{\pm 0.09}$	$4.36_{\pm 0.10}$	$4.70_{\pm 0.11}$	$5.25_{\pm 0.11}$
Society	$4.37_{\pm 0.15}$	$4.90_{\pm 0.16}$	$4.81_{\pm 0.16}$	$3.90_{\pm 0.13}$	$3.29_{\pm 0.08}$	$3.07_{\pm 0.13}$	$3.14_{\pm 0.25}$	$4.10_{\pm 0.13}$	$4.24_{\pm 0.14}$	$4.40_{\pm 0.15}$	$5.41_{\pm 0.17}$
EE & Sys. Sci.	$4.25_{\pm 0.10}$	$4.64_{\pm 0.11}$	$4.59_{\pm 0.10}$	$4.13_{\pm 0.10}$	$4.01_{\pm 0.10}$	$3.64_{\pm 0.43}$	$4.32_{\pm 0.49}$	$3.90_{\pm 0.08}$	$4.43_{\pm 0.11}$	$4.61_{\pm 0.11}$	$5.40_{\pm 0.12}$
HCI	$4.47_{\pm 0.12}$	$4.66_{\pm 0.12}$	$4.57_{\pm 0.12}$	$4.05_{\pm 0.10}$	$3.62_{\pm 0.09}$	$3.10_{\pm 0.33}$	$3.90_{\pm 0.32}$	$4.15_{\pm 0.10}$	$4.66_{\pm 0.12}$	$4.61_{\pm 0.12}$	$5.44_{\pm 0.13}$
Statistics	$4.82_{\pm 0.14}$	$4.91_{\pm 0.13}$	$4.89_{\pm 0.14}$	$4.81_{\pm 0.15}$	$4.63_{\pm 0.14}$	$4.44_{\pm 0.45}$	$5.07_{\pm 0.50}$	$4.19_{\pm 0.11}$	$4.89_{\pm 0.14}$	$5.24_{\pm 0.15}$	$5.42_{\pm 0.14}$
Quant. Bio.	$4.55_{\pm 0.14}$	$4.99_{\pm 0.15}$	$5.06_{\pm 0.16}$	$4.28_{\pm 0.14}$	$3.69_{\pm 0.12}$	$4.23_{\pm 0.32}$	$5.16_{\pm 0.53}$	$4.15_{\pm 0.12}$	$4.56_{\pm 0.15}$	$4.57_{\pm 0.14}$	$5.69_{\pm 0.16}$
CS-Other	$4.61_{\pm 0.15}$	$5.07_{\pm 0.16}$	$5.01_{\pm 0.16}$	$4.03_{\pm 0.13}$	$3.68_{\pm 0.11}$	$3.56_{\pm 0.28}$	$3.80_{\pm 0.35}$	$4.40_{\pm 0.13}$	$4.66_{\pm 0.15}$	$4.79_{\pm 0.16}$	$5.82_{\pm 0.16}$
Overall	$4.21_{\pm 0.03}$	$4.53_{\pm 0.03}$	$4.55_{\pm 0.03}$	$4.09_{\pm 0.03}$	$3.84_{\pm 0.03}$	$3.58_{\pm 0.09}$	$4.16_{\pm 0.11}$	$3.89_{\pm 0.02}$	$4.28_{\pm 0.03}$	$4.49_{\pm 0.03}$	$5.11_{\pm 0.03}$

Table 4: Average similarity scores (mean $\pm$ standard error) per domain on **Test-unseen-parents** ($N = 5{,}294$), which indicates the subset of GIANTSBENCH test set with only unseen parent papers. The judge LM is `gemini-3.1-pro`.

