HorizonBench: Evaluating Ideation Agents on Future Scientific Discoveries

Markus Flicke[1, 2], **Haoyu He**[1, 2], **Kashyap Chitta**[3, 4], **Yong Cao**[1, 2], **Andreas Geiger**[1, 2, 4]
[1]University of Tübingen, [2]Tübingen AI Center, [3]ELLIS Institute Tübingen, [4]KE:SAI
Correspondence: markus.flicke.marburg@gmail.com

Abstract

As ideation agents become more capable, evaluating them poses an increasing challenge. LLM-as-a-judge approaches are brittle and gameable, while human evaluation is expensive and hard to scale. To address this, we introduce HorizonBench, an evaluation methodology leveraging future scientific discoveries. HorizonBench uses published papers as ground truth for ideation quality. Specifically, ideation agents pretraining and context retrieval data are strictly constrained to pre-publication knowledge, and scored on how well their predicted ideas align with subsequently published work. Using HorizonBench, we conduct the first unified comparison of five published ideation agents across multiple frontier model backbones. We verify that our soft recall metric captures meaningful ideation quality aligned to human judgement via a user study with 51 domain experts. We then uncover a surprising finding: a naïve baseline that receives only the research goal as an input consistently matches or outperforms all existing agents that require retrieval or multi-step reasoning, highlighting the importance of the research goal for ideation. Based on this result, we develop OSCAR, an ideation agent that distils retrieved papers into insights and explicitly reinforces the research goal in each step of the pipeline. OSCAR outperforms all evaluated agents on HorizonBench. We publish the code[1] and an online version[2] of HorizonBench.

1 Introduction

Automated ideation is at the forefront of frontier AI research. A typical ideation agent retrieves relevant literature from a database, decomposes the research goal through multi-step reasoning, and synthesises novel hypotheses or experimental plans. Besides ideation, LLM-based agents are now enabling automation across the full scientific pipeline, from literature review through experimentation to manuscript writing (Lu et al., 2024; Yamada et al., 2025; Tang et al., 2025). Yet when such agents are evaluated on completing a research pipeline end-to-end, 73.6% of their errors occur during research planning, when deciding **what** to investigate and **how** (Wang et al., 2026). Ideation shapes every downstream decision, and is thus the primary bottleneck for autonomous science. Despite its importance, the field currently lacks both a generalised, architecture-agnostic benchmark for ideation and a unified comparison of the growing number of ideation agents. Currently, each agent is evaluated under its own metrics, and many forgo human studies entirely (Sanyal et al., 2025; Lu et al., 2024), making it impossible to assess where the field actually stands.

Evaluating ideation is difficult because there is no single correct answer. The two most prominent approaches, LLM-as-a-judge and human evaluation, each have disadvantages. Human evaluation is expensive and hard to scale, while LLM judges are brittle. They can be gamed by surface-level features and are unreliable on out-of-distribution inputs, e.g., Jiang (2026) show that LLM quality judgements are negatively correlated with actual research impact. What is needed is an evaluation grounded in objective outcomes, and we observe that **published research** provides exactly such a ground truth. By constraining an agent to pre-publication knowledge and measuring **soft recall** between the proposed ideas and subsequently published work, we obtain a direct test of an agent's ability to anticipate future discoveries. An agent that solves this benchmark must include all subsequently published work among its set of predicted novel ideas, as peer reviewed publication is the strongest available signifier of idea quality in the sciences. This backtesting based evaluation is contamination-free by design, imposes no constraints on agent architecture, and does not require

[1]github.com/autonomousvision/horizon-bench
[2]scholar-inbox.com/horizon-bench

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 4466–4479
November 6-10, 2026 ©2026 Association for Computational Linguistics

expensive human labeling.

We introduce **HorizonBench** to test an agent's ability to generate novel challenges and methods given a research goal. Framing the right **challenge** is what separates productive inquiry from speculation (Buchanan, 1992), while devising a rigorous **method** is what turns a hypothesis into reliable knowledge (Platt, 1964). Together, these provide a standardised format for research ideas. We normalise both predictions and ground truth through a shared LLM prompt into this format of a challenge and method, and measure prediction quality with an embedding-based metric (Soft Recall) that rewards producing at least one strong idea. We evaluate agents on ICLR 2026 papers, the most recent top-tier conference proceedings available at the time of writing, excluding research papers with preprints released before the HorizonBench knowledge cutoff of January 2025. To prevent HorizonBench from becoming outdated, we release it as a continuously updated service, generating fresh test sets with moving knowledge cutoff dates as new conference proceedings become available.

Using HorizonBench, we conduct the first unified comparison of five published ideation agents across three frontier model backbones. This evaluation reveals a surprising finding: **a naïve baseline that receives only the research goal, without any retrieval or multi-step reasoning, consistently matches or outperforms all retrieval-augmented agents**. We hypothesise that inaccurate retrieved context degrades output quality by pulling generation away from the research goal. Current ideation agents lack this goal-following ability.

Informed by these insights, we develop **OSCAR** as a secondary contribution of our paper. OSCAR is an ideation agent designed around two principles: (1) to build a context relevant for ideation, related work should be distilled into goal-conditioned insights rather than used raw, and (2) the research goal must be explicitly reinforced throughout the pipeline. OSCAR extracts such goal-conditioned insights from retrieved papers, and conditions final idea generation on these insights and the original goal. Figure 1(a) presents qualitative examples from different ideation agents, showing that OSCAR aligns more closely with the research goal. Figures 1(b–c) further demonstrate the advantages of OSCAR over existing agents in both automatic and human evaluations.

The main contributions of this paper are:

1. **HorizonBench.** A backtesting evaluation framework for scientific ideation agents that uses temporal knowledge cutoffs, published papers as ground truth, and a new embedding based scoring metric, released as a continuously updated service.

2. **Human validation.** A ranking-based annotation study (with 318 ratings from 51 domain experts) with personalised sample assignment. Human rankings closely mirror automated HorizonBench rankings (0.86 Spearman correlation).

3. **Finding: ideation agents lack goal-following ability.** A naïve baseline matches or outperforms existing retrieval-augmented agents, revealing that their retrieved context actively degrades quality by distracting from the research goal. Inspired by this finding, we explore a simple design principle using goal-conditioned insights named OSCAR, outperforming existing agents.

We will publicly release the code for HorizonBench and OSCAR, along with the training and anonymised human evaluation datasets, to facilitate reproducibility and future research.

2 Related Work

2.1 Scientific ideation agents

Seminal work. Before LLMs, AI-driven ideation was domain-specific and built on symbolic reasoning. The robot scientist **Adam** (King et al., 2009; King, 2017) was the first AI system to autonomously discover new scientific knowledge, using abductive logic programming to hypothesise gene functions in yeast, design and execute experiments, and update its models from results. Systems relied on hand-crafted ontologies, whereas modern ideation agents use probabilistic inference paired with context retrieval over learned representations.

