Can LLMs predict the Next Generation of AI Research Ideas? A Trend-Aware and Cluster-Based Evaluation

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used for research idea generation, yet evaluating the novelty of open-ended ideas remains difficult. Existing evaluations often rely on embedding-based semantic similarity or LLM-based judges. However, raw similarity can confound topical relevance with conceptual novelty, while LLM judges may introduce model-specific biases that are difficult to diagnose. We introduce TrACE, a training-free **Trend-Aware and Cluster-based Evaluation** framework that estimates topic-specific past-to-future directions in embedding space and evaluates whether generated hypotheses move along these directions. TrACE focuses on *forecastable novelty*: novel contributions that represent plausible next steps in a research topic, as suggested by its past trajectory. Using AI-related papers from 2010–2024, we construct an evaluation dataset covering machine learning, NLP, and data mining. We validate that the required temporal and topical structure is recoverable across two embedding models, and show that Trend Alignment and Progression distinguish ground-truth past and future ideas more clearly than similarity-based metrics. Applying TrACE to multiple LLMs and retrieval settings, we find that generated hypotheses show partial future-oriented movement but remain below ground-truth future ideas under a conservative leave-one-out evaluation. Retrieval-augmented generation does not consistently improve trend alignment and can sometimes anchor models toward past work, although best-of-three sampling can surface more future-aligned candidates. Overall, current single-pass LLM and RAG systems generate topically plausible hypotheses, but do not yet reliably align with observed future research trajectories.

1 Introduction

The rapid growth of scientific literature makes it increasingly difficult for researchers to track emerging ideas and identify promising directions for future work. Large language models (LLMs) have recently been explored as tools for scientific research. There is growing interest in using LLMs and LLM-based agents not only for literature summarization and open-ended scientific question and answer, but also for more generative and complex research tasks, including scientific idea or hypothesis generation (Xu et al., 2025; Patel et al., 2025; Zheng et al., 2025; Panigrahi et al., 2026).

Automated hypothesis generation is not a new problem. Before LLMs, the task was often formulated as a link prediction or relationship discovery task, such as predicting previously unknown associations between biomedical concepts, drugs, treatments, or diseases (Sybrandt et al., 2018, 2020; Tyagin and Safro, 2024). In contrast, recent work often formulates research idea generation as an open-ended text generation task, where systems are expected to produce natural-language scientific ideas or hypotheses conditioned on a research background and relevant literature (Wang et al., 2024; Guo et al., 2025). This shift expands the scope of AI-driven research support, but it also raises a central challenge: *how can open-ended generated hypotheses be evaluated reliably and scalably?*

Generated ideas can be evaluated along several dimensions. Recent work has used human or LLM judges to assess qualities such as utility, feasibility, plausibility, and validity (Baek et al., 2025; Wang et al., 2024; Yang et al., 2024). These approaches are useful for determining whether an idea appears well-formed or promising, but novelty poses a distinct challenge. Unlike many quality dimensions, novelty is inherently relational: an idea is new only relative to existing work. Thus, evaluating novelty requires comparison against reference texts. Existing work has used lexical overlap, embedding similarity, or distance-based metrics to compare generated ideas with ground-truth or related works (Qi et al., 2024; Wang et al., 2024; Qiu et al., 2025; Su et al., 2025). However, lexical and single-reference

metrics are poorly suited for open-ended generation, while similarity- and distance-based metrics may confound topic relevance with conceptual novelty. As a result, automated evaluation of novelty remains challenging.

In this work, we focus on **forecastable novelty**: novel contributions that represent plausible or likely next steps of a research topic, as suggested by its past trajectory. This does not capture all forms of novelty, especially disruptive ideas or breakthroughs that depart from both past and near-future trends. However, it enables a retrospective evaluation in which future literature serves as an empirical reference. Under this framing, hypothesis generation becomes a future-oriented task: *Can LLMs generate ideas aligned with the next generation of research developments?*

Our work presents a training-free and scalable evaluation framework for assessing forecastable novelty that departs from existing similarity-based methods without relying on using LLM as a judge. Our argument is not that pretrained semantic embeddings are unsuitable for novelty evaluation, but that novelty cannot be inferred reliably from raw similarity or distance alone. Since embeddings encode many dimensions of meaning, the key challenge is to identify a latent direction that distinguishes past ideas from future developments within the same topic. Building on this perspective, we make the following contributions:

- We introduce **TrACE**, a **Trend-Aware and Cluster-based Evaluation** framework that uses pretrained semantic embeddings to evaluate *forecastable novelty* by measuring whether generated ideas move along topic-specific past-to-future directions.

- We construct a dataset covering AI, machine learning, NLP, and data mining topics, enabling retrospective evaluation of generated ideas in AI research using the TrACE framework.

- We apply TrACE to multiple LLM-based generation settings, showing that current LLMs generate partially future-aligned ideas but remain below ground-truth future ideas, and that RAG and fine-tuning can affect trend alignment in different ways.

2 Related Work

Prior work evaluates generated research ideas using judge-based, ranking-based, prediction-based, and reference-based approaches. Human and LLM judges have been used to assess qualities such as feasibility, plausibility, validity, and utility (Baek et al., 2025; Wang et al., 2024; Yang et al., 2024); other methods rank generated ideas against alternatives or ground-truth texts (Guo et al., 2024; Li et al., 2025a), or predict downstream utility such as citation potential (Höpner et al., 2025). These approaches are useful for general quality assessment, but provide only indirect evidence of novelty because novelty requires comparison with existing work.

Reference-based metrics address novelty more directly by comparing generated ideas with ground-truth or related texts using lexical overlap, embedding similarity, or distance-based measures (Qi et al., 2024; Wang et al., 2024; Qiu et al., 2025; Su et al., 2025). However, lexical and single-reference metrics are limited for open-ended generation, while similarity- and distance-based metrics may confound topical relevance with conceptual novelty. Concurrent with our work, recent future-oriented metrics such as FAS and LACERScore evaluate generated ideas with respect to later research developments, but they rely on LLM judges (Wang et al., 2026; Ajith et al., 2026). In contrast, TrACE is a training-free, embedding-based alternative.

3 Preliminaries

3.1 Task Formulation

Following prior work on open-ended scientific idea generation (Wang et al., 2024; Qiu et al., 2025), we formulate the task as conditional generation. Given a research background $\mathcal{B}$ and a set of retrieved prior works $\mathcal{A} = \{A_1, A_2, \ldots, A_k\}$, the goal is to generate a candidate research idea or hypothesis $\mathcal{I}_g$:

$$\mathcal{I}_g = f_\theta(\mathcal{B}, \mathcal{A}),$$

where f_θ denotes an LLM or LLM-based generation system. The background $\mathcal{B}$ describes the research problem, motivation, or gap, while $\mathcal{A}$ provides relevant prior literature, typically retrieved from a scientific corpus. The generated idea $\mathcal{I}_g$ is a natural-language research idea, hypothesis, or

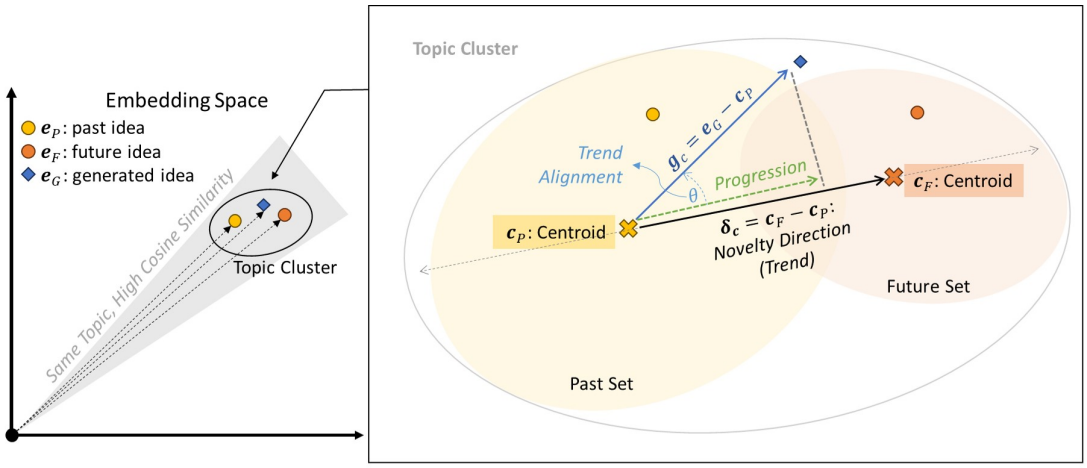

Figure 1: Illustration of the proposed trend-aware and cluster-based evaluation framework. In a semantic embedding space, related past, future, and generated ideas may have similar cosine similarities because they belong to the same topic cluster. We therefore evaluate generated ideas relative to a topic-level temporal direction $\boldsymbol{\delta} = \boldsymbol{c}_F - \boldsymbol{c}_P$, where $\boldsymbol{c}_P$ and $\boldsymbol{c}_F$ denote the centroids of past and future ideas, respectively. For a generated idea embedding $\boldsymbol{e}_G$, the generated shift is defined as $\mathbf{g}_c = \boldsymbol{e}_G - \boldsymbol{c}_P$. Trend alignment measures the angular agreement between $\mathbf{g}_c$ and $\boldsymbol{\delta}_c$, while progression along the trend measures the projection of $\mathbf{g}_c$ onto $\boldsymbol{\delta}_c$.

testable claim relevant to $\mathcal{B}$ while remaining meaningfully distinct from the prior works in $\mathcal{A}$.