Domain	Base	SFT	SFT-think	gemini-2.5-pro	gemini-3-pro	claude-haiku-4-5	claude-opus-4-7	DeepSeek-V2-Lite-Chat	Qwen3-14B	Qwen3-32B	GIANTS-4B
Mathematics	$3.68_{\pm 0.13}$	$4.09_{\pm 0.16}$	$4.24_{\pm 0.16}$	$3.99_{\pm 0.15}$	$3.74_{\pm 0.14}$	$3.00_{\pm 0.26}$	$3.50_{\pm 0.96}$	$3.28_{\pm 0.09}$	$3.83_{\pm 0.15}$	$4.11_{\pm 0.16}$	$4.49_{\pm 0.16}$
Language	$3.97_{\pm 0.16}$	$4.31_{\pm 0.20}$	$4.40_{\pm 0.20}$	$3.94_{\pm 0.18}$	$3.82_{\pm 0.18}$	$3.67_{\pm 0.67}$	$4.67_{\pm 0.88}$	$3.69_{\pm 0.14}$	$4.03_{\pm 0.18}$	$4.29_{\pm 0.19}$	$4.60_{\pm 0.19}$
Theory	$4.06_{\pm 0.21}$	$4.01_{\pm 0.20}$	$4.39_{\pm 0.21}$	$3.83_{\pm 0.18}$	$3.92_{\pm 0.20}$	$4.25_{\pm 1.25}$	$3.75_{\pm 0.75}$	$3.30_{\pm 0.13}$	$3.78_{\pm 0.17}$	$4.22_{\pm 0.22}$	$4.72_{\pm 0.23}$
Vision	$3.80_{\pm 0.18}$	$4.06_{\pm 0.18}$	$4.30_{\pm 0.21}$	$3.72_{\pm 0.20}$	$3.59_{\pm 0.18}$	$3.00_{\pm 0.00}$	$3.00_{\pm 0.32}$	$3.70_{\pm 0.15}$	$3.66_{\pm 0.18}$	$4.07_{\pm 0.21}$	$4.67_{\pm 0.22}$
Economics	$4.29_{\pm 0.31}$	$4.44_{\pm 0.32}$	$4.87_{\pm 0.33}$	$4.48_{\pm 0.34}$	$4.18_{\pm 0.30}$	$3.40_{\pm 0.52}$	$3.00_{\pm 0.30}$	$4.22_{\pm 0.29}$	$4.35_{\pm 0.32}$	$4.62_{\pm 0.33}$	$5.11_{\pm 0.34}$
ML/AI	$4.24_{\pm 0.15}$	$4.51_{\pm 0.17}$	$4.40_{\pm 0.17}$	$4.06_{\pm 0.16}$	$4.24_{\pm 0.17}$	$4.00_{\pm 0.59}$	$4.50_{\pm 0.65}$	$4.05_{\pm 0.14}$	$4.31_{\pm 0.17}$	$4.63_{\pm 0.18}$	$5.06_{\pm 0.18}$
Systems	$4.09_{\pm 0.19}$	$4.25_{\pm 0.19}$	$4.44_{\pm 0.21}$	$3.68_{\pm 0.16}$	$3.53_{\pm 0.17}$	$3.29_{\pm 0.18}$	$3.29_{\pm 0.36}$	$3.90_{\pm 0.17}$	$4.09_{\pm 0.20}$	$4.24_{\pm 0.20}$	$4.99_{\pm 0.22}$
Physics	$4.12_{\pm 0.15}$	$4.62_{\pm 0.18}$	$4.72_{\pm 0.18}$	$4.27_{\pm 0.18}$	$3.68_{\pm 0.13}$	$3.60_{\pm 0.65}$	$4.90_{\pm 0.71}$	$3.93_{\pm 0.13}$	$4.08_{\pm 0.16}$	$4.40_{\pm 0.17}$	$5.16_{\pm 0.19}$
Robotics	$4.06_{\pm 0.16}$	$4.41_{\pm 0.17}$	$4.40_{\pm 0.18}$	$4.10_{\pm 0.16}$	$3.79_{\pm 0.16}$	$3.33_{\pm 0.80}$	$5.67_{\pm 1.12}$	$4.01_{\pm 0.14}$	$4.41_{\pm 0.18}$	$4.49_{\pm 0.18}$	$5.15_{\pm 0.18}$
Society	$4.66_{\pm 0.26}$	$5.12_{\pm 0.28}$	$5.36_{\pm 0.27}$	$4.15_{\pm 0.26}$	$3.67_{\pm 0.21}$	$3.00_{\pm 0.00}$	$3.17_{\pm 0.17}$	$4.40_{\pm 0.22}$	$4.67_{\pm 0.25}$	$4.55_{\pm 0.27}$	$5.89_{\pm 0.29}$
EE & Sys. Sci.	$4.27_{\pm 0.19}$	$4.66_{\pm 0.21}$	$4.55_{\pm 0.20}$	$4.19_{\pm 0.21}$	$3.96_{\pm 0.20}$	$4.75_{\pm 1.89}$	$4.75_{\pm 1.18}$	$3.86_{\pm 0.15}$	$4.42_{\pm 0.21}$	$4.56_{\pm 0.19}$	$5.42_{\pm 0.23}$
HCI	$4.36_{\pm 0.27}$	$4.92_{\pm 0.27}$	$4.44_{\pm 0.23}$	$3.97_{\pm 0.22}$	$3.47_{\pm 0.19}$	$5.50_{\pm 2.50}$	$4.00_{\pm 1.00}$	$4.21_{\pm 0.22}$	$4.62_{\pm 0.29}$	$4.62_{\pm 0.27}$	$5.49_{\pm 0.28}$
Statistics	$4.86_{\pm 0.29}$	$5.18_{\pm 0.28}$	$4.93_{\pm 0.28}$	$4.60_{\pm 0.31}$	$4.75_{\pm 0.31}$	$6.00_{\pm 1.00}$	$5.83_{\pm 1.25}$	$3.92_{\pm 0.17}$	$4.93_{\pm 0.30}$	$5.35_{\pm 0.32}$	$5.44_{\pm 0.29}$
Quant. Bio.	$4.46_{\pm 0.29}$	$4.67_{\pm 0.29}$	$4.98_{\pm 0.33}$	$4.21_{\pm 0.26}$	$3.66_{\pm 0.25}$	$3.50_{\pm 0.38}$	$3.71_{\pm 0.61}$	$4.12_{\pm 0.24}$	$4.28_{\pm 0.27}$	$4.33_{\pm 0.27}$	$5.52_{\pm 0.33}$
CS-Other	$4.48_{\pm 0.27}$	$5.06_{\pm 0.32}$	$4.89_{\pm 0.30}$	$3.65_{\pm 0.21}$	$3.52_{\pm 0.21}$	$3.50_{\pm 0.38}$	$3.75_{\pm 0.77}$	$4.32_{\pm 0.24}$	$4.40_{\pm 0.26}$	$4.59_{\pm 0.30}$	$5.70_{\pm 0.29}$
Overall	$4.17_{\pm 0.05}$	$4.49_{\pm 0.06}$	$4.55_{\pm 0.06}$	$4.04_{\pm 0.05}$	$3.83_{\pm 0.05}$	$3.72_{\pm 0.18}$	$4.07_{\pm 0.20}$	$3.87_{\pm 0.04}$	$4.20_{\pm 0.05}$	$4.44_{\pm 0.06}$	$5.08_{\pm 0.06}$