LLM-based ideation agents. Recently, the prospect of automating scientific discovery has driven a new wave of LLM-based ideation systems. As the first LLM agent to successfully publish a workshop paper, the AI Scientist (Lu et al., 2024) uses automated research, generating ideas, writing code, running experiments, and producing complete manuscripts. Its successor, AI Scientist v2 (Yamada et al., 2025), removes the dependence on human-authored code templates and introduces

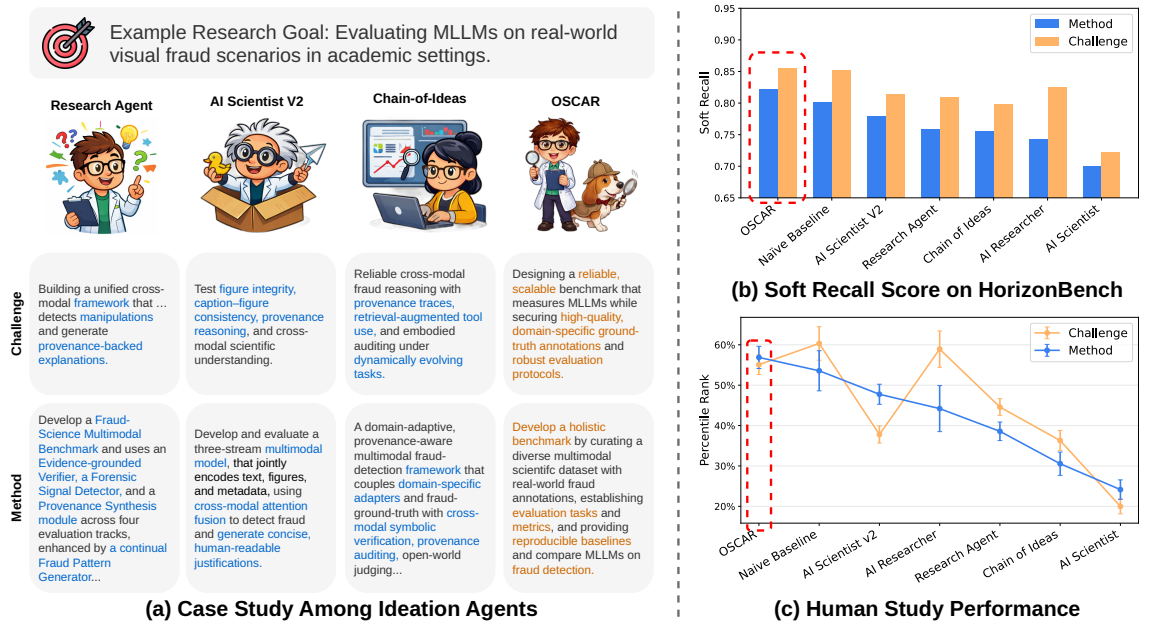


Figure 1: (a) Comparison of existing ideation agents, highlighting that OSCAR adheres more closely to the research goals, while other agents often fail to follow them. Both our proposed automatic (b) and human evaluations (c) demonstrate the effectiveness of OSCAR.

agentic tree search. These were followed by ResearchAgent (Baek et al., 2025), which augments idea generation with academic graph retrieval and the Chain of Ideas agent (Li et al., 2025) which organises retrieved literature into a chain structure that mirrors the progressive development of research. AI-Researcher (Tang et al., 2025) orchestrates the full research pipeline from literature review to manuscript preparation. In our paper, we evaluate all five of these agents under a unified protocol for the first time.

2.2 Evaluating scientific ideation

Scientific agent evaluation. FIRE-Bench (Wang et al., 2026) tasks agents with rediscovering established findings through autonomous experimentation, finding that 73.6% of errors occur during research planning. This motivates our focus on the ideation stage. Several benchmarks also target scientific reproducibility and code reproduction (Siegel et al., 2024; Starace et al., 2025; Xiang et al., 2025; Yan et al., 2025; Kon et al., 2025). These primarily assess **execution** capabilities, whereas HorizonBench targets the upstream **ideation** stage that precedes and informs all downstream research activities.

LLM-as-a-judge and review-based evaluation. Most existing work directly assesses idea quality using LLM judges. AI Researcher (Tang et al., 2025) evaluates generated papers using its own Scientist-Bench with automated review. ResearchBench (Liu et al., 2025) decomposes scientific discovery into sub-tasks and evaluates LLMs across multiple disciplines. Other approaches train dedicated reviewer models (Zhu et al., 2025), ground critiques in web-scale literature through multi-agent frameworks (Goyal et al., 2026), or retrieval-augmented assessment (Moussa et al., 2026). IdeaBench (Guo et al., 2024) and related follow-up work (Qiu et al., 2025; Jiang, 2026) use LLM-as-a-judge protocols while adopting backtesting and a knowledge cutoff like HorizonBench. However, unlike HorizonBench, these approaches are not validated against human judgement (Section 4). In fact, LLM judges have been found to correlate negatively with actual research impact (Jiang, 2026). Furthermore, IdeaBench evaluates ideas at the abstract level, while HorizonBench decomposes them into challenge and method components, following recommendations from scientific ideation theory (Buchanan, 1992; Platt, 1964). This prevents surface-level biases such as length, tone, or formatting from skewing the evaluation results.

3 HorizonBench

We introduce HorizonBench, a backtesting evaluation framework for scientific ideation agents. The central insight is that published research provides a natural ground truth for ideation: if an agent, con-

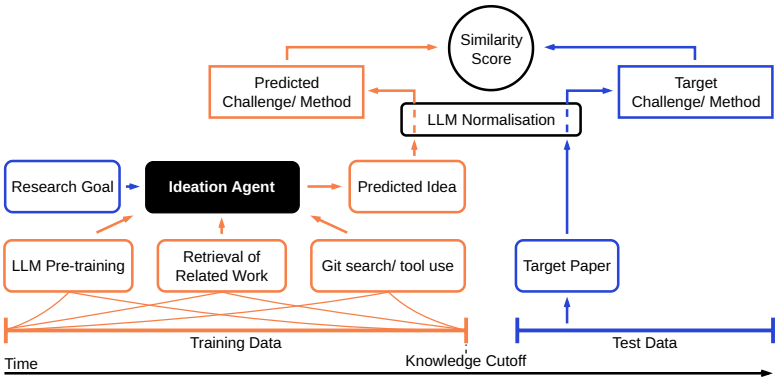


Figure 2: **HorizonBench.** Data is split at the foundation model's knowledge cutoff. The ideation agent operates exclusively on data preceding the cutoff and produces a predicted idea, which is normalised alongside the ground truth target into a challenge and method. Finally, we compute the cosine similarity between the GTE embeddings of the prediction and ground truth target. HorizonBench can evaluate any ideation agent, as long as it obeys the knowledge cutoff.

strained to only know about the past, can predict ideas that were independently discovered and published, it demonstrates genuine ideation capability. By enforcing a temporal knowledge cutoff on training data as well as tool use, we obtain a rigorous, contamination-free evaluation without requiring expensive human judgement.