3.2 Motivation: Novelty Directions in Embedding Space

Our proposed evaluation method is motivated by the linear representation hypothesis, which suggests that high-level concepts can often be represented as directions in neural representation spaces (Park et al., 2024). Prior work has shown that such directions can encode meaningful attributes, including gender relations in word embeddings (e.g., $\mathbf{w}_{queen} - \mathbf{w}_{king}$) (Allen and Hospedales, 2019) and fairness-related attributes in activation spaces estimated by contrasting representation groups (Li et al., 2025c). Building on this perspective, we treat the shift from past to future ideas within a topic cluster as a latent direction of *forecastable novelty*.

For scientific hypothesis generation, this formulation is intuitive. Future ideas often emerge by introducing a new concept, method, or perspective to prior work. Informally,

$$\text{Future Idea} \approx \text{Past Idea} + \text{Novel Direction},$$

which implies

$$\text{Novel Direction} \approx \text{Future Idea} - \text{Past Idea}.$$

We define this contrast at the cluster level. Let $\phi(\cdot)$ be a pretrained text encoder. For each topic cluster c, we partition works into a past set $\mathcal{P}_c$ and a future set $\mathcal{F}_c$, and define their centroids as

$$c_P = \frac{1}{|\mathcal{P}_c|} \sum_{\mathcal{I}_i \in \mathcal{P}_c} \phi(\mathcal{I}_i), \qquad c_F = \frac{1}{|\mathcal{F}_c|} \sum_{\mathcal{I}_i \in \mathcal{F}_c} \phi(\mathcal{I}_i).$$

The topic-level novelty direction is then

$$\boldsymbol{\delta}_c = c_F - c_P.$$

This centroid contrast does not capture all forms of scientific novelty, nor does it model the full variation within past and future works. Instead, it provides a tractable estimate of the dominant past-to-future trend among related ideas. Its usefulness therefore depends on two assumptions: (1) the embedding space contains a latent direction that reasonably separates future ideas from past ideas within a topic, and (2) the topic clusters are compact and interpretable enough to represent specific research areas. Under these assumptions, $\boldsymbol{\delta}_c$ can serve as a meaningful direction of forecastable novelty, forming the basis of our **Trend-Aware and Cluster-based Evaluation** framework, or **TrACE**.

3.3 Proposed Metrics

Given $\boldsymbol{\delta}_c$, TrACE evaluates whether a generated idea moves in the same direction as the observed future trend. For a generated idea $\mathcal{I}_g$ assigned to cluster c, with embedding $e_G = \phi(\mathcal{I}_g)$, we define its shift from past work as

$$\mathbf{g}_c = e_G - c_P.$$

The key question is whether $\mathbf{g}_c$ aligns with $\boldsymbol{\delta}_c$. if it does, the generated idea follows the observed future trajectory of the topic. Otherwise, it may be different from past work without being aligned with the topic's forecastable direction.

We define two complementary metrics. *Trend Alignment* measures directional agreement:

$$TA(\mathcal{I}_g, c) = \cos(\mathbf{g}_c, \boldsymbol{\delta}_c) = \frac{\mathbf{g}_c^\top \boldsymbol{\delta}_c}{\|\mathbf{g}_c\| \|\boldsymbol{\delta}_c\|}.$$

Progression along the Trend measures projected movement along the trend direction:

$$PT(\mathcal{I}_g, c) = \frac{\mathbf{g}_c^\top \boldsymbol{\delta}_c}{\|\boldsymbol{\delta}_c\|^2}.$$

Together, TA and PT separate direction from magnitude. TA asks whether the generated idea points toward future developments, while PT asks how far it progresses along the observed trend. A PT value of 0 indicates no movement along the trend direction, while $PT = 1$ corresponds to reaching the future centroid along $\boldsymbol{\delta}_c$. Figure 1 illustrates the intuition behind these metrics.

4 Methodology

4.1 Data Collection

We construct an idea generation dataset from papers indexed in the Semantic Scholar API (Kinney et al., 2025). For each paper, we collect the title, abstract, and other metadata. We restricted the corpus to papers published from 2010 to 2024 in major venues/journals in machine learning, natural language processing, and data mining. This includes NeurIPS, ICML, KDD, ACL, and EMNLP. We retain only papers with at least three citations and a non-empty abstract of sufficient length. This filtering ensures that each instance contains enough textual information for extracting a research background and idea, while focusing the study on well-represented research areas in AI, ML, NLP, and data mining.

To enable TrACE, we first identify topic clusters and define past and future sets within each cluster. Importantly, these clusters are constructed independently of the proposed metrics. They are used to support topic-level stratified sampling and broad coverage of research areas. We treat papers from 2010–2022 as the past set, which also serves as the training, validation, and retrieval pool, and papers from 2023–2024 as the future set for test-case selection. Topic clusters are obtained by embedding abstracts with `all-MiniLM-L6-v2`, reducing the embeddings with UMAP, and applying HDBSCAN to identify dense groups of related papers. We retain compact micro-clusters with at least 30 neighboring points, yielding 284 clusters with at least one paper in both the past and future sets. Additional details about the dataset is found in Appendix A.

4.2 Background and Idea Extraction

For each selected paper, we construct an idea-generation instance from its abstract by extracting a *background* and an *idea*. The background describes the research problem, motivation, or gap, while the idea summarizes the paper's main hypothesis, method, or contribution. This matches our task definition: given a background, generate a plausible research idea. We use GPT-5 to extract test-set instances and GPT-5-mini for the larger training and validation sets, guided by few-shot examples from the GPT-5 outputs.

To assess extraction quality, we randomly sample 1,000 training instances and 1,000 test instances. Using GPT-5 as an LLM judge, we evaluate each extracted field on a 1–5 scale for accuracy with respect to the source abstract and clarity. Overall, the extracted fields achieve high quality scores: background accuracy = 4.52, background clarity = 4.95, idea accuracy = 4.45, and idea clarity = 4.52. The test set obtains slightly higher scores than the training set, with background accuracy of 4.62 and idea accuracy of 4.56, compared with 4.42 and 4.34 for the training set. As an additional semantic preservation check, we compute the cosine similarity between the original abstract and the concatenation of the extracted background and idea. The mean cosine similarity is 0.887 across the sampled instances, suggesting that the extracted fields preserve the main semantic content of the source abstracts.

4.3 Experimental Design

We design two main sets of experiments. First, we compare conventional similarity-based metrics with the proposed trend-aware metrics using ground-truth past and future ideas. This evaluates whether TrACE better captures forecastable novelty than metrics based only on raw semantic similarity. Second, we compare different LLMs and retrieval settings to examine how retrieval-augmented generation affects trend-aware novelty. We also include

an exploratory analysis of whether task-specific fine-tuning improves performance.

4.3.1 Comparison of Similarity and Trend-Aware Metrics

We compare conventional similarity-based metrics with TrACE metrics using ground-truth past and future ideas within each topic cluster. For similarity-based evaluation, we compute each idea's average pairwise cosine similarity to the cluster's past set and future set, and define a contrastive baseline as $\Delta\text{Sim} = \text{Sim-Future} - \text{Sim-Past}$. We compare this baseline with our proposed trend-aware metrics.

We evaluate these metrics under two implementations. In the leave-one-out (LOO) setting, the evaluated future idea is removed from the future centroid and its corresponding past idea is removed from the past centroid. This yields a more conservative test by preventing the evaluated pair from directly contributing to the estimated trend direction. In the Non-LOO setting, centroids are computed using all available past and future ideas in the cluster, providing the standard estimate of the cluster-level temporal direction. We then assess whether each metric assigns higher scores to ground-truth future ideas than to their corresponding past ideas.

4.3.2 Comparison of Generation Methods

We compare LLMs and retrieval settings to examine whether LLMs can generate trend-aligned research ideas. We evaluate proprietary models, GPT-4o and GPT-4o-mini (Team, 2024); open instruction-tuned models, Llama 3.1 8B Instruct (Grattafiori et al., 2024) and Qwen 2.5 7B Instruct (Team, 2025); and SciLitLLM, a domain-specific model adapted from Qwen 2.5 7B for scientific literature tasks (Li et al., 2025b).

For retrieval-augmented generation, the retrieval pool is restricted to papers from 2010–2022. We use the target background as the query and retrieve the top 30 most similar abstracts from the past pool using `all-mpnet-base-v2` cosine similarity. We then retain only retrieved papers that belong to the same topic cluster as the target background. Thus, retrieved works must be both highly similar to the query and topically consistent with the cluster used for TrACE evaluation.

We report three generation settings. In **No RAG**, the model receives only the background. In **RAG (4 old works)**, the model receives four related works from the older past literature. The model is only given works from a truncated timeline. In **RAG (4 old, 4 influential)**, the model receives four older works and four influential but more recent prior works, where influence is measured by citation count normalized by years since publication. Each model-setting combination is run for three trials, and we report both mean performance and best-of-three performance to capture typical output quality and occasional high-quality generations under stochastic decoding.

Finally, we include a preliminary LoRA fine-tuning experiment with Qwen 2.5 7B Instruct. This experiment is exploratory rather than a full optimization study, and tests whether task-specific adaptation improves trend-aware novelty. We also evaluate a reasoning-oriented variant inspired by Self-Taught Reasoner (Zelikman et al., 2022), using a small set of accepted reasoning traces to train an additional LoRA adapter.

5 Results and Discussion

Before the main experiments, we validate the assumptions required by TrACE. Specifically, we test whether: (1) past and future ideas exhibit linear separability in the embedding space; (2) temporal structure is not limited to a single past–future split; and (3) independently defined topic clusters remain meaningful across the embedding models used for evaluation. We conduct these analyses using `SPECTER2` (Singh et al., 2023), a domain-specific scientific encoder, and `all-mpnet-base-v2` (MPNet) (Reimers and Gurevych, 2019), a general-purpose encoder. Additional details are provided in Appendix E.