Table 5: Average similarity scores (mean $\pm$ standard error) per domain on the subset of GIANTSBENCH test set with only **papers published after the release date of `Qwen3-4B`** ($N = 1{,}596$). The judge LM is `gemini-3.1-pro`.

D EVALUATION AGAINST MULTIPLE DOWNSTREAM INSIGHTS (MULTI-REFERENCE EVALUATION)

Our primary evaluation compares each generated insight $\hat{y}$ to a single ground-truth insight y^* extracted from one downstream paper. Because a fixed pair of parent papers can plausibly give rise to more than one valid downstream insight, this single-target evaluation may understate a model's true performance. To account for this one-to-many structure, we additionally evaluate models against multiple realized downstream insights that build on the same parents.

Concretely, we identify all parent pairs in our test set that have at least four distinct downstream child papers in our corpus, yielding 26 parent pairs. For each pair, we score the model's generated insight against the ground-truth insight of *every* child paper using our LM judge (`gemini-3.1-pro`), and summarize the resulting set of similarity scores by its maximum, median, and mean across children. The maximum is the most informative of the three: it measures whether a model can match *any* of several plausible downstream developments. We then average each statistic over the 26 pairs and report the mean $\pm$ standard error across pairs in Table 6.

`GIANTS-4B` attains the highest similarity under all three aggregations, reaching a maximum similarity of 6.46 compared to 4.46 for `gemini-3-pro` and 5.19 for `DeepSeek-V2-Lite-Chat`. This indicates that the advantage of `GIANTS-4B` is not an artifact of comparing against a single deterministic target: even when a model is credited for matching any of several realized downstream insights, `GIANTS-4B` remains the strongest. These results are consistent with the distributional view of insight anticipation in Section 2, in which a parent pair maps to a distribution of plausible downstream insights rather than a unique output.

Model	Max ↑	Median ↑	Mean ↑
`gemini-2.5-pro`	$4.083_{\pm 0.394}$	$2.958_{\pm 0.141}$	$3.040_{\pm 0.147}$
`gemini-3-pro`	$4.462_{\pm 0.416}$	$3.038_{\pm 0.162}$	$3.105_{\pm 0.176}$
`DeepSeek-V2-Lite-Chat`	$5.192_{\pm 0.388}$	$3.692_{\pm 0.243}$	$3.808_{\pm 0.208}$
`Qwen3-4B`	$5.077_{\pm 0.474}$	$3.346_{\pm 0.267}$	$3.521_{\pm 0.229}$
`Qwen3-14B`	$5.423_{\pm 0.455}$	$3.577_{\pm 0.268}$	$3.742_{\pm 0.249}$
`Qwen3-32B`	$5.385_{\pm 0.484}$	$3.519_{\pm 0.265}$	$3.663_{\pm 0.247}$
`GIANTS-4B`	$6.462_{\pm 0.494}$	$4.058_{\pm 0.309}$	$4.313_{\pm 0.266}$

Table 6: **`GIANTS-4B` outperforms baselines when evaluated against multiple plausible downstream insights.** For the 26 parent-paper pairs that each have at least four distinct downstream child papers, we score each model-generated insight against every child's ground-truth insight and report the maximum, median, and mean similarity across children, averaged over pairs (mean $\pm$ standard error across the 26 pairs). The judge LM is `gemini-3.1-pro`. `GIANTS-4B` achieves the highest similarity under all three aggregation metrics.