3.1 Evaluation framework

Figure 2 illustrates the HorizonBench pipeline. Given a foundation model with a known knowledge cutoff date, we partition the corpus of published papers into a **training & retrieval set** (papers published before the cutoff) and a **test set** (papers published after). The ideation agent under evaluation may be pre-trained on the training data and may employ any form of retrieval, literature search, tool use, or web access, as long as it strictly adheres to the temporal cutoff. This black-box design imposes no constraints on the agent's internal architecture, only on the information it is permitted to access.

For evaluation, we select a high-quality target paper from the test set. The agent receives a research goal description and produces a predicted idea. Both the agent's prediction and the ground truth extracted from the target paper are then passed through a normalisation step (Section 3.2). After this, we compute the similarity between the predicted and target challenge and method by embedding each with a transformer encoder (GTE, (Li et al., 2023)) and computing their cosine similarity.

We evaluate on the top 100 ICLR 2026 papers, based on review score, using LLM backbones with knowledge cutoffs leading up to January 2025 and we reimplement the idea agents' context retrieval code to adhere to this cutoff date.

3.2 Normalisation

A known failure mode of LLM-based evaluation is sensitivity to surface-level features of the response: longer outputs, more confident phrasing, self-enhancement, positioning, emotional sentiment or particular formatting choices can systematically bias both automated metrics and human raters (Ye et al., 2025). To guard against such effects, we pass both the agent's prediction and the ground truth extraction through the same normalisation prompt before comparison. A single LLM reformats both texts into a standardised structure, stripping away differences in length, phrasing, and style.

3.3 Evaluation metric

Ideation is inherently an ambiguous process, as for each research goal there may be multiple valid ideas. But there is only one target paper for each goal, which motivates **Soft Recall** as our primary evaluation metric. Given a test paper i with ground truth target t_i and a set of k predicted ideas $P_i = \{p_i^1, \ldots, p_i^k\}$ from an ideation agent, we define:

$$\text{Soft Recall} = \frac{1}{N} \sum_{i=1}^{N} \max_{j \in \{1,\ldots,k\}} \text{sim}(p_i^j,\ t_i) \quad (1)$$

where N is the number of test papers and sim denotes the cosine similarity between the GTE embeddings of the prediction and ground truth target. For each target, Soft Recall retains only the best-matching prediction, rewarding agents that produce at least one high-quality idea rather than optimising

for the average case. For reference, we also report the average (instead of maximum) similarity and error bars across all experiments.

3.4 Baselines

We implement and evaluate five recently published ideation agents on HorizonBench: AI Scientist (Lu et al., 2024), AI Scientist v2 (Yamada et al., 2025), AI Researcher (Tang et al., 2025), Research Agent (Baek et al., 2025), and the Chain of Ideas (COI) agent (Li et al., 2025). AI Scientist does not support an explicit goal input, but instead expects a starting codebase. In this case we prompt the backbone LLM to generate the required code from the goal description, ensuring a fair and uniform evaluation across all agents.

In addition, we include a **naïve baseline** that receives only the goal description and directly predicts the idea without any retrieval or multi-step reasoning. This baseline isolates the contribution of the foundation model's parametric knowledge from the agent scaffolding built on top of it, and reveals a key insight that current research ideation agents lack research goal following capabilities. Results and analysis are presented in Section 5.

3.5 Withstanding the test of time

A common criticism of history-based benchmarks is that they become stale. A benchmark constructed from ICLR 2026 papers will be obsolete within a year, as state-of-the-art foundation models start to incorporate these publications into their training data. To address this, we release HorizonBench not only as a static dataset but as a continuously updated service. We scrape the proceedings of the top machine learning venues including NeurIPS, CVPR, ACL, ICLR, and COLM and will incorporate all future publications from these conferences as they become available. Researchers then simply specify the desired knowledge cutoff of the ideation agent they wish to evaluate and a publication venue, and our platform generates a suitable test set containing only papers published after that date. This design ensures that HorizonBench remains a valid evaluation benchmark and minimises the effort for future research with HorizonBench.

3.6 OSCAR

The evaluation results on HorizonBench (Section 5) reveal two key findings that motivate the design of OSCAR, our ideation agent. First, **existing research agents lack a goal-following ability**. Their multi-step retrieval and tool-use pipelines progressively lose sight of the original research goal as they iterate. Prior evaluation frameworks such as IdeaBench (Guo et al., 2024) do not provide a research goal, which obscures whether the agent stayed on task during ideation. HorizonBench is able to expose this behaviour because it provides concrete target ideas against which goal adherence can be measured. Second, while existing agents retrieve related work, mirroring the human research process, they do so without a quantitative analysis on whether the retrieved material actually helps. In creating OSCAR, we were able to test different structured contexts (use retrieved papers or abstracts directly, using extracted goals and methods) and verify them against a validation set.

As shown in Figure 3, OSCAR proceeds in four stages:

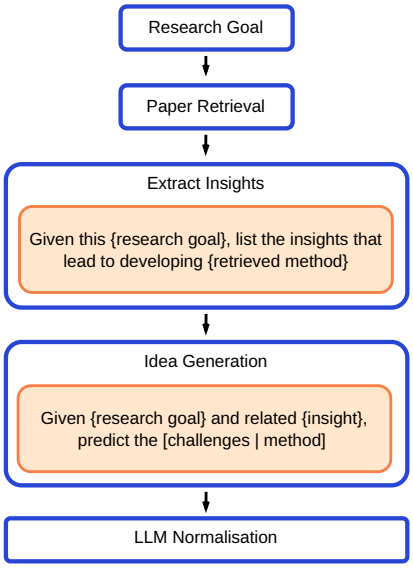


Figure 3: **OSCAR**, our ideation agent built on the insights gained from HorizonBench.

Query rewriting. The backbone LLM rewrites the given goal g into a set of keyword search queries to retrieve papers whose methods or findings could inform a solution to the goal.

Paper retrieval. Each query is issued to the Semantic Scholar API to retrieve the top three relevant papers. For each retrieved paper, we extract its method using the same structured extraction functions in Section 3. **Insight extraction.** For every retrieved paper, we prompt the LLM to list the insights that lead to developing the method, given the target research goal. We find that aligning the

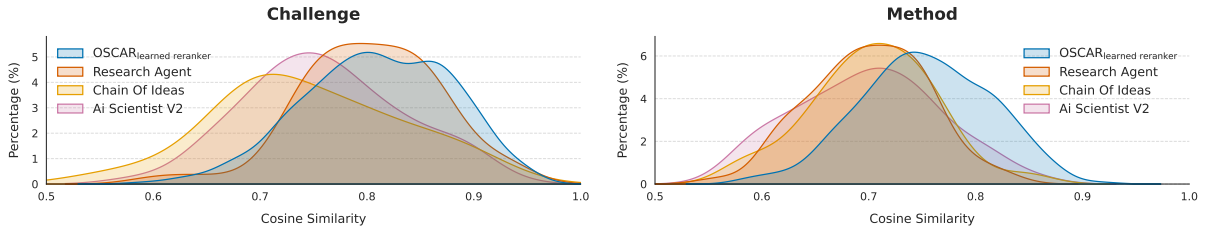


Figure 4: OSCAR outperforms other ideation agents on HorizonBench, with significant improvements stemming from the goal-conditioned ideation process using insights of related works.

insight extraction prompt with the target research goal, rather than the retrieved paper's own goal, yields a 1–2 percentage point improvement in the final output's similarity to the ground truth, confirming that goal conditioning is beneficial even during intermediate extraction steps. This step distils each retrieved paper into a set of actionable insights grounded in the retrieved work and the target research goal.