Overall, the validation results support the plausibility of applying TrACE to both embedding spaces. Past and future ideas show substantial separation, with temporal splits predicted by a logistic model using MPNet (test AUC: 0.758) and SPECTER2 (test AUC: 0.819) as feature inputs. We also estimated a linear direction that is moderately correlated with publication year, suggesting that the structure is not specific to the chosen split (See Appendix, Table 7). Finally, since clusters were constructed independently, we verify that they remain predictable in both MPNet (test accuracy: 0.779) and SPECTER2 (test accuracy: 0.716) embedding spaces (See Appendix, Table 8).

Model	Case	Ground Truth	Sim. to Past (SP)		Sim. to Future (SF)		ΔSim (SF–SP)		Trend Alignment		Progression	
			Avg.	95% CI	Avg.	95% CI	Avg.	95% CI	Avg.	95% CI	Avg.	95% CI
MPNet	LOO	Future	0.482	(0.479, 0.485)	0.485	(0.482, 0.488)	0.003	(0.001, 0.005)	0.168	(0.161, 0.175)	0.414	(0.396, 0.432)
	LOO	Past	0.439	(0.435, 0.444)	0.425	(0.420, 0.429)	-0.015	(-0.016, -0.013)	0.090	(0.084, 0.097)	0.231	(0.214, 0.248)
	Non-LOO	Future	0.483	(0.480, 0.486)	0.530	(0.527, 0.533)	0.047	(0.046, 0.049)	0.379	(0.373, 0.385)	1.002	(0.986, 1.018)
	Non-LOO	Past	0.445	(0.440, 0.450)	0.430	(0.425, 0.434)	-0.016	(-0.017, -0.014)	0.088	(0.082, 0.094)	0.251	(0.234, 0.269)
SPECTER2	LOO	Future	0.887	(0.886, 0.888)	0.889	(0.888, 0.889)	0.001	(0.001, 0.002)	0.172	(0.165, 0.179)	0.428	(0.410, 0.445)
	LOO	Past	0.879	(0.878, 0.880)	0.876	(0.876, 0.877)	-0.002	(-0.003, -0.002)	0.093	(0.086, 0.099)	0.227	(0.209, 0.245)
	Non-LOO	Future	0.887	(0.887, 0.888)	0.898	(0.898, 0.899)	0.011	(0.011, 0.012)	0.381	(0.375, 0.387)	1.004	(0.988, 1.020)
	Non-LOO	Past	0.880	(0.879, 0.881)	0.877	(0.876, 0.878)	-0.003	(-0.003, -0.002)	0.087	(0.080, 0.093)	0.235	(0.217, 0.253)

Table 1: Comparison of metric behavior on ground-truth ideas under leave-one-out (LOO) and non-leave-one-out (Non-LOO) settings using all-mpnet-base-v2 (MPNet) and SPECTER2 embeddings. The table reports average pairwise cosine similarity to related past works (SP), average pairwise cosine similarity to related future works (SF), their difference (SF–SP), Trend Alignment, and Progression along the Trend. Statistics show the average metric value on the test set and the corresponding 95% confidence interval.

Model	Metric	LOO		Non-LOO	
		Acc.	Mean Diff.	Acc.	Mean Diff.
MPNet	ΔSim	63.90%	0.018	91.40%	0.063
	TA	65.80%	0.078	94.30%	0.291
	PT	63.90%	0.183	91.40%	0.750
SPECTER2	ΔSim	63.60%	0.004	90.80%	0.014
	TA	65.20%	0.079	93.10%	0.294
	PT	63.50%	0.201	90.90%	0.769

Table 2: Pairwise comparison between ground-truth future and past ideas for the same background under leave-one-out (LOO) and non-leave-one-out (Non-LOO) settings. Acc. denotes the percentage of cases where the metric assigns a higher score to the ground-truth future idea than to the ground-truth past idea. Mean Diff. is the average paired difference between the scores assigned to the future and past ideas. ΔSim is defined as Sim-Future minus Sim-Past. TA denotes Trend Alignment, PT denotes Progression along the Trend, and MPNet refers to `all-mpnet-base-v2`.

5.1 Comparison of Similarity and Trend-Aware Metrics

Table 1 compares the behavior of similarity-based and trend-aware metrics on ground-truth past and future ideas. The results show that average pairwise cosine similarity can be misleading when used alone. Both past and future ideas remain highly similar to both past and future comparison sets, especially under SPECTER2. This suggests that raw similarity mainly reflects topical relatedness within a cluster rather than temporal movement.

The contrastive similarity metric, $\Delta\text{Sim} = \text{SF} - \text{SP}$, provides a stronger baseline than using similarity to only one reference set. This is partly because ΔSim is itself related to a past–future contrast direction. We provide a formal explanation of why ΔSim can be viewed as a projection-like score onto a normalized past–future contrast direction in Appendix D. Empirically, ground-truth future ideas have slightly positive ΔSim values on average, while past ideas have slightly negative values. However, these differences are small, especially under SPECTER2. For example, in the LOO setting, future ideas have an average ΔSim of only 0.001 under SPECTER2 and 0.003 under MPNet. Thus, while future ideas tend to have higher contrastive similarity than past ideas, the magnitude is difficult to interpret in practice. The higher overall similarity values under SPECTER2 also suggest that it better captures the topical relatedness of scientific texts.

In contrast, Trend Alignment (TA) and Progression along the Trend (PT) provide clearer separation between past and future ideas across both MPNet and SPECTER2, and under both LOO and Non-LOO settings. We use the LOO averages for ground-truth future ideas as conservative reference values for interpreting LLM-generated ideas in the next experiment. For example, under SPECTER2, a generated idea with an average TA close to 0.172 is comparable to the average ground-truth future idea under the LOO setting. By contrast, the Non-LOO averages for the future set describe the relative dispersion of individual future ideas around the cluster-specific novelty direction.

Table 2 further evaluates whether each metric assigns higher scores to ground-truth future ideas than to their corresponding past ideas. In both LOO and Non-LOO settings, TA achieves higher accuracy than ΔSim across both embedding models. Although the accuracy differences are modest, the mean differences are much larger for TA and PT than for ΔSim. This indicates a stronger effect size and a more interpretable separation between past and future ideas. In particular, near-zero or negative TA and PT values indicate that an idea is not moving toward the estimated future direction,

Method		Mean of 3 Trials				Best of 3 Trials			
Model	Retrieval Case	TA		PT		TA		PT	
		Avg.	95% CI	Avg.	95% CI	Avg.	95% CI	Avg.	95% CI
gpt-4o	No RAG	0.128	(0.121, 0.134)	0.325	(0.308, 0.341)	0.185	(0.179, 0.192)	0.497	(0.480, 0.514)
	RAG (4 old works)	0.128	(0.121, 0.135)	0.321	(0.305, 0.337)	0.188	(0.182, 0.195)	0.497	(0.480, 0.514)
	RAG (4 old, 4 influential)	0.131	(0.125, 0.138)	0.332	(0.315, 0.348)	0.195	(0.188, 0.201)	0.516	(0.499, 0.533)
gpt-4o-mini	No RAG	0.122	(0.116, 0.129)	0.309	(0.291, 0.326)	0.177	(0.170, 0.184)	0.472	(0.454, 0.489)
	RAG (4 old works)	0.117	(0.110, 0.123)	0.288	(0.272, 0.305)	0.175	(0.168, 0.182)	0.456	(0.440, 0.473)
	RAG (4 old, 4 influential)	0.121	(0.115, 0.128)	0.299	(0.283, 0.315)	0.180	(0.173, 0.187)	0.466	(0.449, 0.482)
Llama 3.1 8B	No RAG	0.103	(0.097, 0.110)	0.244	(0.229, 0.260)	0.177	(0.170, 0.183)	0.457	(0.441, 0.473)
	RAG (4 old works)	0.095	(0.088, 0.101)	0.218	(0.203, 0.233)	0.170	(0.164, 0.177)	0.439	(0.423, 0.454)
	RAG (4 old, 4 influential)	0.092	(0.086, 0.099)	0.213	(0.198, 0.228)	0.171	(0.164, 0.177)	0.441	(0.425, 0.456)
SciLitLLM	No RAG	0.118	(0.112, 0.124)	0.322	(0.305, 0.338)	0.195	(0.189, 0.201)	0.575	(0.558, 0.592)
	RAG (4 old works)	0.112	(0.107, 0.118)	0.314	(0.298, 0.330)	0.194	(0.189, 0.200)	0.595	(0.578, 0.612)
	RAG (4 old, 4 influential)	0.125	(0.119, 0.130)	0.356	(0.340, 0.372)	0.206	(0.200, 0.211)	0.631	(0.614, 0.648)
Qwen 2.5 7B	No RAG	0.127	(0.120, 0.134)	0.322	(0.305, 0.339)	0.181	(0.174, 0.188)	0.479	(0.462, 0.497)
	RAG (4 old works)	0.128	(0.121, 0.134)	0.330	(0.313, 0.346)	0.188	(0.182, 0.195)	0.510	(0.492, 0.527)
	RAG (4 old, 4 influential)	0.131	(0.124, 0.137)	0.340	(0.324, 0.356)	0.191	(0.185, 0.198)	0.519	(0.502, 0.536)
LoRA fine-tuned Qwen variants									
Qwen (Task FT)	No RAG	0.108	(0.101, 0.115)	0.245	(0.230, 0.261)	0.191	(0.184, 0.198)	0.479	(0.463, 0.495)
	RAG (4 old works)	0.105	(0.099, 0.112)	0.237	(0.222, 0.252)	0.190	(0.183, 0.197)	0.475	(0.459, 0.491)
	RAG (4 old, 4 influential)	0.109	(0.103, 0.116)	0.249	(0.233, 0.264)	0.193	(0.186, 0.199)	0.483	(0.467, 0.499)
Qwen (Task+Reasoning FT)	No RAG	0.106	(0.100, 0.113)	0.238	(0.223, 0.253)	0.198	(0.191, 0.205)	0.491	(0.475, 0.506)
	RAG (4 old works)	0.107	(0.100, 0.113)	0.243	(0.228, 0.258)	0.194	(0.188, 0.201)	0.492	(0.476, 0.508)
	RAG (4 old, 4 influential)	0.107	(0.100, 0.113)	0.239	(0.224, 0.254)	0.201	(0.195, 0.208)	0.502	(0.486, 0.517)

Table 3: Performance of different models and selected retrieval settings under leave-one-out trend-aware evaluation. Results are reported for the mean of three trials and the best of three trials. TA denotes Trend Alignment and PT denotes Progression along the Trend. Qwen (Task FT) uses a LoRA adapter fine-tuned for the idea generation task. Qwen (Task+Reasoning FT) uses the task adapter together with an additional LoRA adapter trained to generate reasoning before producing the final response.

and may instead remain closer to the past trajectory. Overall, these results support the use of trend-aware metrics as a more interpretable alternative to raw or contrastive similarity for evaluating forecastable novelty.

