Yale THE UNIVERSITY OF CHICAGO

Measuring the Gap Between Human and LLM Research Ideas

Ziyu Chen [C] **Yilun Zhao** [Y] **Arman Cohan** [Y]
[Y] Yale University [C] University of Chicago

ziyuuc/TasteGap IdeaLand/IdeaSeed

Abstract

LLMs are increasingly used to brainstorm research ideas, but existing evaluations mostly judge individual ideas by novelty, feasibility, or expert preference. We instead ask: how far are current LLM-generated ideas from human researchers? To characterize this gap, we build a large-scale evaluation framework for ideation from high-quality human research papers. For each paper, we reverse-engineer a small set of closely related prior works that likely inspired its core idea. LLMs are then prompted to generate a new idea from the set of paper titles and summaries. We introduce a two-axis research-taste taxonomy to profile each idea by its opportunity pattern and research paradigm, and use it to quantify the divergence between human and LLM ideas. Across idea sets generated by different LLMs, we observe a consistent distributional gap: LLM ideas are disproportionately concentrated around bridge-like opportunities and synthesis methods, whereas the human paper reference distribution spreads more broadly across ways of framing gaps and constructing contributions. This result suggests that strong LLMs can produce a range of reasonable ideas, but that range remains narrower than, and systematically shifted relative to, human research taste.

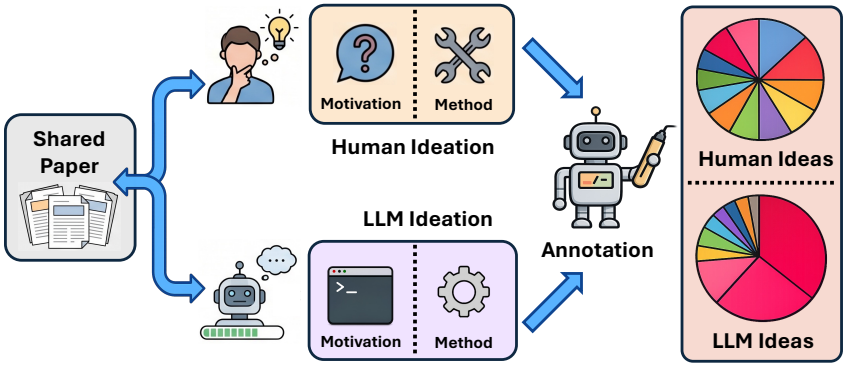

Figure 1: Overview of our research-taste gap analysis. From a shared literature context, humans contribute the paper idea while LLMs generate new ideas from the same prior works. Each idea is decomposed into a *motivation* and a *method*, then annotated with a research-taste taxonomy. Comparing the resulting distributions reveals that LLM ideas are substantially narrower than human ideas, with strong biases toward bridge-like motivations and explicit synthesis methods.

1 Introduction

Research ideation is one of the most ambitious proposed uses of LLMs. Some controlled human studies have shown that LLM-generated ideas can match or approach those from human experts in terms of judged novelty and feasibility, demonstrating the potential of LLMs as ideation tools [Si et al., 2025a, Baek et al., 2025, Si et al., 2025b]. More recently, AI-scientist systems have already begun adapting LLMs to generate research ideas, execute experiments, and write paper drafts, bringing this capability into scientific process [Boiko et al., 2023, Lu et al., 2024, Zhang et al., 2024, Zhao et al., 2025, Vasu et al., 2025, Zhao et al., 2026, Wu et al., 2026]. Despite this progress, a basic empirical question remains unanswered: *What kinds of research ideas do LLMs tend to produce, and how far are current LLM-generated ideas from human researchers?*

Most existing evaluations of LLM ideation judge ideas individually, using criteria such as novelty, feasibility, impact, or preference [Si et al., 2025a, Baek et al., 2025, Garikaparthi et al., 2025, Tong et al., 2026]. In this work, we take a complementary distributional view. We use *research taste* to refer to the kinds of problems, gaps, and contributions that a source tends to produce across many comparable literature-grounded ideation contexts. Under this view, research taste concerns not only whether an individual idea is reasonable, but also what kinds of gap framings and contribution strategies repeatedly appear when the same source is asked to generate ideas under comparable constraints.