Idea generation. OSCAR samples from the insights and generates the final challenge and method predictions, conditioned on the original research goal g. By re-introducing g at this final stage, the pipeline explicitly reinforces goal following implementing the key insight gathered from HorizonBench.

4 Human Evaluation of the Benchmark Quality

To assess the perceived quality of ideation outputs across agents, we conduct a human evaluation study with well-defined criteria and guidelines (see Appendix A.7 for details).

4.1 Human study design

Our evaluation follows the ranking-based annotation protocol introduced by Ouyang et al. (2022). Rather than collecting pairwise preference judgements, we present annotators with the outputs of all six agents described in Section 3.4 for a given goal simultaneously and ask them to rank the outputs from best to worst by dragging and dropping items into their preferred ordering (see Figure 5 in the Appendix A.7 for a screenshot of the interface). This approach is substantially more efficient than binary comparisons, because ranking six agents requires a single annotation, whereas exhaustive pairwise comparisons would require $\binom{6}{2} = 15$ judgements per sample. The study is conducted blind, with annotators unaware of the producing agent.

A distinguishing feature of our study is its targeted, personalised sample assignment. We leverage a personalised research paper recommender that learns user interests from their interactions, and tailor each evaluation sample directly to the annotator's research expertise, ensuring that participants are domain experts for the samples they evaluate. To further ensure annotation quality, we exclude users whose predicted research interest is dissimilar to the ICLR 2026 target papers in our study, and we ask each annotator to rate their confidence in their rankings. Participants agreed to anonymous data collection as part of the study.

4.2 Human study results

We collected 318 ratings from 51 PhD level domain experts. Annotators show high inter-annotator agreement, as reflected by the small 2σ error bars in Figure 1(c).

The human rankings for the "method" task are, from best to worst: (1) OSCAR (2) Naïve Baseline, (3) AI Scientist v2, (4) AI Researcher, (5) Research Agent, (6) Chain of Ideas, and (7) AI Scientist. This ordering shows strong agreement (0.86 Spearman correlation) with the automated HorizonBench results (Table 4), where the same six agents are ranked identically by challenge Soft Recall. The agreement is particularly notable at the extremes: both human annotators and the automated metric place the naïve baseline first and AI Scientist last, corroborating the finding that current retrieval-augmented ideation agents struggle to outperform simple goal-conditioned generation and that AI Scientist produces substantially lower-quality outputs. The close correspondence between human judgement and our embedding-based metric provides evidence that HorizonBench captures meaningful aspects of ideation quality.

5 Results

We evaluate OSCAR and five published ideation agents on HorizonBench across three frontier model backbones. We report Soft Recall (Equation 1) and, where applicable, mean similarity (in-

	Challenge		Method	
	Soft Recall	Mean Sim.	Soft Recall	Mean Sim.
OSCAR $_{\text{learned reranker}}$	**0.865** $\pm$ 0.05	**0.810** $\pm$ 0.05	**0.807** $\pm$ 0.05	0.753 $\pm$ 0.05
Naïve Baseline	0.798 $\pm$ 0.05	0.761 $\pm$ 0.05	0.805 $\pm$ 0.05	**0.759** $\pm$ 0.05
AI Scientist v2	0.791 $\pm$ 0.07	0.758 $\pm$ 0.06	0.774 $\pm$ 0.05	0.700 $\pm$ 0.05
Research Agent	0.821 $\pm$ 0.06	0.793 $\pm$ 0.06	0.770 $\pm$ 0.04	0.700 $\pm$ 0.03
Chain of Ideas	0.764 $\pm$ 0.08	0.753 $\pm$ 0.08	0.705 $\pm$ 0.06	0.703 $\pm$ 0.06
AI Scientist	0.761 $\pm$ 0.05	0.749 $\pm$ 0.04	0.699 $\pm$ 0.05	0.704 $\pm$ 0.04

Table 1: Comparison of idea generation agents on the challenge and method prediction tasks. Error bars are $\pm$ 1 standard deviation.

stead of the maximum over k ideas used in Soft Recall). All results use cosine similarity of GTE embeddings (Li et al., 2023) between the prediction and ground truth.

OSCAR outperforms existing ideation agents. Figure 4 shows histograms of per-paper similarity scores for the method and challenge tasks. OSCAR's distribution is shifted further to the right compared to all other agents, confirming that OSCAR produces higher-quality predictions. Because Soft Recall takes the maximum similarity across an agent's outputs for each goal, it rewards agents that generate at least one strong idea per paper. We ran each agent for ten inference runs. Under this metric, OSCAR outperforms all other ideation agents significantly on the method prediction task and outperforms the naïve baseline, with improvements coming from accurately using insights extracted from related works.

Qualitative analysis. As shown in Figure 1 (b), OSCAR stands out as the strongest ideation agent primarily due to its close alignment with the research goal and its higher feasibility. Intuitively, as shown in the case of Figure 1 (a), unlike other agents that shift focus toward model design or intermediate reasoning components, OSCAR directly addresses the core objective of the given task. This alignment ensures that proposed dataset construction, annotation protocols and evaluation metrics all contribute directly to the end goal without introducing unnecessary complexity. In addition, OSCAR is more feasible because it builds on straightforward method with essential components, but avoids overly ambitious or technically uncertain designs.

5.1 Ablation study

To isolate the contribution of each component, we evaluate several variants of OSCAR (Table 2): **OSCAR $_{\textbf{no reranker}}$** generates a separate method prediction from every extracted insight without any reranking or selection. This produces many outputs, some of which are excellent, maximising Soft Recall at the cost of average performance. **OSCAR $_{\textbf{random}}$** selects insights uniformly at random, providing a lower bound on reranking quality. **OSCAR $_{\textbf{LLM}}$** uses the backbone LLM to rerank insights in a zero-shot fashion, without any learned scoring. **OSCAR $_{\textbf{learned reranker}}$** uses the trained reranker from Section A.2, with training data generated by running OSCAR $_{\text{no reranker}}$ on a training dataset. **OSCAR $_{\textbf{cheating reranker}}$** selects retrieved insights that are most semantically similar to the ground-truth insights, ablating both mean and max similarity. This variant measures how well semantic similarity proxies for true insight utility. **OSCAR $_{\textbf{gt insights}}$** is an oracle setting in which insights are extracted from the ground-truth method rather than from retrieved papers. This upper-bounds the contribution of retrieval and isolates the quality of the idea generation stage.