5.2 Comparison of Generation Methods

Table 3 reports the main generation results using SPECTER2 embeddings under the LOO setting. We focus on SPECTER2 because ideas within the same topic cluster appear more similar in this embedding space, suggesting stronger sensitivity to scientific-text relatedness. We use the LOO setting so that generated ideas can be compared against the conservative LOO averages for ground-truth future ideas. The complete results across embeddings, LOO/Non-LOO settings, and retrieval variants are provided in Appendix F. Also, the prompt design for the experiment are found in Appendix B.

Overall, the mean-of-three results show that single-pass generation remains below the ground-truth future reference. Based on Table 1, ground-truth future ideas obtain an average TA of 0.172 and PT of 0.428. In comparison, the highest mean TA among generated ideas is 0.131, achieved by GPT-4o and Qwen 2.5 7B under RAG with old and influential works. Thus, LLMs produce partially future-aligned ideas on average, but their typical outputs remain less aligned than actual future ideas.

However, the best-of-three results show that LLMs can sometimes generate ideas comparable to, or better than, the average ground-truth future idea. Several models exceed the LOO future-reference TA of 0.172, and SciLitLLM reaches the strongest best-of-three performance with TA = 0.206 and PT = 0.631 under RAG with old and influential works. This suggests that LLMs are capable of sampling trend-aligned ideas, but may require iteration or candidate selection to reliably surface them.

The effect of RAG is mixed. Across models, RAG does not provide a consistent average improvement over No RAG, and retrieval using old works can sometimes reduce TA and PT, suggesting that retrieved prior works may pull generations toward older ideas. Including influential works occasionally improves performance, but the gains are generally small. This indicates that RAG design matters: retrieval can provide useful context, but it can also anchor the model to past literature rather than help it move toward the future direction.

Fine-tuning also shows mixed effects. Compared with the base Qwen 2.5 7B model, task-specific

LoRA fine-tuning does not improve mean performance and often lowers average TA and PT. Under best-of-three selection, however, fine-tuned variants become more competitive. The reasoning-oriented variant performs slightly better than task-only fine-tuning, reaching TA = 0.201 under RAG with old and influential works. These results suggest that reasoning supervision, careful RAG design, and iterative generation may be important for improving trend-aware hypothesis generation, although the observed differences are still modest.

5.3 Examples of Generated Ideas

This section illustrates how TrACE scores correspond to qualitative differences in generated ideas. We compare examples with different leave-one-out Trend Alignment (TA) scores.

Example 1: LLM Watermarking, Attacks, and Robust Decoding
Background: Existing watermarking methods are largely zero-bit and only support binary detection, limiting their usefulness for tracing adversarial users.
Lower-score idea (Qwen2.5-7B; TA = 0.154): Embedding cryptographically secure identifiers within LLM outputs can enable both detection and traceability of adversarial users.
Higher-score idea (GPT-4o; TA = 0.472): Embedding multi-bit digital watermarks into LLM outputs can enable tracing of individual users by encoding unique identifiers while preserving text quality.
Future ground truth (TA = 0.505): Allocating token positions across a generated sequence to independently encode sub-units of a multi-bit watermark enables robust recovery of long identifiers under corruption while preserving text quality and retaining zero-bit detectability.

For the example 1, the lower-score idea is relevant but broad: it mentions traceability without specifying multi-bit encoding or robustness. The higher-score idea is closer to the future direction because it explicitly proposes multi-bit watermarks for user tracing while preserving text quality. This illustrates that higher TA reflects alignment with a more specific technical direction, not only topical similarity.

Example 2 shows how retrieval can anchor generation. The RAG output adopts retrieved concepts such as predictive uncertainty and confidence-based weighting, making it plausible but less aligned with the future direction centered on modality-specific encoders, learned fusion, multi-loss training, and contextual representation. Additional examples are provided in Appendix G.

Example 2: Multimodal Sentiment Analysis and Fusion
Background: Selecting and combining feature encoders from heterogeneous modalities remains challenging.
No RAG (GPT-4o; TA = 0.176): Incorporating attention mechanisms that dynamically weigh feature encoders based on contextual sentiment shifts can enhance sentiment detection across heterogeneous modalities.
Retrieved old works: Retrieved abstracts emphasize "predictive uncertainty estimates," "well-calibrated uncertainty estimates," and uncertainty on "out-of-distribution examples."
RAG with 4 old works (GPT-4o; TA = 0.154): Incorporating predictive uncertainty estimates from non-Bayesian neural networks into the fusion process can enhance sentiment detection by dynamically weighting each modality based on confidence levels.

6 Conclusion

We presented TrACE, a training-free framework that evaluates scientific hypothesis generation by measuring whether generated ideas move along empirically observed past-to-future directions within topic clusters. TrACE can be used as an alternative or complement to recent LLM-based evaluation methods because it assesses forecastable novelty without relying on LLM-as-a-judge. While the proposed metrics do not promise perfect discrimination, they are grounded in explicit and verifiable assumptions: temporal directions should be recoverable in the embedding space, and topic clusters should be compact and interpretable. This makes the framework more diagnosable than LLM-based judges, whose model-specific biases can be difficult to identify or validate. Our results show that current LLM-based systems can generate topically plausible ideas with some future-oriented movement, but single-pass generations only partially align with observed future research trajectories. Overall, TrACE provides a relatively simple tool for evaluating whether generated scientific ideas follow forecastable directions of research development.

7 Limitations

This work has several limitations. First, TrACE evaluates *forecastable novelty*: ideas that align with future developments observed in the literature. It is not intended to capture all forms of novelty, especially disruptive ideas that depart from both past and near-future trends.

Second, TrACE depends on the quality of the embedding space, topic clusters, and estimated past-to-future directions. We address this through assumption-validation, robustness checks, and both LOO and non-LOO evaluation. However, because

our clusters are designed to represent specific research topics, some clusters contain relatively few papers within each temporal split. This can make the estimated novelty direction sensitive to individual samples, partly explaining the gap between LOO and non-LOO results. Future work may improve stability using larger per-cluster samples or more robust alternatives to centroids, such as medoids, median centers, or representations that better account for within-cluster variation.

Third, our experiments focus on AI, machine learning, NLP, and data mining papers from selected venues. Future work should test whether similar temporal structure can be recovered in other scientific domains and publication settings.

Fourth, the dataset uses LLM-extracted background and idea fields from abstracts. While we perform automatic validation and consistency checks, future work should include human evaluation of these extracted fields to further verify their accuracy and usefulness.

Finally, TrACE measures directional alignment with future trends, but it does not directly evaluate feasibility, testability, causal validity, or broader scientific value. It should therefore be used as a complementary metric alongside expert judgment and other forms of hypothesis evaluation.

8 Ethical Considerations

This work studies the evaluation of scientific hypothesis generation. The generated outputs should not be treated as validated scientific claims. Rather, they should be viewed as candidate ideas that require expert review, empirical validation, and responsible interpretation. Because automated idea generation may produce plausible but incorrect, derivative, or misleading hypotheses, we emphasize that such systems should support, not replace, human scientific judgment.

We used formal model access channels, including official APIs for proprietary models and publicly released model checkpoints for open models. For LLM-based extraction and generation, we avoided using private or sensitive data and restricted the corpus to publicly available scientific metadata and abstracts. To support transparency and reproducibility, we plan to release the dataset construction scripts, prompts, evaluation code, and processed resources in a public GitHub repository after the blind review period. This release will follow the licensing and redistribution constraints of the original data sources and model providers.

9 Use of AI in this Research

AI tools were used as research assistants during the preparation of this work. Specifically, LLMs were used to support code optimization, debugging, prompt refinement, and writing assistance, including grammar correction. LLMs were also used as part of the data construction pipeline for extracting background and idea fields, as described in the methodology. All experimental design choices, metric definitions, interpretation of results, and final manuscript decisions were made and reviewed by the authors.

References

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

Carl Allen and Timothy Hospedales. 2019. Analogies explained: Towards understanding word embeddings. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 223–231. PMLR.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Myles Kim, Corey M. Williams, Stefan Bekiranov, and Aidong Zhang. 2025. Ideabench: Benchmarking large language models for research idea generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 5888–5899, New York, NY, USA. Association for Computing Machinery.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, and Aidong Zhang. 2024. Embracing Foundation Models for Advancing Scientific Discovery. In *2024 IEEE International Conference on Big Data (BigData)*, pages 1746–1755, Washington, DC, USA. IEEE.