E Robustness of the Evaluation

E.1 Style and Length Control

To verify that the similarity metric rewards the content of an insight rather than its surface form, we use `gemini-3.1-pro` to rewrite every model's generated insight into a standardized style and length (see the prompt in Figure 9), then recompute similarity to the ground-truth insights. As shown in Table 7, the rewriting leaves both the ranking and `GIANTS-4B`'s margin essentially unchanged, indicating that the metric is not driven by stylistic or length differences between models.

Model	Raw insights	Rewritten insights
`GIANTS-4B`	$4.904_{\pm 0.103}$	$4.929_{\pm 0.103}$
`Qwen3-4B`	$4.105_{\pm 0.096}$	$4.094_{\pm 0.095}$
`gemini-3-pro`	$3.804_{\pm 0.090}$	$3.906_{\pm 0.094}$

Table 7: **Model rankings are unchanged after standardizing style and length.** We use `gemini-3.1-pro` to rewrite each model's generated insight into a standardized surface form and recompute similarity to the ground-truth insights. The ranking and `GIANTS-4B`'s margin are essentially unchanged (mean $\pm$ standard error; a subset of 480 samples judged by `gemini-3.1-pro`).

You are a careful copy-editor. You will rewrite a research insight to a standardized length and style so that downstream similarity judgments are not confounded by surface phrasing. Preserve meaning exactly; change only length and style.

ORIGINAL STATEMENT:
<original>
{model-generated insight}
</original>

TARGET LENGTH
- Approximately 150 words (acceptable range: 130-170).
- One single paragraph. No section breaks.

TARGET STYLE
- Plain declarative academic prose.
- No markdown of any kind: no bold, italics, asterisks for emphasis, bullet points, numbered lists, headers, or block quotes.
- No em-dashes, no scare quotes, no rhetorical questions, no parenthetical asides used for emphasis.
- No TitleCased or quoted concept names invented by the author (e.g., rewrite '"Crowdsourced Pedagogical Competition" model' as 'a model in which ...'). Keep proper nouns and standard technical terms (e.g., Bloom's Taxonomy, Paper 1, Paper 2) as-is.
- No hedging adverbs beyond what the original uses ("clearly", "obviously", "crucially", "importantly" should be dropped unless the original load-bearingly hedges).
- Do not address the reader, do not use first person.

CONTENT RULES (these override the length target)
- Preserve every substantive claim, mechanism, condition, scope, and quantitative detail. If a claim cannot fit the word budget, exceed the budget rather than drop the claim.
- Do not add new claims, examples, motivations, justifications, caveats, or interpretations not present in the original.
- Do not strengthen or weaken any claim (preserve modality: "could" stays "could", "is" stays "is").
- Preserve cross-paper attributions (e.g., "from Paper 1", "in Paper 2").
- Numerical values, names, and citations must appear unchanged.

OUTPUT FORMAT
Output only the rewritten statement, wrapped in <statement></statement> tags. No preamble, no commentary, no explanation.

Figure 9: Prompt for rewriting a generated insight to a standardized length and style for the style and length control experiment (Table 7). The model-generated insight is in orange.

E.2 Parent-Ablation Controls

To test whether `GIANTS-4B` genuinely synthesizes both parent papers rather than exploiting shortcuts, we compare the standard setting (*Original*) against four controls. *Single* provides only one parent and isolates the contribution of a single paper; *Shuffled* replaces the second parent with a random paper and isolates topical-cue exploitation; *Reversed* swaps the order of the two parent summaries and isolates positional bias; and *No-parent* provides no parents and isolates the prior over target insights. Table 8 shows that performance drops sharply under Single, Shuffled, and No-parent, but is preserved under Reversed, indicating that the model relies on synthesizing both parents rather than on positional or distributional shortcuts.

Model	Single	Shuffled	Reversed	No-parent	Original
`GIANTS-4B`	$2.990_{\pm 0.043}$	$1.981_{\pm 0.037}$	$4.946_{\pm 0.106}$	$1.004_{\pm 0.003}$	$4.904_{\pm 0.103}$
`Qwen3-4B`	$2.460_{\pm 0.043}$	$1.667_{\pm 0.034}$	$4.067_{\pm 0.095}$	$1.004_{\pm 0.004}$	$4.169_{\pm 0.097}$

Table 8: **Parent-ablation controls confirm genuine two-paper synthesis.** *Single* provides one parent; *Shuffled* replaces the second parent with a random paper; *Reversed* swaps the order of the two parent summaries; *No-parent* provides no parents; *Original* is the standard setting. Performance drops sharply under Single, Shuffled, and No-parent but is preserved under Reversed, showing that the model synthesizes both parents rather than exploiting topical cues, positional bias, or the target prior (mean $\pm$ standard error; a subset of 480 samples judged by `gemini-3.1-pro`).