Table 2 reports results on the method task. The oracle setting (OSCAR $_{\text{gt insights}}$), which uses insights extracted from the ground-truth method, achieves the highest Soft Recall (0.848) and average similarity (0.792), confirming that high-quality insights are the primary driver of idea generation quality. OSCAR $_{\text{no reranker}}$ achieves a Soft Recall of 0.837, outperforming the naïve baseline (0.829) while also incorporating retrieved content. Notably OSCAR $_{\text{no reranker}}$ doesn't beat the naïve baseline on average similarity which suggests room for a reranking component to improve average perfor-

	Challenge		Method	
	Soft Recall	Mean Sim.	Soft Recall	Mean Sim.
OSCAR $_{\textit{gt insights}}$	*0.867* $\pm$ *0.04*	*0.817* $\pm$ *0.04*	*0.852* $\pm$ *0.04*	*0.800* $\pm$ *0.04*
OSCAR $_{\textit{cheating reranker}}$	*0.863* $\pm$ *0.04*	*0.811* $\pm$ *0.05*	*0.839* $\pm$ *0.05*	*0.783* $\pm$ *0.04*
OSCAR $_{\text{learned reranker}}$	**0.865** $\pm$ 0.05	**0.810** $\pm$ 0.05	**0.807** $\pm$ 0.05	**0.753** $\pm$ 0.05
OSCAR $_{\text{random}}$	0.865 $\pm$ 0.05	0.806 $\pm$ 0.04	0.806 $\pm$ 0.05	0.731 $\pm$ 0.05
OSCAR $_{\text{no reranker}}$	0.862 $\pm$ 0.05	0.801 $\pm$ 0.05	0.804 $\pm$ 0.05	0.748 $\pm$ 0.05
OSCAR $_{\text{LLM}}$	0.861 $\pm$ 0.05	0.807 $\pm$ 0.04	0.803 $\pm$ 0.05	0.739 $\pm$ 0.05

Table 2: Ablations of OSCAR on the method and challenge tasks.

mance.

A key finding is that **current scientific idea agents are poor at goal following**, and inaccurate retrieved context easily distracts them. All models that incorporate retrieved context into the generation, including several OSCAR ablations, underperform the naïve baseline on average similarity. The naïve baseline, which receives only the goal description without any retrieval, achieves an average similarity of 0.778, higher than OSCAR $_{\text{no reranker}}$ (0.770), OSCAR $_{\text{cheating reranker}}$ (0.771), OSCAR $_{\text{LLM}}$ (0.757), and OSCAR $_{\text{random}}$ (0.754). This indicates that when retrieved material is not precisely relevant, it actively degrades output quality by pulling the generator away from the research goal.

The LLM-as-a-reranker approach (OSCAR $_{\text{LLM}}$) performs poorly, and does not beat the random reranker baseline, suggesting that current LLMs lack the ability to reliably judge which insights will be most useful for idea generation without task-specific training.

5.2 Comparison across frontier models

Tables 3, 4, and 5 in the Appendix A.1 report Soft Recall across three frontier model backbones: GPT5-nano, GPT5, and Gemini Flash Lite Preview.

Across all three backbones, the naïve baseline is consistently among the top performers, ranking first on both challenge and method with GPT5-nano (0.853 and 0.802) and GPT5 (0.782 and 0.759), and first on method with Gemini Flash Lite (0.740). This reinforces the finding from the ablation study: retrieval-augmented agents frequently underperform a simple goal-conditioned generation, because inaccurate context degrades goal following.

AI Scientist consistently ranks last or near-last across all backbone models and tasks. With GPT5-nano, it achieves only 0.723 on challenge and 0.701 on method, far below all other agents. With GPT5, the gap persists (0.699 challenge, 0.679 method). AI Scientist v2 addresses these issues significantly, improving to 0.815 and 0.780 with GPT5-nano, and 0.767 and 0.719 with GPT5. This is corroborated by the human evaluation (Figure 1(c)), which also places AI Scientist's generations in last place. We observe that AI Scientist's poor performance on the challenge task stems largely from rarely outputting challenge descriptions at all, while on the method task its generations are simply lower quality.

6 Conclusion

We introduced HorizonBench, a backtesting evaluation framework that uses published papers as ground truth to assess scientific ideation agents. By enforcing temporal knowledge cutoffs and measuring alignment between predicted and published ideas, HorizonBench provides a rigorous, contamination-free, and scalable alternative to LLM-as-a-judge and human evaluation. A ranking-based human study with 318 ratings from 51 domain experts confirms that HorizonBench captures meaningful ideation quality, achieving a 0.86 Spearman correlation with the automated evaluation.

Using HorizonBench, we conducted the first unified comparison of five published ideation agents across multiple frontier model backbones found that a naïve baseline that receives only the research goal consistently matches or outperforms all retrieval-augmented agents. This reveals that current ideation agents lack a goal-following ability and we developed OSCAR, an ideation agent that distils retrieved papers into goal-conditioned insights and explicitly reinforces the research goal throughout its pipeline. OSCAR outperforms all evaluated agents on HorizonBench and in human evaluation.

7 Limitations

HorizonBench currently evaluates ideation on top-tier ML venues and measures alignment with a single published paper per goal. Extending to other scientific domains and capturing the multiplicity of valid ideas remain open directions. Additionally, the backtesting methodology inherently measures an agent's ability to anticipate ideas that were published, which may not fully reflect the value of ideas that are novel but have not yet appeared in the literature.

Task definition. We frame ideation evaluation as a soft recall problem against a single ground truth paper per goal. In practice, ideation is inherently a one-to-many process: for a given research goal there may be many distinct, equally valid ideas, only one of which corresponds to the published paper. Soft Recall mitigates this by rewarding the best of k predictions, but it cannot credit valuable ideas that diverge from the published target. Future work could extend HorizonBench to multi-reference targets, for example by aggregating ideas across multiple papers that share a common goal.

Scope of evaluated agents. We evaluate five published ideation agents alongside OSCAR and a naïve baseline. While these agents cover the most prominent end-to-end ideation systems at the time of writing, the broader ideation literature includes complementary design choices (multi-agent architectures, bandit-style exploration, Bayesian search, preference-tuned ideation, etc.) that we do not benchmark.

Domain and venue coverage. All test papers in this work are drawn from top-tier machine learning venues (with the headline evaluation on ICLR 2026), and all goals and ideas are expressed in English. Our results therefore exhibit limited generalisability to other scientific fields, such as the natural and life sciences, where ideation conventions, the role of literature, and acceptable framings of a research goal can differ substantially. Extending HorizonBench beyond ML to broader scientific domains remains an open direction.

Human evaluation. Our human study uses ranking-based annotations from 51 PhD-level domain experts, with personalised sample assignment via a recommender system. This design biases the participant pool towards researchers reachable through the recommender platform and may not be representative of the broader ML research community. Ranking is also a relative judgement and does not capture absolute idea quality: if all six agents produce weak ideas for a given goal, the ranking still produces a winner.