Niklas Höpner, Leon Eshuijs, Dimitrios Alivanistos, Giacomo Zamprogno, and Ilaria Tiddi. 2025. Automatic evaluation metrics for artificially generated scientific research. *Preprint*, arXiv:2503.05712.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David Graham, Fangzhou Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Chris Newell, Smita Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, Amber Tanaka, Alex D. Wade, Linda Wagner, Lucy Lu Wang, Chris Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine Van Zuylen, and Daniel S. Weld. 2025. The semantic scholar open data platform. *Preprint*, arXiv:2301.10140.

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. 2025a. 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.

Sihang Li, Jin Huang, Jiaxi Zhuang, Yaorui Shi, Xiaochen Cai, Mingjun Xu, Xiang Wang, Linfeng Zhang, Guolin Ke, and Hengxing Cai. 2025b. ScilitLLM: How to adapt LLMs for scientific literature understanding. In *The Thirteenth International Conference on Learning Representations*.

Yichen Li, Zhiting Fan, Ruizhe Chen, Xiaotang Gai, Luqi Gong, Yan Zhang, and Zuozhu Liu. 2025c. FairSteer: Inference time debiasing for LLMs with dynamic activation steering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11293–11312, Vienna, Austria. Association for Computational Linguistics.

Siba Smarak Panigrahi, Jovana Videnović, and Maria Brbic. 2026. Heurekabench: A benchmarking framework for AI co-scientist. In *The Fourteenth International Conference on Learning Representations*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Liana Patel, Negar Arabzadeh, Harshit Gupta, Ankita Sundar, Ion Stoica, Matei Zaharia, and Carlos Guestrin. 2025. Deepscholar-bench: A live benchmark and automated evaluation for generative research synthesis. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. 2024. Large language models as biomedical hypothesis generators: A comprehensive evaluation. *Preprint*, arXiv:2407.08940.

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. *Preprint*, arXiv:2504.14191.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566, Singapore. Association for Computational Linguistics.

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, Philip Torr, Bowen Zhou, and Nanqing Dong. 2025. Many heads are better than one: Improved scientific idea generation by a LLM-based multi-agent system. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28201–28240, Vienna, Austria. Association for Computational Linguistics.

Justin Sybrandt, Micheal Shtutman, and Ilya Safro. 2018. Large-scale validation of hypothesis generation systems via candidate ranking. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1494–1503.

Justin Sybrandt, Ilya Tyagin, Michael Shtutman, and Ilya Safro. 2020. Agatha: Automatic graph mining and transformer based hypothesis generation approach. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2757–2764, New York, NY, USA. Association for Computing Machinery.

OpenAI Team. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Qwen Team. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Ilya Tyagin and Ilya Safro. 2024. Dyport: dynamic importance-based biomedical hypothesis generation benchmarking technique. *BMC Bioinformatics*, 25(1):213.

Heng Wang, Pengcheng Jiang, Jiashuo Sun, Zhiyi Shi, Haofei Yu, Jiawei Han, and Heng Ji. 2026. Learning to predict future-aligned research proposals with language models. *Preprint*, arXiv:2603.27146.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. SciMON: Scientific Inspiration Machines Optimized for Novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–299, Bangkok, Thailand. Association for Computational Linguistics.

Tianze Xu, Pengrui Lu, Lyumanshan Ye, Xiangkun Hu, and Pengfei Liu. 2025. Researcherbench: Evaluating deep ai research systems on the frontiers of scientific inquiry. *Preprint*, arXiv:2507.16280.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. Large Language Models for Automated Open-domain Scientific Hypotheses Discovery. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13545–13565, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. STar: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From automation to autonomy: A survey on large language models in scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17733–17750, Suzhou, China. Association for Computational Linguistics.

A Additional details about the Data

As described above, we collected paper metadata using the Semantic Scholar API (Kinney et al., 2025). The starting corpus consists of 65,473 papers published between 2010 and 2024 from major AI, machine learning, natural language processing, and data mining venues: NeurIPS, ACL, EMNLP, ICML, ICLR, KDD, NAACL, AISTATS, JMLR, EACL, TACL, and IJCNLP/AACL. This larger corpus is used to construct topic clusters and to form the retrieval pool for the RAG-based generation settings. For retrieval, we use papers from 2010–2022 with complete metadata, ensuring that retrieved inspirations are drawn only from works available before the future evaluation period.

We define papers from 2010–2022 as the past set and papers from 2023–2024 as the future set. Topic clusters are constructed from the starting corpus by embedding paper abstracts with `all-MiniLM-L6-v2`, reducing the embeddings with UMAP, and applying HDBSCAN to identify dense groups of related papers. We retain compact micro-clusters with at least 30 neighboring points, which allows us to estimate topic-level past-to-future directions. This process yields 284 clusters with sufficient representation in both temporal periods.

Year	Test	Train	Valid
2011	0	10	0
2012	0	14	0
2013	0	21	0
2014	0	59	0
2015	0	59	0
2016	0	66	0
2017	0	1,022	94
2018	0	1,738	163
2019	0	2,569	269
2020	0	2,984	279
2021	0	3,967	413
2022	0	4,683	482
2023	1,398	0	0
2024	1,526	0	0
Total	2,924	17,192	1,700

Table 4: Year distribution of papers retained for background–idea extraction, grouped by dataset split.

A smaller subset of papers is then used to construct the background–idea extraction dataset. We select test cases from future-set papers using stratified sampling within each topic cluster. For clusters with fewer than 10 future-set papers, all available cases are included. For larger clusters, we sample either 10 papers or 10% of the cluster's future-set papers, whichever is larger. The remaining past-set papers are used for training and validation, with 10% assigned to validation using proportional allocation across clusters. This produces 21,816 extracted cases: 17,192 for training, 1,700 for validation, and 2,924 for testing. After filtering out cases from clusters with insufficient samples for metric computation, the final experimental test set contains 2,895 cases. Tables 4 and 5 report the distribution of the extracted dataset by year, split, and venue.

B Prompt Design

Prompt design. We use a shared prompt design across all LLM generation experiments to ensure that outputs are comparable across models and retrieval settings. Each prompt provides a research background and asks the model to generate one concise, novel hypothesis. The no-retrieval setting uses

Venue	Test	Train	Valid	Total
NeurIPS	647	3,939	401	4,987
EMNLP	533	3,019	284	3,836
ACL	437	2,490	245	3,172
ICML	406	2,268	237	2,911
ICLR	411	1,812	172	2,395
NAACL	150	1,070	118	1,338
AISTATS	98	906	67	1,071
KDD	103	509	55	667
JMLR	21	578	53	652
EACL	77	366	52	495
TACL	22	204	15	241
IJCNLP	19	31	1	51
Total	2,924	17,192	1,700	21,816

Table 5: Venue distribution of papers retained for background–idea extraction, grouped by dataset split.

only the background, while retrieval-augmented settings additionally provide related works as inspiration and explicitly instruct the model to produce a hypothesis that is distinct from the retrieved works. In the main experiments, we focus on three representative settings: no RAG, RAG with four older related works, and RAG with four older works plus four influential recent works. Additional retrieval variants follow the same general template and are reported in the appendix.

For all LLMs, we generate three independent trials per test case and retrieval setting. We use the same decoding configuration across models: temperature $= 0.7$, top-$p = 0.9$, and a maximum generation length of 500 tokens. Each output is required to contain only one concise hypothesis, without explanations or additional details.

The system-level prompt is:

System prompt

You are a research assistant for scientific ideation. Given a research background and related works, propose one concise, plausible, and distinct new hypothesis. Do not repeat the related works verbatim.

Below are the prompt outlines used for the three main generation settings.

No RAG

The task is to propose a new hypothesis for a given research background.
Research motivation/background: {background}
Return one concise novel hypothesis for the given background. No additional details.

RAG: 4 old works

The task is to propose a new hypothesis for a given research background. Use the list of related works as inspiration, but ensure that the proposed hypothesis is novel or distinct from existing works.
Research motivation/background: {background}
Past works have shown the following: {*abstracts, 4 old works*}
Return one concise novel hypothesis for the given background. No additional details.

RAG: 4 old + 4 influential

The task is to propose a new hypothesis for a given research background. Use the list of related works as inspiration, but ensure that the proposed hypothesis is novel or distinct from existing works.
Research motivation/background: {background}
Past works have shown the following: {*abstracts, 4 old works*}
But more recently, the following works have shown: {*abstracts, 4 recent but relatively high-citation works*}
Return one concise novel hypothesis for the given background. No additional details.

C Hyperparameters and Computational Resources

Fine-tuning and compute settings. For the fine-tuned model variants, we train LoRA adapters on top of `Qwen2.5-7B-Instruct` using supervised fine-tuning. We use rank $r = 16$, $\alpha = 32$, dropout $= 0.05$, a learning rate of 2×10^{-5}, cosine scheduling with a warmup ratio of 0.03, weight decay of 0.01, and a maximum sequence length of 4096 tokens. Training uses gradient accumulation with an effective batch size of 16, BF16 precision, evaluation and checkpointing every 500 steps,

and early stopping based on validation loss. The LoRA adapters are applied to all attention projection matrices and feed-forward components, including `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. The task-only adapter is trained for one epoch, while the task+reasoning adapter is trained for two epochs because it uses a smaller set of accepted reasoning traces.

For the task+reasoning setting, we construct the training data using a rejection-sampling procedure inspired by STaR. Candidate reasoning traces are retained only when including the trace leads to a generated hypothesis with Trend Alignment of at least 0.500 and Progression of at least 0.700. To avoid circularity between trace selection and final evaluation, this rejection-sampling step uses `all-mpnet-base-v2` embeddings, whereas the main reported evaluation uses SPECTER2 embeddings.

All open-model inference and fine-tuning experiments were run on laboratory computing resources equipped with an NVIDIA RTX A6000 GPU with 49GB of memory.