F Dataset Quality Validation

We conduct two human expert studies to validate the quality of GiantsBench: one verifying the fidelity of the extracted ground-truth insight y^*, and one verifying that the selected parent papers are the true conceptual foundations of the downstream paper. Each study involves four human experts, and each of the 50 sampled papers is rated by all four experts, yielding $N = 200$ ratings per study. Table 9 summarizes the results.

Validation study	Avg. score (1–10)	N
y^* extraction fidelity (§F.1)	$9.46_{\pm 0.77}$	200
Parent-paper lineage (§F.2)	$8.26_{\pm 1.97}$	200

Table 9: **Human experts validate the quality of GIANTSBENCH.** Each study samples 50 papers, rated by four experts on a 1-10 scale ($N = 200$). The high y^*-fidelity score confirms that extracted insights match papers' stated contributions, and the high lineage score confirms that selected parents capture the downstream paper's conceptual foundations. Values are mean $\pm$ standard deviation.

F.1 VALIDATING THE GROUND-TRUTH INSIGHT EXTRACTION PIPELINE

We verify that our extraction pipeline accurately captures what downstream papers explicitly claim as their contributions. We filter 50 papers that contain a clearly worded contribution statement (e.g., "We summarize our main contributions as follows…") and ask four human experts to rate the similarity between the paper's stated contributions and our extracted y^*, using the same similarity rubric given to the LM judge (Figure 14). The average similarity score is 9.46 (SD $= 0.77$, $N = 200$), corresponding to a level of agreement between near-identical and perfectly matched. This indicates that y^* faithfully reflects the contributions that downstream authors themselves report.

F.2 VALIDATING THE PARENT-PAPER LINEAGE

We assess whether the selected parent papers genuinely serve as the foundational conceptual building blocks of the downstream paper. We sample 50 papers and ask four human experts to rate, on a 1–10 scale, how well the selected parent papers serve as the conceptual foundations of the downstream paper, using the rubric shown in Figure 10. The average conceptual-foundation score is 8.26 (SD $= 1.97$, $N = 200$), indicating that the selected parents capture the main conceptual lineage of the downstream work. We note that automated conceptual attribution is fundamentally difficult even for human experts, and we discuss this further as a limitation in Section 6.

Task: Rate how well the selected parent papers serve as the conceptual foundations of the downstream paper on a scale of 1–10.

STRICT RULES:
- A high score requires that the downstream paper's core contribution meaningfully builds on, extends, combines, or depends on the selected parent papers.
- Shared topic, shared keywords, or citation relationship alone is not enough.
- Do not judge whether the parent papers are high quality or highly cited.
- Focus only on whether these parents explain the conceptual origin of the downstream paper's main idea.

Compare explicitly:
1.Core problem or motivation
2.Key method/mechanism
3.Conceptual dependency
4.Whether the downstream paper combines or extends both parents
5.Whether more obvious foundational papers seem missing

Downgrade if the selected parents:
- Are only background or related work
- Share a broad topic but not the core contribution
- Explain only one small component of the downstream paper
- Do not jointly support the downstream insight
- Seem replaceable by many other papers
- Miss a clearly more important conceptual foundation

Scale:
1–2: Not foundational.
The selected parent papers are unrelated or only superficially related to the downstream paper. They do not explain the downstream paper's main contribution.
3–4: Topically related but not conceptually foundational.
The parents share the same area, keywords, dataset, or general problem, but they mainly serve as background or related work. The downstream paper's core insight does not clearly depend on them.
5–6: Partial conceptual foundation.
At least one parent paper contributes meaningfully to the downstream paper's idea, but the pair only partially explains the downstream contribution. One parent may be weakly relevant, or important conceptual foundations may be missing.
7–8: Strong conceptual foundation.
Both parents are clearly relevant, and the downstream paper can be reasonably understood as extending, adapting, or combining their ideas. There may be some missing context, but the selected parents capture the main conceptual lineage.
9: Near-complete conceptual foundation.
The selected parents are central to the downstream paper's contribution. The downstream insight is strongly motivated by their combination or extension, with only minor missing context.
10: Essential conceptual foundation.
The downstream paper's core insight directly emerges from synthesizing or extending these two parent papers. Replacing either parent would substantially weaken the explanation of where the downstream contribution comes from.

Figure 10: **Rubric for rating parent-paper lineage.** Human experts use this rubric to rate, on a 1–10 scale, how well the selected parent papers serve as the conceptual foundations of the downstream paper.

G QUALITATIVE COMPARISONS

We present an additional qualitative example comparing `GIANTS-4B`, the base model (`Qwen3-4B`), and `gemini-3-pro` on the same pair of parent papers (Figure 11). The parents are the DEtection TRansformer (DETR) (Carion et al., 2020), an end-to-end, anchor-free, NMS-free detector that frames detection as direct set prediction, and STAC (Sohn et al., 2020), a simple semi-supervised object detection framework that trains a student on pseudo-labels generated by a teacher under strong augmentations. The realized downstream paper combines them by applying STAC-style teacher–student pseudo-labeling to a DETR-based detector, yielding a data-efficient, end-to-end detector that retains DETR's NMS-free design (realized as SAM-DETR for semi-supervised table detection). The reference (ground-truth) insight is precisely this combination.