This distributional view matters because a single idea may appear novel, feasible, and coherent, while the broader set of ideas from the same source may still reflect a narrow range of research taste. Research communities generate many kinds of contributions: some papers discover a failure mode, some relax an assumption, some build a measurement instrument, some introduce a formal explanation, and others construct a system or artifact. An LLM that generates reasonable ideas one at a time may therefore still be behaviorally narrow if its outputs repeatedly identify the same kinds of gaps, use the same methodological paradigms, or rely on the same contribution templates [Meincke et al., 2024, Smith et al., 2025, Sourati et al., 2026]. Such concentration would affect how LLMs are used for brainstorming, literature exploration, and automated research agents, even when many individual outputs appear coherent.

We study this question through a constrained literature-grounded ideation task. Each example consists of a small set of closely related prior papers, represented by their titles and abstracts. The target output is a new research idea, separated into a *motivation* and a *method*. This differs from open-ended ideation such as *write an idea about topic X*. Grounding each generation in a small related-work context makes the task comparable across human and LLM outputs. The human idea is the idea realized in the real paper, while the LLM idea is generated from the reconstructed local context that seemingly reasonably preceded that paper. This setting also reduces the chance that differences are driven only by broad topic choice or by an LLM's preferred generic paper format.

To build the evaluation framework, we collect real papers in machine learning and natural science domains. For each paper, we use a strong LLM-assisted extraction pipeline to identify the paper's core human idea and to reverse-engineer several closely related prior works from which that idea can be understood. We retrieve these prior works and prompt evaluated LLMs to produce a new motivation and method from the same prior-work context. This yields paired human and LLM idea corpora over the same inputs.

We compare these corpora at the distributional level through a two-dimensional view of research taste. One dimension characterizes how a proposal frames the underlying research opportunity, ranging from identifying missing explanations or overlooked failures to exposing structural disconnects or limitations in existing understanding. The other captures the style of intellectual contribution through which the opportunity is developed into a research idea, including analytical, constructive, integrative, and exploratory forms of proposed methods. We introduce a taxonomy along these two axes, constructed by human experts through a review of research guidance from NSF, NIH, AHRQ and DARPA, and then iteratively refined using a held-out set of papers to ensure applicability across both machine learning and natural science domains. We apply the taxonomy at scale using an LLM annotator validated against independent human judgments.

We find that LLM-generated ideas occupy a substantially narrower region of the research-taste taxonomy than human ideas. This narrowing is most visible in an ideation pattern centered on connection, where model ideas more often frame the motivation as a need to link previously literatures, methods, or evidence streams, and more often develop the method by integrating, reconciling, or unifying existing approaches. In our evaluation, only 12.1% of human ideas motivated by the pattern of connection, and only 5.1% use synthesis or unification as the central method paradigm. By contrast, across the nine main evaluated LLMs, the corresponding rates range from 47.1% to 64.2% and from 22.5% to 38.7%, respectively. Human ideas also exhibit consistently higher normalized entropy on both taxonomy axes. This pattern remains stable across model families and scientific domains, indicating that current LLM ideation is disproportionately concentrated around integrative and synthesis-oriented types, while human research ideas span a substantially broader range of opportunity patterns and methodological paradigms.

2 Related Work

LLMs for Research Ideation. Recent work has explored LLMs for scientific ideation, including generating, refining, and evaluating research hypotheses and directions. Early studies show that directly prompted LLMs can produce ideas perceived as highly novel, though often less feasible or well-grounded than human proposals [Si et al., 2025a]. Building on this, subsequent work has explored iterative refinement, retrieval-augmented generation, and search-based ideation pipelines [Wang et al., 2024, Baek et al., 2025, Sanyal et al., 2025]. Other approaches ground generation in external scientific signals such as retrieved literature, knowledge graphs, or emerging research trends [Ghafarollahi and Buehler, 2024, Hu et al., 2024, Pu et al., 2025]. Multi-agent collaboration has also become a common paradigm for improving idea diversity and critique through simulated scientific discussion [Gu et al.,

2024, Su et al., 2025]. Benchmarks are proposed to evaluate dimensions including novelty, feasibility, and impact of generated ideas [Ruan et al., 2026, Guo et al., 2025a, Liu et al., 2026]. Our work differs from this line of research by studying whether the overall *distribution* of LLM-generated ideas resembles the distribution of ideas realized in human-written scientific papers.