D Relationship Between Contrastive Similarity and Novelty Direction

Mean pairwise cosine similarity can be interpreted as a centroid-based comparison when embeddings are normalized. Let $\hat{e} = e/\|e\|$ be the normalized embedding of an evaluated idea, and let $\hat{e}_i = e_i/\|e_i\|$ denote the normalized embedding of idea i in a reference set $\mathcal{S}$. The average cosine similarity between e and all ideas in $\mathcal{S}$ is

$$\begin{aligned}\mathrm{Sim}(e, \mathcal{S}) &= \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \cos(e, e_i) \\ &= \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \hat{e}^\top \hat{e}_i \\ &= \hat{e}^\top \left(\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \hat{e}_i \right).\end{aligned}$$

Thus, mean pairwise cosine similarity is equivalent to the inner product between the normalized evaluated idea and the centroid of normalized reference embeddings.

This relationship helps explain why contrastive similarity ΔSim can distinguish past and future ideas. Let $\bar{c}_P$ and $\bar{c}_F$ denote the centroids of normalized embeddings in the past and future sets, respectively:

$$\bar{c}_P = \frac{1}{|\mathcal{P}_c|} \sum_{i \in \mathcal{P}_c} \hat{e}_i, \qquad \bar{c}_F = \frac{1}{|\mathcal{F}_c|} \sum_{i \in \mathcal{F}_c} \hat{e}_i.$$

Then,

$$\begin{aligned}\Delta\mathrm{Sim} &= \text{Sim-Future} - \text{Sim-Past} \\ &= \hat{e}^\top \bar{c}_F - \hat{e}^\top \bar{c}_P \\ &= \hat{e}^\top (\bar{c}_F - \bar{c}_P).\end{aligned}$$

Therefore, ΔSim can be viewed as an inner product with a normalized-space past–future direction. If the TrACE novelty direction is defined as

$$\boldsymbol{\delta}_c = c_F - c_P,$$

then the contrastive similarity direction

$$\tilde{\boldsymbol{\delta}}_c = \bar{c}_F - \bar{c}_P$$

can be viewed as an alternative approximation of a topic-specific novelty direction, computed from normalized embeddings. In this sense, ΔSim is not merely a heuristic baseline: it is also a function of a possible past–future novelty direction.

However, ΔSim remains different from the proposed trend-aware metrics. First, ΔSim takes the inner product directly with the normalized evaluated idea $\hat{e}$, whereas TrACE uses the displacement from the past centroid,

$$\mathbf{g}_c = e - c_P.$$

Second, ΔSim uses the normalized-space direction $\tilde{\boldsymbol{\delta}}_c = \bar{c}_F - \bar{c}_P$, whereas TrACE uses the topic-specific novelty direction $\boldsymbol{\delta}_c = c_F - c_P$.

Third, ΔSim is proportional to the cosine similarity between $\hat{e}$ and $\tilde{\boldsymbol{\delta}}_c$:

$$\begin{aligned}\Delta\mathrm{Sim} &= \hat{e}^\top \tilde{\boldsymbol{\delta}}_c \\ &= \|\tilde{\boldsymbol{\delta}}_c\| \cos(\hat{e}, \tilde{\boldsymbol{\delta}}_c),\end{aligned}$$

because $\|\hat{e}\| = 1$. By contrast, Trend Alignment is the cosine similarity between the generated shift $\mathbf{g}_c$ and the TrACE novelty direction $\boldsymbol{\delta}_c$.

Thus, both ΔSim and TA use a past–future contrast direction, but they apply it differently. ΔSim evaluates whether the idea is relatively closer to future than past works in normalized embedding space, while TA evaluates whether the idea's shift from the past centroid points along the estimated novelty direction.

Model Type	Embedding	Train	Test
Global Linear Model			
	all-mpnet-base-v2	0.845	0.758
	SPECTER2	0.888	0.819
Cluster-Specific Model			
	all-mpnet-base-v2	0.996	0.675
	SPECTER2	0.996	0.677

Table 6: Linear separability analysis of past and future ideas using logistic regression. AUC values are reported for global linear models and cluster-specific models under both training and test splits.

Progression further differs from ΔSim because it is normalized by the squared length of the TrACE novelty direction:

$$PT(e, c) = \frac{\mathbf{g}_c^\top \boldsymbol{\delta}_c}{\|\boldsymbol{\delta}_c\|^2}.$$

This denominator rescales the score by the size of the estimated temporal shift, giving PT a direct trajectory interpretation. A value of $PT = 0$ corresponds to the past centroid, while $PT = 1$ corresponds to the future centroid along $\boldsymbol{\delta}_c$. In contrast, ΔSim does not measure the idea's displacement from the past centroid or rescale that displacement by the size of the past–future shift. Thus, contrastive similarity and TrACE metrics are related through a past–future directional contrast, but TrACE more directly evaluates movement along the topic-specific novelty direction.

E Robustness and Assumption Validation

E.1 Linear Separability of Temporal Splits

We first test whether past and future ideas are linearly separable in the embedding space. We train logistic regression models to distinguish past from future ideas using either a global temporal boundary or cluster-specific temporal boundaries. AUC is used because it evaluates ranking performance without requiring a fixed decision threshold.

Table 6 shows that both embedding models encode recoverable temporal information. The global model achieves strong test AUCs, especially with SPECTER2, suggesting that past–future movement is partly captured by a shared direction in embedding space. Cluster-specific models achieve near-perfect training AUCs but lower test AUCs, indicating that local temporal boundaries are more sensitive to sample composition. These results support the use of temporal directions while motivating the more conservative LOO evaluation.

Embedding	Correlation	
	Train	Test
all-mpnet-base-v2	0.510	0.439
SPECTER2	0.516	0.473

Table 7: Temporal feature sensitivity analysis. Values report the correlation between the linear temporal predictor and the true temporal value.

Embedding	Macro F1-score		Accuracy	
	Train	Test	Train	Test
all-mpnet-base-v2	0.987	0.742	0.978	0.779
SPECTER2	0.974	0.675	0.956	0.716

Table 8: Transferability of cluster structure across embedding models. Macro F1-score and accuracy measure how well cluster labels remain detectable when represented using another embedding space.

E.2 General Latent Direction for Temporal Feature

We next test whether the learned temporal direction reflects ordering across time rather than only binary separability between two periods. We compute the correlation between the linear temporal predictor and the true temporal value.

Table 7 shows moderate positive correlations across embeddings and splits. This suggests that the learned direction captures a temporal gradient in the embedding space. SPECTER2 gives slightly stronger test correlation than all-mpnet-base-v2, indicating better preservation of temporal structure for this task. Although the train–test gap reflects the noisy and topic-dependent nature of scientific change, the positive test correlations support the assumption that temporal information is recoverable.

E.3 Transferability of Clusters

Finally, we test whether the independently defined clusters remain meaningful across embedding models. We train classifiers to predict cluster labels from each evaluation embedding space. Because clusters are imbalanced, we report both accuracy and macro F1-score.

Table 8 shows that cluster labels remain strongly recoverable. Test performance is far above random prediction despite the large number of imbalanced clusters, with all-mpnet-base-v2 reaching 0.742 macro F1 and 0.779 accuracy, and

SPECTER2 reaching 0.675 macro F1 and 0.716 accuracy. These results suggest that the clusters capture stable topical structure rather than artifacts of a single embedding model. The train–test gap also indicates that the clusters are not perfectly transferable, which is expected for fine-grained scientific topics.

F Complete Experiment Results

This appendix reports the complete set of experimental results across all evaluated models, retrieval settings, and scoring metrics. These tables extend the main results by including additional RAG variants and both mean-of-three and best-of-three generation settings. They are provided to support transparency and to show that the main findings are consistent across alternative prompting and retrieval configurations.

G Examples of Generated Ideas

This appendix presents representative examples of generated hypotheses from different models and retrieval settings. For each example, we show the input background, the reference future idea, and selected model-generated ideas. These examples are intended to illustrate how the quantitative metrics correspond to qualitative differences in specificity, novelty, and alignment with future research directions.