All three models stay on-topic, but they differ in whether they preserve the reference's central mechanism. `GIANTS-4B` reconstructs the same synergy, a DETR teacher generates confidence-filtered pseudo-labels that train a student under strong augmentations via a consistency loss, while keeping DETR's anchor-free and NMS-free pipeline, and is scored $10/10$ by the LM judge. The base model and `gemini-3-pro` instead drift toward different contributions. The base model (`Qwen3-4B`) narrows the synergy to a mechanism the reference does not emphasize, reframing it as a fix for small-object detection through DETR's matching cost, and receives $4/10$. `gemini-3-pro`, despite producing the most polished and technically detailed output, over-reaches: it proposes an entirely new training paradigm, "Teacher–Student Bipartite Matching," that explicitly eliminates pseudo-label generation and instead matches the teacher's and student's output sets with the Hungarian algorithm, motivated by a problem (DETR's slow convergence) that the realized paper never addresses. Because this discards the reference's core mechanism, it receives only $3/10$, scoring even below the smaller, untuned base model.

This example illustrates two points. First, the similarity metric does not simply reward topical overlap or polished technical language; it rewards preservation of the reference insight's core mechanism and causal structure, which is why a fluent frontier model can score below a weaker model whose answer stays closer to the realized contribution. Second, it characterizes a recurring failure mode of strong frontier models on this task: their outputs are fluent and internally coherent, but tend to drift from the realized contribution. In particular, frontier models tend to narrow it to a different mechanism or over-reach toward a more ambitious or speculative research project. `GIANTS-4B`, by contrast, remains anchored to the synergy actually realized by the downstream paper.



Figure 11: **Qualitative comparison of `GIANTS-4B`, the base model, and `gemini-3-pro`.** Given the same two parent papers, DETR (Carion et al., 2020) and STAC (Sohn et al., 2020), only `GIANTS-4B` preserves the reference insight's central mechanism, applying STAC-style teacher-student pseudo-labeling to a DETR-based NMS-free detector, earning a $10/10$ similarity score. The base model narrows the contribution to small-object detection ($4/10$), while `gemini-3-pro` over-reaches to a new, pseudo-label-free training paradigm ($3/10$). The metric rewards mechanism preservation rather than topical overlap or surface fluency.

H PROMPTS

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

H.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 12: Prompt for insight anticipation. The summaries of two parent papers are in red.

H.2 Parent 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 13: 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^*.

H.3 Paper Summarization Prompt

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.*"

H.4 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 14: 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**.

H.5 Rewriting Insight Prompt

To ensure the generated insights could be evaluated consistently, we employed a rewriting step to standardize their format. Figure 15 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 15: 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.

I 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 10 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 (Shao et al., 2024)
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 10: Key reinforcement learning hyperparameters used in our experiments.

Supervised Learning Hyperparameters. Table 11 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 11: Key supervised fine-tuning hyperparameters used in our experiments.

Decoding and Temperature Protocol. For every model, we use the default or recommended sampling hyperparameters specified in its corresponding Hugging Face model card or API documentation, rather than tuning decoding settings per model. This follows common practice in prior benchmarking work such as HealthBench (Arora et al., 2025). Consequently, the temperature differences between the Gemini, Claude, DeepSeek baselines (temperature $= 0$) and the Qwen-based models (temperature $= 0.6$, `top-p` $= 0.95$, `top-k` $= 20$, `min-p` $= 0$) arise from each model's own recommended configuration, not from model-specific tuning: Qwen explicitly recommends these thinking-mode settings, which we follow for both training and inference. To further rule out that our gains are an artifact of these temperature differences, the best-of-k comparison in Figure 2 evaluates *all* models at a common temperature of 0.6, and `GIANTS-4B` remains the top-performing model across every k.