Gaps between Human and LLM-Generated Content. Even when LLM outputs appear fluent and useful, research shows that they can differ systematically from human outputs. Early detection work found that neural generations contain statistical artifacts in token ranks, sampling behavior, and likelihood geometry that distinguish them from human text [Gehrmann et al., 2019, Ippolito et al., 2020, Mitchell et al., 2023]; broader comparison corpora and distributional metrics similarly show measurable gaps between ChatGPT or neural text and human expert writing [Pillutla et al., 2021, Guo et al., 2023]. These gaps become more consequential when LLM outputs are used as substitutes for human populations or human judgments. In social simulation, LLMs can reproduce some aggregate patterns while still exhibiting distortions or demographic misalignment relative to real human responses [Aher et al., 2023, Santurkar et al., 2023]. In evaluation and review settings, LLM judges vary substantially across tasks and require validation against human annotations, while LLM-generated paper reviews over-focus on technical validity and under-attend to novelty compared with expert reviewers [Bavaresco et al., 2025, Shin et al., 2025]. Our work follows this perspective of comparison between human and LLMs, shifting the target to the distribution of research ideas.

3 Evaluation Framework for Ideation

We propose an idea-level evaluation framework to systematically analyze what kinds of research opportunities and contribution strategies LLMs emphasize or overlook during scientific ideation. We first define a literature-grounded ideation task and construct paired human and LLM idea corpora accordingly. We represent each idea through its motivation and method, annotate these ideas with a research-taste taxonomy, and analyze.

3.1 Literature-Grounded Ideation Task

Each instance contains a set of related prior works $X_i = \{(t_{i1}, a_{i1}), \ldots, (t_{ik}, a_{ik})\}$, with t denoting the title and a the abstract. The target output is a research idea $y_i = (m_i, s_i)$, where m_i refers to the motivation and s_i stands for the proposed method. Guided by the provided literature context, the prompt directs models to identify research gaps across papers and generate a coherent research idea. This task is constrained. In an open-ended ideation setting, differences between human and model outputs can be confounded by topic selection, prior knowledge, and generic paper-writing templates, making the comparison difficult to interpret [Si et al., 2025a, Ruan et al., 2026]. By anchoring both human and model ideas to the same set of related prior works, we focus the analysis on how each source identifies research gaps and constructs contributions from a shared local context.

3.2 Idea Corpus

Human Idea. We build our evaluation data from published research papers drawn from two sources: machine learning conference proceedings from ICLR, ICML and NeurIPS published between 2023 and 2026 and Nature Communications released from 2023 to 2025, covering 71 major scientific disciplines such as physics, chemistry, and biology. For each paper, the work itself represents the human endpoint, which is the research idea originally devised by authors for publication. We extract this idea into a structured representation using an LLM-assisted pipeline. The extraction prompt asks for the innovation, departure from prior work, and key insight, then rewrites the result into a proposal-style motivation and method. We then reverse-engineer 4 to 8 highly relevant prior studies based on the extracted idea and the paper's related-work section. For each related work, we retrieve the title and abstract. The final input contains only the prior-work titles and abstracts as main idea summaries. The mixed corpus contains 11,683 valid human ideas in total. Full prompts and additional data details are provided in Appendix A.

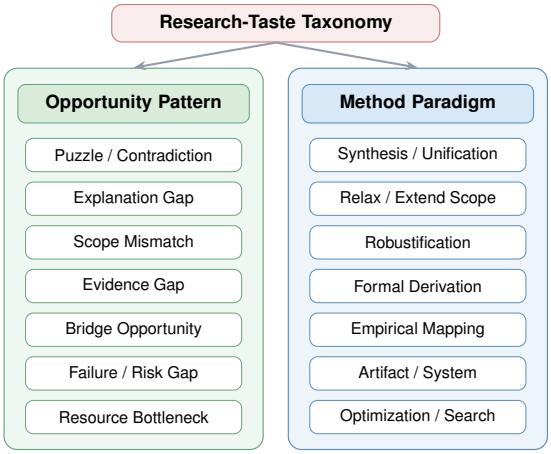

Figure 2: The two-axis research-taste taxonomy. The opportunity pattern axis labels *why* a new study is needed and captures a research gap, while the method paradigm axis labels *how* the proposed work turns that gap into a contribution.

LLM Idea. For each input, we prompt LLMs to generate a new research idea in the same structured format with two components. The motivation synthesizes research gaps across the provided papers, and the method outlines a concrete high-level approach. We evaluate a set of mainstream LLM families including `Claude`, `Gemini`, `GPT`, `DeepSeek` and `Qwen`.