Method		Mean of 3 Trials				Best of 3 Trials			
Model	Retrieval Case	TA		PT		TA		PT	
		Avg.	95% CI	Avg.	95% CI	Avg.	95% CI	Avg.	95% CI
gpt-4o	No RAG	0.179	(0.173, 0.184)	0.499	(0.483, 0.515)	0.237	(0.231, 0.243)	0.678	(0.661, 0.695)
	RAG (4 old works)	0.172	(0.167, 0.178)	0.473	(0.458, 0.489)	0.234	(0.228, 0.240)	0.657	(0.641, 0.673)
	RAG (4 old, 4 recent)	0.170	(0.164, 0.175)	0.469	(0.454, 0.485)	0.233	(0.228, 0.239)	0.659	(0.643, 0.676)
	RAG (any 8 works)	0.172	(0.166, 0.177)	0.476	(0.460, 0.491)	0.235	(0.229, 0.240)	0.666	(0.650, 0.682)
	RAG (4 old, 4 influential)	0.172	(0.167, 0.178)	0.475	(0.459, 0.490)	0.237	(0.231, 0.243)	0.667	(0.651, 0.683)
gpt-4o-mini	No RAG	0.171	(0.165, 0.177)	0.475	(0.458, 0.491)	0.227	(0.221, 0.232)	0.646	(0.629, 0.663)
	RAG (4 old works)	0.160	(0.154, 0.165)	0.434	(0.419, 0.450)	0.219	(0.213, 0.225)	0.611	(0.595, 0.627)
	RAG (4 old, 4 recent)	0.159	(0.153, 0.164)	0.428	(0.413, 0.444)	0.221	(0.215, 0.227)	0.611	(0.595, 0.628)
	RAG (any 8 works)	0.161	(0.156, 0.167)	0.438	(0.422, 0.453)	0.222	(0.217, 0.228)	0.616	(0.600, 0.633)
	RAG (4 old, 4 influential)	0.161	(0.155, 0.167)	0.435	(0.419, 0.451)	0.221	(0.215, 0.227)	0.609	(0.592, 0.625)
Llama 3.1 8B	No RAG	0.146	(0.141, 0.152)	0.391	(0.376, 0.406)	0.221	(0.215, 0.227)	0.612	(0.597, 0.628)
	RAG (4 old works)	0.124	(0.118, 0.129)	0.324	(0.309, 0.339)	0.201	(0.196, 0.207)	0.555	(0.539, 0.570)
	RAG (4 old, 4 recent)	0.115	(0.110, 0.121)	0.304	(0.289, 0.318)	0.197	(0.191, 0.202)	0.546	(0.531, 0.562)
	RAG (any 8 works)	0.117	(0.111, 0.122)	0.309	(0.294, 0.324)	0.198	(0.192, 0.203)	0.553	(0.537, 0.568)
	RAG (4 old, 4 influential)	0.116	(0.110, 0.121)	0.303	(0.289, 0.318)	0.197	(0.191, 0.202)	0.545	(0.530, 0.560)
SciLitLLM	No RAG	0.157	(0.152, 0.163)	0.475	(0.459, 0.491)	0.237	(0.232, 0.242)	0.744	(0.727, 0.760)
	RAG (4 old works)	0.133	(0.128, 0.138)	0.409	(0.393, 0.425)	0.218	(0.213, 0.223)	0.707	(0.689, 0.724)
	RAG (4 old, 4 recent)	0.133	(0.128, 0.138)	0.412	(0.396, 0.428)	0.217	(0.212, 0.222)	0.702	(0.685, 0.719)
	RAG (any 8 works)	0.134	(0.129, 0.139)	0.413	(0.396, 0.429)	0.218	(0.213, 0.223)	0.704	(0.687, 0.721)
	RAG (4 old, 4 influential)	0.138	(0.134, 0.143)	0.430	(0.414, 0.446)	0.222	(0.217, 0.227)	0.722	(0.705, 0.739)
Qwen 2.5 7B	No RAG	0.182	(0.176, 0.188)	0.507	(0.491, 0.524)	0.237	(0.231, 0.243)	0.673	(0.656, 0.689)
	RAG (4 old works)	0.171	(0.165, 0.176)	0.481	(0.465, 0.497)	0.232	(0.226, 0.238)	0.669	(0.652, 0.686)
	RAG (4 old, 4 recent)	0.164	(0.159, 0.170)	0.466	(0.450, 0.482)	0.228	(0.222, 0.233)	0.658	(0.641, 0.675)
	RAG (any 8 works)	0.170	(0.165, 0.176)	0.487	(0.471, 0.503)	0.231	(0.225, 0.236)	0.673	(0.657, 0.690)
	RAG (4 old, 4 influential)	0.166	(0.160, 0.171)	0.469	(0.453, 0.485)	0.228	(0.222, 0.234)	0.656	(0.639, 0.673)
LoRA fine-tuned Qwen variants									
Qwen (Task FT)	No RAG	0.160	(0.154, 0.166)	0.411	(0.396, 0.426)	0.245	(0.239, 0.251)	0.658	(0.642, 0.674)
	RAG (4 old works)	0.155	(0.149, 0.161)	0.396	(0.381, 0.411)	0.242	(0.236, 0.248)	0.648	(0.632, 0.663)
	RAG (4 old, 4 recent)	0.159	(0.153, 0.165)	0.407	(0.392, 0.421)	0.244	(0.238, 0.250)	0.655	(0.639, 0.670)
	RAG (any 8 works)	0.160	(0.154, 0.165)	0.407	(0.392, 0.421)	0.246	(0.240, 0.252)	0.656	(0.640, 0.671)
	RAG (4 old, 4 influential)	0.161	(0.155, 0.166)	0.412	(0.397, 0.427)	0.247	(0.241, 0.253)	0.662	(0.646, 0.677)
Qwen (Task+Reasoning FT)	No RAG	0.158	(0.152, 0.163)	0.398	(0.384, 0.413)	0.251	(0.245, 0.257)	0.663	(0.648, 0.679)
	RAG (4 old works)	0.154	(0.148, 0.160)	0.393	(0.378, 0.407)	0.244	(0.238, 0.249)	0.653	(0.637, 0.669)
	RAG (4 old, 4 recent)	0.153	(0.147, 0.159)	0.389	(0.374, 0.404)	0.243	(0.237, 0.249)	0.650	(0.635, 0.666)
	RAG (any 8 works)	0.154	(0.149, 0.160)	0.393	(0.378, 0.408)	0.245	(0.239, 0.250)	0.654	(0.638, 0.670)
	RAG (4 old, 4 influential)	0.154	(0.148, 0.160)	0.390	(0.376, 0.404)	0.252	(0.246, 0.258)	0.667	(0.652, 0.682)

Table 9: Performance of different models and retrieval settings under trend-aware evaluation, non-LOO (Leave One Out) case. Results are reported for the mean of three trials and the best of three trials. TA denotes Trend Alignment and PT denotes Progression along the Trend. Qwen (Task FT) uses a LoRA adapter fine-tuned for the idea generation task. Qwen (Task+Reasoning FT) uses the task adapter together with an additional LoRA adapter trained to generate reasoning before producing the final response.

Method		Mean of 3 Trials				Best of 3 Trials			
Model	Retrieval Case	TA		PT		TA		PT	
		Avg.	95% CI	Avg.	95% CI	Avg.	95% CI	Avg.	95% CI
gpt-4o	No RAG	0.128	(0.121, 0.134)	0.325	(0.308, 0.341)	0.185	(0.179, 0.192)	0.497	(0.480, 0.514)
	RAG (4 old works)	0.128	(0.121, 0.135)	0.321	(0.305, 0.337)	0.188	(0.182, 0.195)	0.497	(0.480, 0.514)
	RAG (4 old, 4 recent)	0.128	(0.122, 0.135)	0.326	(0.310, 0.342)	0.190	(0.184, 0.197)	0.507	(0.490, 0.524)
	RAG (any 8 works)	0.129	(0.123, 0.136)	0.328	(0.312, 0.344)	0.191	(0.184, 0.197)	0.509	(0.492, 0.526)
	RAG (4 old, 4 influential)	0.131	(0.125, 0.138)	0.332	(0.315, 0.348)	0.195	(0.188, 0.201)	0.516	(0.499, 0.533)
gpt-4o-mini	No RAG	0.122	(0.116, 0.129)	0.309	(0.291, 0.326)	0.177	(0.170, 0.184)	0.472	(0.454, 0.489)
	RAG (4 old works)	0.117	(0.110, 0.123)	0.288	(0.272, 0.305)	0.175	(0.168, 0.182)	0.456	(0.440, 0.473)
	RAG (4 old, 4 recent)	0.119	(0.112, 0.125)	0.292	(0.276, 0.308)	0.180	(0.173, 0.186)	0.467	(0.450, 0.484)
	RAG (any 8 works)	0.121	(0.114, 0.128)	0.299	(0.283, 0.315)	0.181	(0.174, 0.188)	0.471	(0.455, 0.488)
	RAG (4 old, 4 influential)	0.121	(0.115, 0.128)	0.299	(0.283, 0.315)	0.180	(0.173, 0.187)	0.466	(0.449, 0.482)
Llama 3.1 8B	No RAG	0.103	(0.097, 0.110)	0.244	(0.229, 0.260)	0.177	(0.170, 0.183)	0.457	(0.441, 0.473)
	RAG (4 old works)	0.095	(0.088, 0.101)	0.218	(0.203, 0.233)	0.170	(0.164, 0.177)	0.439	(0.423, 0.454)
	RAG (4 old, 4 recent)	0.091	(0.085, 0.098)	0.213	(0.198, 0.228)	0.170	(0.164, 0.177)	0.444	(0.429, 0.460)
	RAG (any 8 works)	0.094	(0.087, 0.100)	0.220	(0.205, 0.235)	0.172	(0.166, 0.178)	0.450	(0.434, 0.466)
	RAG (4 old, 4 influential)	0.092	(0.086, 0.099)	0.213	(0.198, 0.228)	0.171	(0.164, 0.177)	0.441	(0.425, 0.456)
SciLitLLM	No RAG	0.118	(0.112, 0.124)	0.322	(0.305, 0.338)	0.195	(0.189, 0.201)	0.575	(0.558, 0.592)
	RAG (4 old works)	0.112	(0.107, 0.118)	0.314	(0.298, 0.330)	0.194	(0.189, 0.200)	0.595	(0.578, 0.612)
	RAG (4 old, 4 recent)	0.120	(0.114, 0.126)	0.341	(0.325, 0.357)	0.201	(0.195, 0.207)	0.615	(0.598, 0.632)
	RAG (any 8 works)	0.119	(0.113, 0.125)	0.336	(0.320, 0.352)	0.200	(0.195, 0.206)	0.612	(0.595, 0.629)
	RAG (4 old, 4 influential)	0.125	(0.119, 0.130)	0.356	(0.340, 0.372)	0.206	(0.200, 0.211)	0.631	(0.614, 0.648)
Qwen 2.5 7B	No RAG	0.127	(0.120, 0.134)	0.322	(0.305, 0.339)	0.181	(0.174, 0.188)	0.479	(0.462, 0.497)
	RAG (4 old works)	0.128	(0.121, 0.134)	0.330	(0.313, 0.346)	0.188	(0.182, 0.195)	0.510	(0.492, 0.527)
	RAG (4 old, 4 recent)	0.130	(0.123, 0.136)	0.337	(0.320, 0.353)	0.191	(0.184, 0.197)	0.519	(0.502, 0.537)
	RAG (any 8 works)	0.132	(0.125, 0.138)	0.346	(0.329, 0.362)	0.191	(0.185, 0.198)	0.525	(0.508, 0.542)
	RAG (4 old, 4 influential)	0.131	(0.124, 0.137)	0.340	(0.324, 0.356)	0.191	(0.185, 0.198)	0.519	(0.502, 0.536)
LoRA fine-tuned Qwen variants									
Qwen (Task FT)	No RAG	0.108	(0.101, 0.115)	0.245	(0.230, 0.261)	0.191	(0.184, 0.198)	0.479	(0.463, 0.495)
	RAG (4 old works)	0.105	(0.099, 0.112)	0.237	(0.222, 0.252)	0.190	(0.183, 0.197)	0.475	(0.459, 0.491)
	RAG (4 old, 4 recent)	0.108	(0.101, 0.115)	0.245	(0.229, 0.260)	0.192	(0.185, 0.199)	0.480	(0.464, 0.496)
	RAG (any 8 works)	0.108	(0.102, 0.115)	0.243	(0.228, 0.258)	0.192	(0.185, 0.199)	0.478	(0.463, 0.494)
	RAG (4 old, 4 influential)	0.109	(0.103, 0.116)	0.249	(0.233, 0.264)	0.193	(0.186, 0.199)	0.483	(0.467, 0.499)
Qwen (Task+Reasoning FT)	No RAG	0.106	(0.100, 0.113)	0.238	(0.223, 0.253)	0.198	(0.191, 0.205)	0.491	(0.475, 0.506)
	RAG (4 old works)	0.107	(0.100, 0.113)	0.243	(0.228, 0.258)	0.194	(0.188, 0.201)	0.492	(0.476, 0.508)
	RAG (4 old, 4 recent)	0.105	(0.098, 0.112)	0.237	(0.222, 0.252)	0.192	(0.186, 0.199)	0.485	(0.468, 0.501)
	RAG (any 8 works)	0.107	(0.100, 0.114)	0.242	(0.227, 0.257)	0.194	(0.187, 0.201)	0.488	(0.472, 0.504)
	RAG (4 old, 4 influential)	0.107	(0.100, 0.113)	0.239	(0.224, 0.254)	0.201	(0.195, 0.208)	0.502	(0.486, 0.517)