J Ablations with Diverse LM Judges

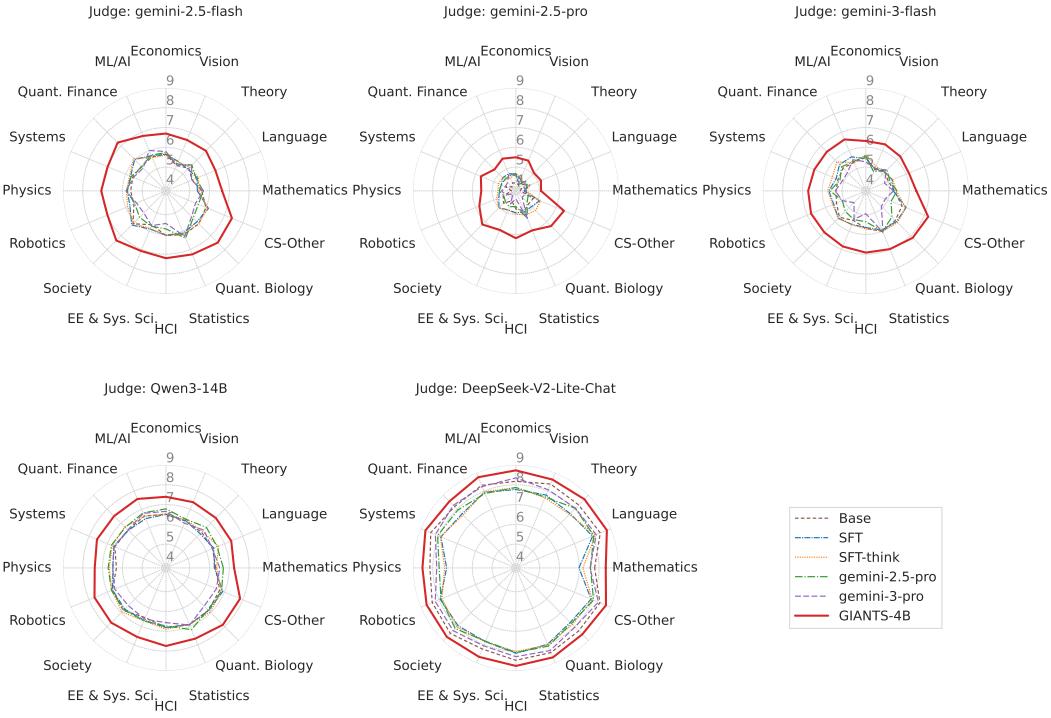


Figure 16: **Similarity scores on *Test-unseen-parents***. This is the subset of GIANTSBENCH test set with only unseen parent papers. Across all five LM judges, `GIANTS-4B` is ranked as the top-performing model.

K HUMAN EVALUATION DETAILS

We conduct two human studies. The first, on insight similarity, involves 10 graduate students with expertise across Electrical Engineering & Systems Science, Vision, Robotics, HCI, and Mathematics. The second, on insight feasibility, involves two graduate students in Computer Science as annotators.

K.1 HUMAN EVALUATION OF INSIGHT SIMILARITY

To validate the reliability of our automated metrics, we conducted a human evaluation study focusing on insight similarity. Ten graduate students, with expertise spanning Electrical Engineering & Systems Science, Vision, Robotics, HCI, and Mathematics, scored 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 17. 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.

In total, the annotators scored 180 insights, each generated by one of three models (the base model, `GIANTS-4B`, or `gemini-3-pro`), against its corresponding ground-truth insight. Each insight was independently rated by two to three annotators. For each insight, we average the similarity scores across its annotators and correlate these averages with the LM judge's scores to obtain the Spearman correlations in Table 1. To compute the win rate under the human judge, we organize the insights into 90 head-to-head pairs comparing `GIANTS-4B` against the base model (30 pairs) or against `gemini-3-pro` (60 pairs), and for each pair compare the two models' averaged scores, ignoring ties.

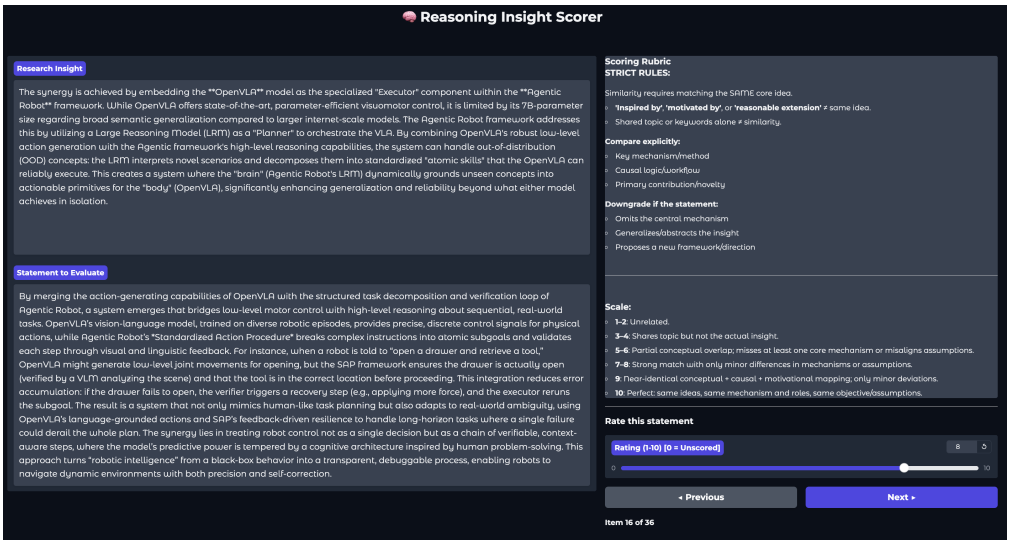


Figure 17: **Human eval labeling interface for assessing similarity.** This is the Gradio interface that was used by human annotators to measure the consistency of LM similarity scores with human ratings. The rubric provided to the humans exactly matched that provided to the LM judge for consistency (with markdown formatting for human interpretability).