3.3 Research-Taste Taxonomy

We label each idea with a two-axis research-taste taxonomy matching an idea's structure of *motivation* and *method*. The *opportunity pattern* axis corresponds to the motivation: it asks what kind of gap makes the proposal worth pursuing, such as a contradiction, missing explanation, scope mismatch, evidence gap, disconnected literature, failure risk, or resource bottleneck. The *method paradigm* axis corresponds to the method: it asks what high-level contribution strategy turns that gap into a paper, such as synthesis, scope extension, robustification, formal derivation, empirical mapping, artifact construction, or optimization.

We designed this taxonomy to support interpretable distributional comparison. We first reviewed guidance from NSF, NIH, AHRQ, and DARPA on research proposals and problem formulation. From these sources, we extracted recurring elements organized around two broad dimensions: the opportunity that motivates a study and the method by which the study addresses it. This initial taxonomy contained 11 opportunity elements and 9 method elements. We refined these candidate elements using a held-out set of 150 papers. For each paper, we allowed up to two closest labels on each axis and included an *other* option for cases not covered by the initial taxonomy. We then examined coverage and label distributions, merged near-duplicate labels, split labels that conflated motivation and method, removed labels tied to a specific domain or technical substrate, and renamed categories using terms applicable across ML and natural-science papers. The final taxonomy contains seven opportunity patterns and seven method paradigms as shown in Figure 2. Our final taxonomy meets three requirements: categories correspond to recurring gap framings and contribution strategies, generalizable across different research domains, and were checked during human validation to avoid systematic category collapse. This ensures frequency discrepancies stem from distinct research tendencies, not field differences or wording variations. We provide details in Appendix B.

3.4 Automated Annotation

We use an LLM as an automated annotator to label every human and LLM idea [Zheng et al., 2023, Liu et al., 2023, Kim et al., 2024]. The annotator receives the taxonomy, the prior-work titles for context, and the proposal motivation and method. It returns a primary and secondary label for each axis, confidence scores, and three diagnostic scores: surface stitching, bottleneck specificity, and boilerplate. We use primary labels for distributional comparisons and diagnostic scores for mechanism analyses.

Before applying the annotator at scale, we validated it on the same held-out set of 150 papers used for taxonomy calibration. Two authors examined ideas from this set and audited three outputs from each annotation pass including the opportunity-pattern label, the method-paradigm label, and the diagnostic-score profile. As a label-level reliability measure, we compute Cohen's κ [Cohen, 1960] between the `GPT-5.4-mini`[1] annotation and each author's judgment, then average over the two human-LLM pairs. The resulting κ are 0.84, 0.81, and 0.93, respectively. We also inspect disagreements and confusion matrices to ensure that errors concentrate in semantically adjacent labels rather than reflecting systematic category collapse. Given this high agreement, we use `GPT-5.4-mini` as the automatic annotator for evaluation.

4 Experiments

Our experiments ask whether LLMs generate the same *kinds* of research ideas as human researchers when both are conditioned on the same local literature context. We first compare human and model label distributions on the two taxonomy axes. We then use the diagnostic scores to examine whether the same gap appears as lower specificity or more template-like proposals. Finally, we study how reasoning changes the distribution and use mechanism analyses to probe why the gap appears.

4.1 Setup

Data. We evaluate on the data introduced in Section 3.2. The full evaluation set contains 11,683 ground-truth human ideas from diverse fields mixed together, one per source paper, matched to generations from each evaluated model. We present overall results, while detailed analyses and results for each individual domain are provided in Section E.1.

[1]We chose `GPT-5.4-mini` as it provided an appropriate balance between quality and cost.

Table 1: Distributional distance between human and LLM ideas on our evaluation set. **Opportunity Pattern** captures why the idea is worth pursuing, while **Method Paradigm** captures how the proposal turns that gap into a contribution. Ent. is normalized entropy. All model distributions remain far from the human distribution, especially on the opportunity axis. ■ marks the best non-human score, and ■ marks the clearest model-side degradation.