References

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6709–6738, Albuquerque, New Mexico. Association for Computational Linguistics.

Richard Buchanan. 1992. Wicked Problems in Design Thinking. *Design Issues*, 8(2):5.

Palash Goyal, Mihir Parmar, Yiwen Song, Hamid Palangi, Tomas Pfister, and Jinsung Yoon. 2026. ScholarPeer: A Context-Aware Multi-Agent Framework for Automated Peer Review. *arXiv preprint*. ArXiv:2601.22638 [cs].

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Eric Xie, Stefan Bekiranov, and Aidong Zhang. 2024. IdeaBench: Benchmarking Large Language Models for Research Idea Generation. *arXiv preprint*. ArXiv:2411.02429 [cs].

Bo Jiang. 2026. HindSight: Evaluating LLM-Generated Research Ideas via Future Impact. *arXiv preprint*. ArXiv:2603.15164 [cs].

Ross King. 2017. The Adam and Eve Robot Scientists for the Automated Discovery of Scientific Knowledge. volume 2017, page X49.001. ADS Bibcode: 2017APS..MARX49001K.

Ross D. King, Jem Rowland, Wayne Aubrey, Maria Liakata, Magdalena Markham, Larisa N. Soldatova, Ken E. Whelan, Amanda Clare, Mike Young, Andrew Sparkes, Stephen G. Oliver, and Pinar Pir. 2009. The Robot Scientist Adam. *Computer*, 42(8):46–54.

Patrick Tser Jern Kon, Jiachen Liu, Xinyi Zhu, Qiuyi Ding, Jingjia Peng, Jiarong Xing, Yibo Huang, Yiming Qiu, Jayanth Srinivasa, Myungjin Lee, Mosharaf Chowdhury, Matei Zaharia, and Ang Chen. 2025. EXP-Bench: Can AI Conduct AI Research Experiments? *arXiv preprint*. ArXiv:2505.24785 [cs].

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Yu Rong, Deli Zhao, Tian Feng, and Lidong Bing. 2025. Chain of Ideas: Revolutionizing Research Via Novel Idea Development with LLM Agents. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8971–9004, Suzhou, China. Association for Computational Linguistics.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards General Text Embeddings with Multi-stage Contrastive Learning. *arXiv preprint*. ArXiv:2308.03281 [cs].

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. 2025. ResearchBench: Benchmarking LLMs in Scientific Discovery via Inspiration-Based Task Decomposition. *arXiv preprint*. ArXiv:2503.21248 [cs].

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *arXiv preprint*. ArXiv:2408.06292 [cs].

Hanane Nour Moussa, Patrick Queiroz Da Silva, Daniel Adu-Ampratwum, Alyson East, Zitong Lu, Nikki Puccetti, Mingyi Xue, Huan Sun, Bodhisattwa Prasad Majumder, and Sachin Kumar. 2026. ScholarEval: Research Idea Evaluation Grounded in Literature. *arXiv preprint*. ArXiv:2510.16234 [cs].

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Changhua Pei, Yi Zhang, Yongfeng Zhang, Fei Sun, Xiao Lin, Hanxiao Sun, Jian Wu, Peng Jiang, Junfeng Ge, Wenwu Ou, and Dan Pei. 2019. Personalized re-ranking for recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, pages 3–11, New York, NY, USA. Association for Computing Machinery.

John R. Platt. 1964. Strong Inference. *Science*, 146(3642):347–353.

Yansheng Qiu, Haoquan Zhang, Zhaopan Xu, Ming Li, Diping Song, Zheng Wang, and Kaipeng Zhang. 2025. AI Idea Bench 2025: AI Research Idea Generation Benchmark. *arXiv preprint*. ArXiv:2504.14191 [cs].

Aishik Sanyal, Samuel Schapiro, Sumuk Shashidhar, Royce Moon, Lav R. Varshney, and Dilek Hakkani-Tur. 2025. Spark: A System for Scientifically Creative Idea Generation. *arXiv preprint*. ArXiv:2504.20090 [cs].

Zachary S. Siegel, Sayash Kapoor, Nitya Nagdir, Benedikt Stroebl, and Arvind Narayanan. 2024. CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark. *arXiv preprint*. ArXiv:2409.11363 [cs].

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. 2025. PaperBench: Evaluating AI's Ability to Replicate AI Research. *arXiv preprint*. ArXiv:2504.01848 [cs].

Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. 2025. AI-Researcher: Autonomous Scientific Innovation. *CoRR*, abs/2505.18705. ArXiv: 2505.18705.

Jingqi Tong, Mingzhe Li, Hangcheng Li, Yongzhuo Yang, Yurong Mou, Weijie Ma, Zhiheng Xi, Hongji Chen, Xiaoran Liu, Qinyuan Cheng, Ming Zhang, Qiguang Chen, Weifeng Ge, Qipeng Guo, Tianlei Ying, Tianxiang Sun, Yining Zheng, Xinchi Chen, Jun Zhao, and 4 others. 2026. AI Can Learn Scientific Taste. *arXiv preprint*. ArXiv:2603.14473 [cs] version: 1.

Zhen Wang, Fan Bai, Zhongyan Luo, Jinyan Su, Kaiser Sun, Xinle Yu, Jieyuan Liu, Kun Zhou, Claire Cardie, Mark Dredze, Eric P. Xing, and Zhiting Hu. 2026. FIRE-Bench: Evaluating Agents on the Rediscovery of Scientific Insights. *arXiv preprint*. Version Number: 1.

Yanzheng Xiang, Hanqi Yan, Shuyin Ouyang, Lin Gui, and Yulan He. 2025. SciReplicate-Bench: Benchmarking LLMs in Agent-driven Algorithmic Reproduction from Research Papers. *arXiv preprint*. ArXiv:2504.00255 [cs].

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search. *arXiv preprint*. ArXiv:2504.08066 [cs].

Shuo Yan, Ruochen Li, Ziming Luo, Zimu Wang, Daoyang Li, Liqiang Jing, Kaiyu He, Peilin Wu, George Michalopoulos, Yue Zhang, Ziyang Zhang, Mian Zhang, Zhiyu Chen, and Xinya Du. 2025. LMR-BENCH: Evaluating LLM Agent's Ability on Reproducing Language Modeling Research. *arXiv preprint*. ArXiv:2506.17335 [cs].

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2025. Justice or prejudice? quantifying biases in llm-as-a-judge. In *International Conference on Learning Representations (ICLR)*.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. DeepReview: Improving LLM-based Paper Review with Human-like Deep Thinking Process. *arXiv preprint*. ArXiv:2503.08569 [cs].