K.2 HUMAN EVALUATION OF INSIGHT FEASIBILITY

In the study on insight feasibility, two annotators rated insights generated by three models (the base model, `GIANTS-4B`, and `gemini-3-pro`). For each of 15 examples, every model generates one insight, yielding 45 insights in total. Each insight is rated along two dimensions: engineering / algorithmic complexity (rated 1–10, where higher scores indicate easier implementation) and conceptual clarity (rated 1–10, where higher scores indicate clearer and better-scoped ideas). For each insight and dimension, we report the average rating across the two annotators. Figure 18 shows the rubric for the two dimensions.

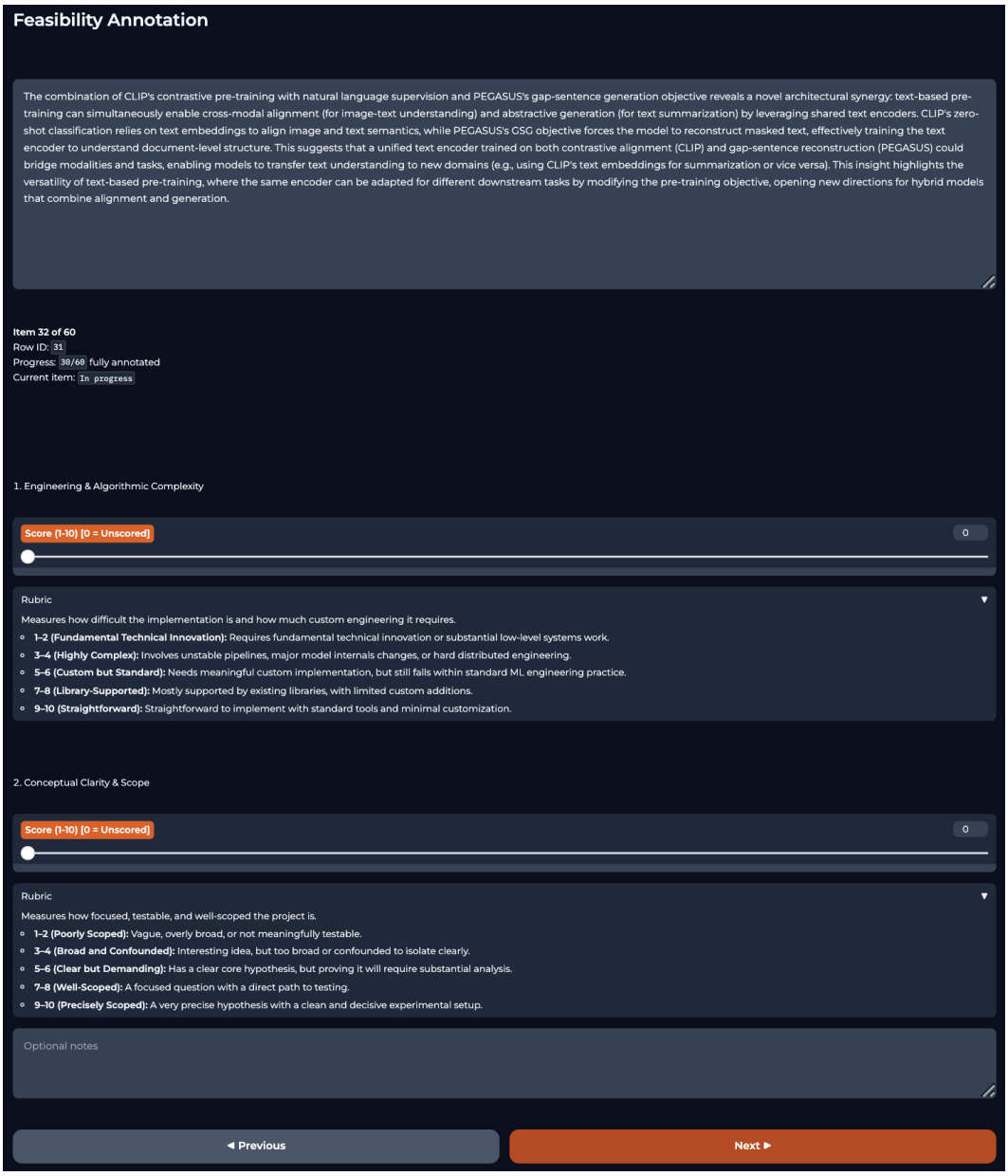


Figure 18: **Human eval labeling app for assessing feasibility.** This is the Gradio interface that was used by human annotators to measure the feasibility of a research insight along two dimensions: engineering/algorithmic complexity and conceptual clarity.