	Opportunity Pattern			Method Paradigm		
Source	TVD ↓	JSD ↓	Ent. ↑	TVD ↓	JSD ↓	Ent. ↑
Human	—	—	0.926	—	—	0.920
`Claude-Sonnet-4.6`	0.351	0.130	0.737	**0.211**	**0.070**	**0.879**
`Gemini-3.1-Pro`	**0.348**	**0.128**	**0.758**	0.227	0.092	0.874
`GPT-OSS-20B`	0.456	0.218	0.598	0.378	0.158	0.723
`GPT-OSS-120B`	0.521	0.259	0.550	0.391	0.170	0.735
`GPT-5.4-mini`	0.512	0.243	0.568	0.339	0.119	0.814
`Qwen3-8B`	0.382	0.179	0.658	0.368	0.190	0.734
`Qwen3-32B`	0.417	0.191	0.640	0.364	0.183	0.745
`DeepSeek-V4-Flash`	0.400	0.167	0.683	0.246	0.086	0.845
`DeepSeek-V4-Pro`	0.436	0.208	0.642	0.258	0.108	0.828

Models. The main comparison includes nine model settings: `Claude-Sonnet-4.6`, `Gemini-3.1-Pro`, `GPT-OSS-20B`, `GPT-OSS-120B`, `GPT-5.4-mini`, `Qwen3-8B`, `Qwen3-32B`, `DeepSeek-V4-Flash`, and `DeepSeek-V4-Pro` [Anthropic, 2026, Google, 2026, OpenAI, 2025, 2026, Yang et al., 2025, DeepSeek-AI, 2026]. For the reasoning ablation, we additionally evaluate `Qwen3-8B` and `DeepSeek-V4-Flash` with thinking mode. Each model is prompted with the same input and asked to produce a structured idea.

Metrics. For each source and taxonomy axis, we estimate the empirical distribution of primary labels. We use three distributional measures to characterize how model-generated ideas differ from human ideas. Total Variation Distance (TVD) measures the amount of label mass that would need to be reallocated for a model distribution to match the corresponding human distribution. Jensen-Shannon Divergence (JSD) [Lin, 1991] provides a bounded and symmetric measure of divergence between the model and human distributions. And normalized entropy, which measures how concentrated a source is over taxonomy labels. We report TVD as,

$$\mathrm{TVD}(\hat{P}, \hat{Q}) = \frac{1}{2} \sum_{c \in A} |\hat{P}(c) - \hat{Q}(c)|,$$

where $\hat{P}$ and $\hat{Q}$ denote empirical label distributions over the label set A, and c indexes individual labels. We report JSD using base-2 logarithms,

$$\mathrm{JSD}(\hat{P}, \hat{Q}) = \frac{1}{2}\mathrm{KL}(\hat{P} \| M) + \frac{1}{2}\mathrm{KL}(\hat{Q} \| M),$$

where $M = (\hat{P} + \hat{Q})/2$. Normalized entropy over the labels

$$H_{\mathrm{norm}}(\hat{P}) = -\frac{1}{\log_2 |A|} \sum_{c \in A} \hat{P}(c) \log_2 \hat{P}(c).$$

Lower normalized entropy indicates greater concentration over a smaller set of research moves [Sajjadi et al., 2018, Pillutla et al., 2021].

4.2 Main Distributional Gap

Figure 3 and Table 1 show a consistent distributional gap across tested model families. Human ideas have high normalized entropy on both taxonomy axes, above 0.92. Model distributions are generally more concentrated than the human reference distribution, especially on the opportunity axis: opportunity entropy ranges from 0.550 to 0.758, while method-paradigm entropy ranges from 0.723 to 0.879. Even the closest model on the opportunity axis, `Gemini-3.1-Pro`, has TVD 0.348, meaning that over a third of the distributional mass would need to move to match human outputs. On the method axis, `Claude-Sonnet-4.6` is closest, but its TVD remains 0.211.

The largest gap is a shift toward bridge-and-synthesis ideas. Only 12.1% of human opportunities are labeled as fragmentation or bridge opportunities, compared with 47.1 to 64.2% for the main LLMs. The same pattern appears on the method axis: explicit synthesis or unification accounts for 5.1% of human ideas but 22.5 to 38.7% of LLM ideas. This does not mean that synthesis is never a valid research contribution. Rather, it shows that when models are given nearby papers, they frequently turn ideation into a generic move of connecting or combining

Table 3: Diagnostic scores from the automatic annotator. The annotator assigns three ordinal 0 to 3 ratings from the proposal motivation and method: surface stitching measures whether the idea is a superficial combination, bottleneck specificity measures whether it identifies a precise mechanism or limiting factor, and boilerplate measures generality. ■ marks the clearest model-side degradation, and ■ marks the best non-human diagnostic scores.