Table 10: Performance of different models and retrieval settings under leave-one-out trend-aware evaluation, LOO (Leave One Out) case. Results are reported for the mean of three trials and the best of three trials. TA denotes Trend Alignment and PT denotes Progression along the Trend. Qwen (Task FT) uses a LoRA adapter fine-tuned for the idea generation task. Qwen (Task+Reasoning FT) uses the task adapter together with an additional LoRA adapter trained to generate reasoning before producing the final response.

Method	Generated idea	TA	PT
Cluster: LLM Watermarking, Attacks, and Robust Decoding *Background:* Large language models can be misused, and simply detecting machine-generated text is often insufficient for response and accountability. Prevailing watermarking methods are largely zero-bit, providing only binary detection signals that do not support tracing specific adversarial users.			
Qwen2.5-7B-Instruct	Hypothesis: Embedding unique, cryptographically secure identifiers within large language model outputs can enable both detection and traceability of adversarial users, enhancing accountability in machine-generated content.	0.154	0.226
Llama-3.1-8B-Instruct	Hypothesis: Adversarially robust, multi-bit watermarking of large language models can be achieved through the strategic incorporation of a hidden, gradient-encoded "fingerprint" that adapts to user behavior, allowing for both detection and attribution of malicious content.	0.344	0.529
GPT-4o	Embedding multi-bit digital watermarks into large language model outputs can enable the tracing of individual users by encoding unique identifiers, thus enhancing accountability without significantly degrading text quality.	0.472	0.708
Future Ground Truth	Hypothesis: Allocating token positions across a generated sequence to independently encode sub-units of a multi-bit watermark enables robust, model-access-free recovery of long identifiers under corruption without finetuning or added latency, while preserving text quality and retaining zero-bit detectability.	0.505	0.735
Cluster: Robust Evaluation and Benchmarking of LLMs *Background:* LLM-based judges are increasingly used as scalable substitutes for human evaluation, but their reliability on tasks demanding factual accuracy and complex reasoning, such as knowledge, math, and coding, is underexamined, while existing benchmarks emphasize crowd-aligned preferences that can fail to capture objective correctness.			
GPT-4o (trial 1)	In tasks requiring factual accuracy and complex reasoning, LLM-based judges exhibit improved reliability when supplemented with modular expert systems tailored for specific domains such as knowledge retrieval, mathematical computation, and code analysis.	0.548	0.970
GPT-4o (trial 2)	LLM-based judges' ability to evaluate tasks demanding factual accuracy and complex reasoning improves significantly when supplemented with an external verification system that cross-references with authoritative databases.	0.463	0.723
GPT-4o (trial 3)	LLM-based judges can improve their reliability in tasks requiring factual accuracy and complex reasoning by incorporating a multi-model consensus approach, where outputs from multiple independent LLMs are aggregated to enhance decision-making accuracy.	0.277	0.388
Future Ground Truth	Hypothesis: Converting challenging knowledge, reasoning, math, and coding instances into paired responses labeled by objective correctness produces a more rigorous, discriminative benchmark for assessing the reliability of LLM-based judges than benchmarks based on crowdsourced preferences.	0.592	0.851
Cluster: Legal NLP for Case Retrieval, Judgment Prediction and Explainability *Background:* Legal case outcomes frequently exhibit split judicial votes that reflect inherent human label variation, but common NLP calibration approaches evaluate models only against the majority label and thus miss this variation, limiting reliable assessment of model trustworthiness in high-stakes legal settings.			
Qwen2.5-7B-Instruct	A novel hypothesis: Evaluating NLP models using a range of metrics that account for both majority and minority judicial votes will provide a more accurate assessment of model trustworthiness in legal settings with split votes.	0.340	0.633
SciLitLLM	A novel approach that measures model trustworthiness by considering both majority and minority labels can reduce bias and improve performance of models used in high-stakes legal settings.	0.191	0.421
GPT-4o-mini	Incorporating minority label perspectives into model evaluation metrics will enhance the trustworthiness of NLP systems in legal case outcome predictions by more accurately reflecting judicial vote variability.	0.186	0.363
Future Ground Truth	Hypothesis: Assembling judges' vote distributions into the SVA-ECHR dataset, creating a taxonomy of split-vote subtypes, and evaluating case-outcome classifiers against the full judicial vote distribution rather than the majority label will reveal misalignment between model confidence and judicial disagreement and expose calibration gaps.	0.229	0.499

Table 11: Representative generated hypotheses and leave-one-out Trend Alignment (TA) and Progression (PT) scores.

Table 12: Qualitative example comparing No-RAG and RAG generations under TrACE. For retrieved contexts, we show short quoted phrases from retrieved abstracts and their source papers to illustrate how retrieval can influence the generated hypothesis. Scores are computed under the leave-one-out setting. TA denotes Trend Alignment and PT denotes Progression along the Trend.

Method / Source	Hypothesis or Retrieval Signal	TA	PT
Cluster: Multimodal Sentiment Analysis and Fusion. *Background:* Selecting and combining feature encoders from heterogeneous modalities for sentiment detection remains challenging, with open questions about optimal fusion methods, multi-loss training effects, and how to assess individual subnet contributions and contextual influence.			
Future Ground Truth	Hypothesis: Training modality-specific feature encoders and learning their fusion within a single neural architecture, using multi-loss training and an auxiliary contextual representation, improves sentiment detection and clarifies subnet contributions, achieving state-of-the-art performance on CMU-MOSI, CMU-MOSEI, and CH-SIMS.	0.395	0.887
GPT-4o (No RAG)	Incorporating attention mechanisms that dynamically weigh feature encoders based on contextual sentiment shifts can enhance the accuracy and robustness of sentiment detection across heterogeneous modalities.	0.176	0.375
Retrieved signal: RAG (4 old works)	Retrieved abstracts include Pedregosa et al. (2011), describing *Scikit-learn* as a Python module integrating a "wide range of state-of-the-art machine learning algorithms"; and Lakshminarayanan et al. (2016), proposing an alternative to Bayesian neural networks that is "simple to implement" and yields "high quality predictive uncertainty estimates." The retrieved context also emphasizes "well-calibrated uncertainty estimates" and higher uncertainty on "out-of-distribution examples."	–	–
GPT-4o (RAG, 4 old works)	Incorporating predictive uncertainty estimates from non-Bayesian neural networks into the fusion process of heterogeneous feature encoders will enhance sentiment detection accuracy by dynamically weighting contributions from each modality based on their confidence levels.	0.154	0.377
Retrieved signal: RAG (4 old, 4 influential)	Retrieved abstracts further emphasize uncertainty and calibration: Ovadia et al. (2019) discuss "predictive uncertainty" under "dataset shift" and evaluate "accuracy and calibration"; Kuleshov et al. (2018) propose "calibrating any regression algorithm" to produce "well-calibrated credible intervals"; Han et al. (2022) introduce models for "conditional distribution prediction" and "instance-level confidence assessment"; and Rahaman et al. (2020) discuss "calibration issues", "deep-ensembles", and "Expected Calibration Error."	–	–
GPT-4o (RAG, 4 old, 4 influential)	In sentiment detection tasks using heterogeneous modalities, employing a dynamic fusion strategy that adaptively adjusts the contribution of each modality based on real-time uncertainty estimates can improve model calibration and performance in variable input conditions.	0.317	0.765