A Appendix

A.1 Results across frontier models

Agent	Challenge	Method
Naïve Baseline	**0.782** ± 0.127	**0.759** ± 0.148
Research Agent	0.771 ± 0.118	0.708 ± 0.159
AI Scientist v2	0.767 ± 0.141	0.719 ± 0.155
AI Scientist	0.699 ± 0.149	0.679 ± 0.133

Table 3: Soft Recall ($\pm 2\sigma$) results using GPT5

Agent	Challenge	Method	Research Question
Naïve Baseline	**0.853** ± 0.04	**0.802** ± 0.04	**0.897** ± 0.04
AI Researcher	0.826 ± 0.05	0.743 ± 0.05	0.860 ± 0.05
AI Scientist v2	0.815 ± 0.06	0.780 ± 0.06	0.842 ± 0.06
Research Agent	0.810 ± 0.05	0.759 ± 0.05	0.843 ± 0.05
Chain of Ideas	0.799 ± 0.04	0.755 ± 0.04	0.819 ± 0.05
AI Scientist	0.723 ± 0.08	0.701 ± 0.07	0.694 ± 0.08

Table 4: Soft Recall results on HorizonBench using GPT5-nano as the goal proposer, for the backbone of the ideation agents, and as the evaluator.

Agent	Challenge	Method
AI Scientist v2	**0.796** ± 0.118	0.700 ± 0.113
Research Agent	0.780 ± 0.122	0.685 ± 0.123
Naïve Baseline	0.779 ± 0.105	**0.740** ± 0.114
AI Scientist	0.741 ± 0.124	0.690 ± 0.124

Table 5: Soft Recall ($\pm 2\sigma$) results using Gemini Flash Lite Preview

A.2 Learned reranker

For sampling insights that yield better idea generation, we train a *pointwise* reranker. Given a candidate set of insights $S = \{s_j\}_{j=1}^{m}$ for goal g, this reranker predicts a utility score $\hat{u}_j \in [0, 1]$ for each goal–insight pair (g, s_j) and sorts insights by decreasing $\hat{u}_j$.

Data construction To supervise the reranker, we collect a dataset of 3591 insights by running OSCAR, initially without reranking (variant OSCAR $_{\text{no reranker}}$, Section 5.1), to produce ground truth utility scores by evaluating the generated methods and challenges following the same evaluation protocol as defined in Equation 1. The resulting data are stored goal-wise as tuples of one goal with aligned insight and score arrays, $(g, \{s_j\}_{j=1}^{m}, \{u_j\}_{j=1}^{m})$, and then flattened into pointwise training triples (g, s_j, u_j).

Training We train the reranker on pointwise triples (g, s_j, u_j) as a weighted regression problem over utility scores. Concretely, a parametric scoring function f_θ maps each goal–insight pair to a scalar prediction, $\hat{u}_j = f_\theta(g, s_j)$. Let $u_j^\star$ denote the training target after normalisation (see Appendix A.3). We optimise:

$$\mathcal{L}_{\text{rerank}} = \frac{\sum_j \omega_j \big(\hat{u}_j - u_j^\star\big)^2}{\sum_j \omega_j}, \tag{2}$$

where ω_j is a target-dependent weight that emphasises high-utility insights.

At inference time, we score every extracted insight independently with the trained pointwise reranker and sort insights by decreasing the predicted utility scores. Then OSCAR passes top-ranked insights to the idea generator.

A.3 Pointwise reranker implementation details

Per-goal target normalisation. Following our implementation, utility labels are optionally min–max normalised *within each goal*:

$$u_j^\star = \begin{cases} \dfrac{u_j - \min_k u_k}{\max_k u_k - \min_k u_k}, & \max_k u_k > \min_k u_k, \\ 0.5, & \text{otherwise}, \end{cases} \tag{3}$$

which keeps targets in $[0, 1]$ while preserving relative ordering among insights for the same goal.

Target-dependent weighting. For the weighted regression objective in Section A.2, we use

$$\omega_j = \begin{cases} \beta, & u_j^\star \geq \tau, \\ 1, & \text{otherwise}, \end{cases} \tag{4}$$

with τ=0.7 and β=2.0 by default, so that high-utility insights receive larger training weight.

Score instantiation and inference. Section A.2 defines the reranker abstractly as $\hat{u}_j = f_\theta(g, s_j)$. In our concrete instantiation, we use an instruction-style input with fields `<Instruct>`, `<Query>` and `<Document>`, and compute

$$\hat{u}_j = P_\theta(\texttt{yes} \mid g, s_j) = \frac{\exp(\ell_{\texttt{yes}})}{\exp(\ell_{\texttt{yes}}) + \exp(\ell_{\texttt{no}})}, \tag{5}$$

where $\ell_{\texttt{yes}}$ and $\ell_{\texttt{no}}$ are the logits of the `"yes"`/`"no"` tokens at the scoring position. At inference time, insights are ranked in descending order of $\hat{u}_j$.

Concrete training configuration. In our experiments, the reranker is instantiated with `Qwen/Qwen3-Reranker-8B`. We optimise with AdamW, linear warmup/decay scheduling, and gradient clipping, and choose checkpoints by validation NDCG@5.

Optional parameter-efficient tuning (LoRA). Using LoRA is optional in our implementation. By default, we enable LoRA adapters on `q_proj`, `k_proj`, `v_proj`, and `o_proj` (rank r=64, α=128, dropout 0.1), with a command-line switch to disable LoRA and fine-tune without adapters. This choice does not change the reranking objective or inference procedure described in Section A.2.

We explore two reranker architectures:

Encoder-based reranker. Following Pei et al. (2019), we encode the research goal and each insight with a transformer encoder and predict utility with a multilayer perceptron over the concatenated embeddings.

LLM-based reranker. Following Tong et al. (2026), we post-train an LLM with comparison-based Group Relative Policy Optimisation (GRPO), teaching it to prefer insights that lead to higher-quality idea generation.

A.4 Used Prompts

The goal prompts were carefully designed for each target type seperately to prevent leaking information about the target (Challenge, Method, Research Question).

Prompt to extract the goal

System Prompt:
You are a PhD level researcher with extensive experience in reading and summarising papers.
User Prompt:
What is the main research goal of the following research paper?
Exclude any information about the {Challenge, Method, Target Type} of this research paper.

Title: {title}
Abstract: {abstract}
Full Text: {full text}

Example Output:
To establish sample-complexity conditions for sparse recovery from mixed-quality measurements by quantifying the price-of-quality trade-offs between high- and low-noise data and contrasting information-theoretic limits with practical algorithmic recovery under both agnostic and informed decoding.

Prompt for extracting the challenge

System Prompt:
You are a PhD level researcher with extensive experience in reading and summarising papers.
User Prompt:
What is the main challenge of the following research paper?

Title: {title}
Abstract: {abstract}
Full Text: {full text}

Example Output:
The core difficulty is rigorously quantifying how heterogeneous noise and the mix of high- and low-quality measurements govern sample complexity, while reconciling information-theoretic limits with the practical performance and robustness of algorithms like LASSO in the agnostic setting.

Prompt for extracting the method

System Prompt:
You are a PhD level researcher with extensive experience in reading and summarising papers.
User Prompt:
What is the main method proposed in the following research paper?

Title: {title}
Abstract: {abstract}
Full Text: {full text}

Example Output:
To build a theory using linearized dynamics and random-matrix analysis to predict temporal–spatial superposition under varying memory loads, then train families of RNNs across controlled task/data/capacity grids, quantify representational geometry (dimensionality, subspace overlap, manifold capacity, time-resolved decoding), and validate by mapping empirical phase diagrams of regimes and tradeoffs against theoretical predictions.

A.5 Results on Physics

To assess whether our findings transfer beyond machine learning, we evaluate the idea generation agents on 20 papers published by the American Physical Society in December 2025, using the same setting as for the results in Table 1. Table 6 shows that the ranking of methods largely matches the one observed on ICLR 2026, with the exception that AI Scientist and AI Scientist v2 score at the bottom of the leaderboard. This is expected, as both agents are specialised for the machine learning domain and their internal prompts are optimised for machine learning research.

Agent	Challenge	Method
Research Agent	**0.772** ± 0.085	0.758 ± 0.148
Naïve Baseline	0.767 ± 0.152	**0.789** ± 0.113
Chain of Ideas	0.749 ± 0.095	0.729 ± 0.111
AI Scientist	0.732 ± 0.182	0.716 ± 0.091
AI Scientist v2	0.703 ± 0.181	0.723 ± 0.094

Table 6: Soft Recall on 20 American Physical Society papers from December 2025. Error bars are ± 1 standard deviation.

A.6 Failure modes of ideation agents

Table 7 summarises the characteristic failure modes we observe for each ideation agent when its predictions diverge from the ground truth.

A.7 Human evaluation interface

We show a screenshot of our human evaluation interface in Figure 5. Before starting, annotators are told that they will evaluate research idea agents by how well the generated ideas match a given research goal, and are shown the list of agents covered in the study. Each sample then displays the research goal together with the task, i.e. whether the agents predicted the challenge or the method, and the agents' predictions in random order. Annotators drag each prediction into a ranking slot from best to worst and submit once all predictions are ranked, while a progress bar tracks their position in the survey. After completing the study, annotators are shown which agents they preferred and how their preferences compare to those of other participants.

Agent (retrieval)	**Primary failure mode**
Naïve Baseline *None*	**Under-specification.** Captures the topic but stays at a high level of abstraction and occasionally misses the paper's novel angle (e.g. answering "robust normalization factors" instead of "does loss-curve collapse *persist at scale*").
Chain of Ideas *Idea-chain retrieval of related papers*	**Wrong-neighbour grafting.** Anchors on an adjacent paper and imports the wrong specific mechanism, or invents a named framework absent from the paper (e.g. "sparse reward problem", "computational thrashing", "mode collapse").
AI Scientist *Semantic Scholar literature search*	**Question inversion.** Reframes the challenge as carrying out its own proposed method (e.g. "to implement a gauge-invariant training objective", "to determine how feedback granularity impacts…"), stating a research plan rather than the paper's gap.
Research Agent *Heavy, structured retrieval*	**Mild over-specification.** Mostly succeeds with ground-truth-like framing (e.g. "absence of a formal theoretical framework…", "current turn-based benchmarks fail to…"), but occasionally injects an invented verification or benchmark detail the paper does not stress.
AI Scientist v2 *Retrieval and agentic experimentation loop*	**Speculative over-elaboration.** Describes the risks of *its own* proposed method rather than the paper's stated challenge, in colourful compound sentences (e.g. "Zeno-like effects", "blind leading the blind", "chimeric side-chains", "trajectory curvature destabilizing the straight-path inductive bias", "feedback-loop instabilities").

Table 7: Summary of the primary failure modes of each ideation agent, with representative phrases taken from agent predictions.

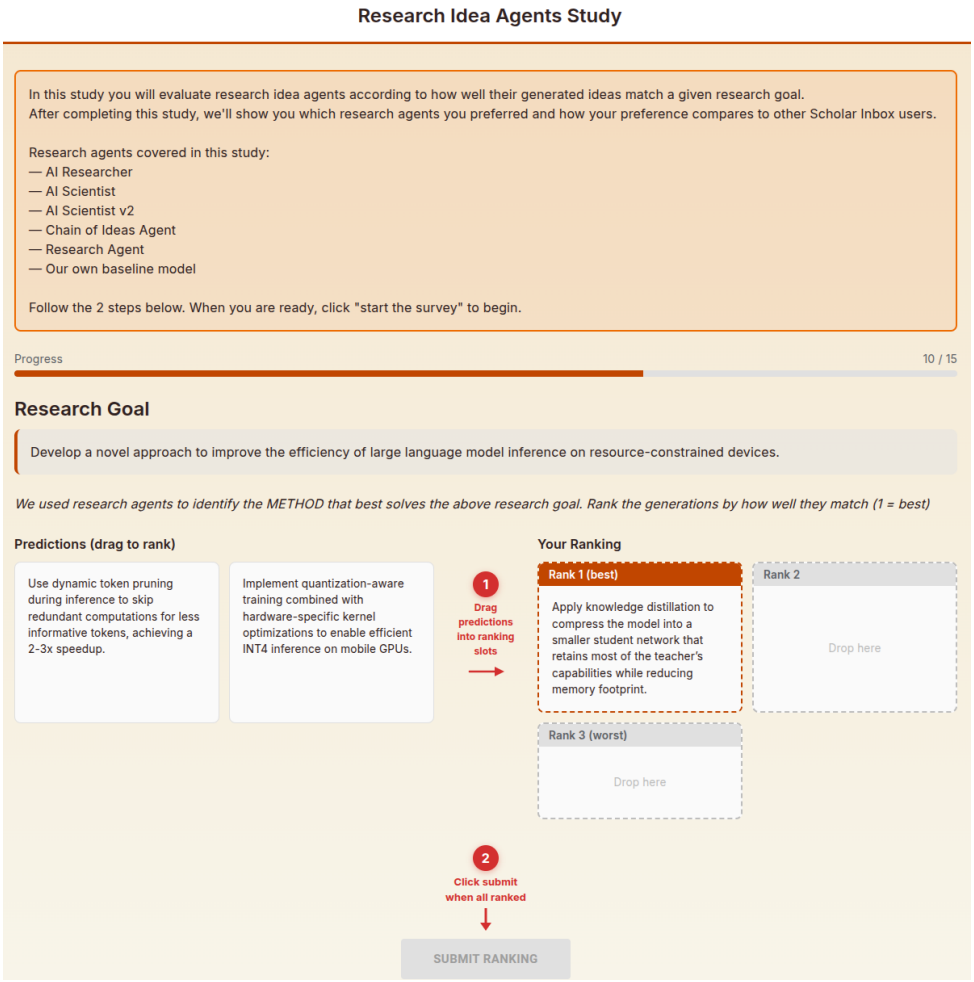


Figure 5: Screenshot of the human evaluation interface. Annotators are presented with an unordered list of predictions, each produced by a different agent, and rank them by dragging and dropping into their preferred ordering. Agent identities are not revealed.