Source	Surf. Score ↓	Surf. Flag (%) ↓	Bottleneck ↑	Boilerplate ↓
Human	**0.00**	**0.0**	**2.56**	**0.48**
`Claude-Sonnet-4.6`	**0.02**	**0.1**	**2.60**	**0.37**
`Gemini-3.1-Pro`	0.09	0.4	2.34	0.79
`GPT-OSS-20B`	0.09	1.1	2.07	0.97
`GPT-OSS-120B`	0.07	0.3	2.16	0.87
`GPT-5.4-mini`	0.02	0.1	2.21	0.75
`Qwen3-8B`	0.58	20.6	1.76	1.25
`Qwen3-8B-Think`	0.45	11.0	1.90	1.11
`Qwen3-32B`	0.44	13.7	1.87	1.15
`DeepSeek-V4-Flash`	0.10	1.2	2.12	0.92
`DeepSeek-V4-Flash-Think`	0.10	0.7	2.12	0.89
`DeepSeek-V4-Pro`	0.04	0.2	2.34	0.69

prior work, while human papers distribute more mass to explanation, measurement, risk, scope, artifacts, and optimization-style contributions. The mechanism analyses in Section 4.5 ask whether this gap reflects many diverse forms of synthesis or a narrower template for constructing proposals. We also test alternative prompts, and this preference for organizing new ideas through bridging and synthesis does not appear sensitive to the specific wording of the generation prompt. Details in Section E.3.

Full-paper context ablation. To better approximate richer ideation settings, we also evaluate a full-paper context condition in which models can draw on the original related-work papers rather than only abstract summaries. In this setting, for each related work paper, LLMs first read the full text and produce their own summary; these model-generated summaries are then used in place of abstracts as the input context for idea generation. We run this ablation on a subset sampling 500 papers from each of the two domains and evaluate `Qwen3-8B` and `DeepSeek-V4-Flash`. Table 2 shows that richer context does not move either model closer to the human reference distribution. For both models, TVD and JSD increase on both taxonomy axes under full-paper summaries, while entropy decreases or remains nearly unchanged. The full label distributions in Section E.2 show the same qualitative pattern: even when models are given a richer representation of the related literature, their generated ideas continue to exhibit the same preference for organizing new proposals through bridging and synthesis.

Table 2: Full-paper context ablation on a 1,000 paper subset. Full context replaces abstracts with model-generated full-paper summaries with detailed *motivation*, *method* and *insight*. TVD / JSD are computed against the human reference distribution, and Ent. is normalized entropy. Full label counts are in Table 10.

		Opportunity Pattern			**Method Paradigm**		
Model	Context	TVD ↓	JSD ↓	Ent. ↑	TVD ↓	JSD ↓	Ent. ↑
`Qwen3-8B`	Abstract	0.376	0.165	0.669	0.338	0.182	0.752
`Qwen3-8B`	Full	0.430 (+.054)	0.205 (+.040)	0.623 (-.046)	0.400 (+.062)	0.229 (+.047)	0.699 (-.053)
`DeepSeek-V4-Flash`	Abstract	0.368	0.152	0.706	0.213	0.079	0.867
`DeepSeek-V4-Flash`	Full	0.400 (+.032)	0.160 (+.008)	0.701 (-.005)	0.236 (+.023)	0.093 (+.014)	0.860 (-.007)

4.3 Diagnostic Scores

We next analyze the three diagnostic scores defined in Section 3.4 and reported in Table 3: surface stitching, bottleneck specificity, and boilerplate. These annotator-assigned scores characterize the specificity and template-likeness of each proposal. Surface stitching measures whether the idea is a superficial combination of prior work, bottleneck specificity measures whether it identifies a precise mechanism or limiting factor, and boilerplate measures generic phrasing. In Table 3, most model outputs receive lower specificity and higher boilerplate scores than human ideas, with especially strong degradation for the `Qwen` models, which also have the highest surface-stitching scores and flags. `Claude-Sonnet-4.6` is an exception on these diagnostic dimensions: it has slightly higher bottleneck specificity and lower boilerplate than the human baseline, while still remaining distributionally shifted in Table 1. Taken as supporting evidence, these annotator-based diagnostic scores suggest that the research-taste gap is not

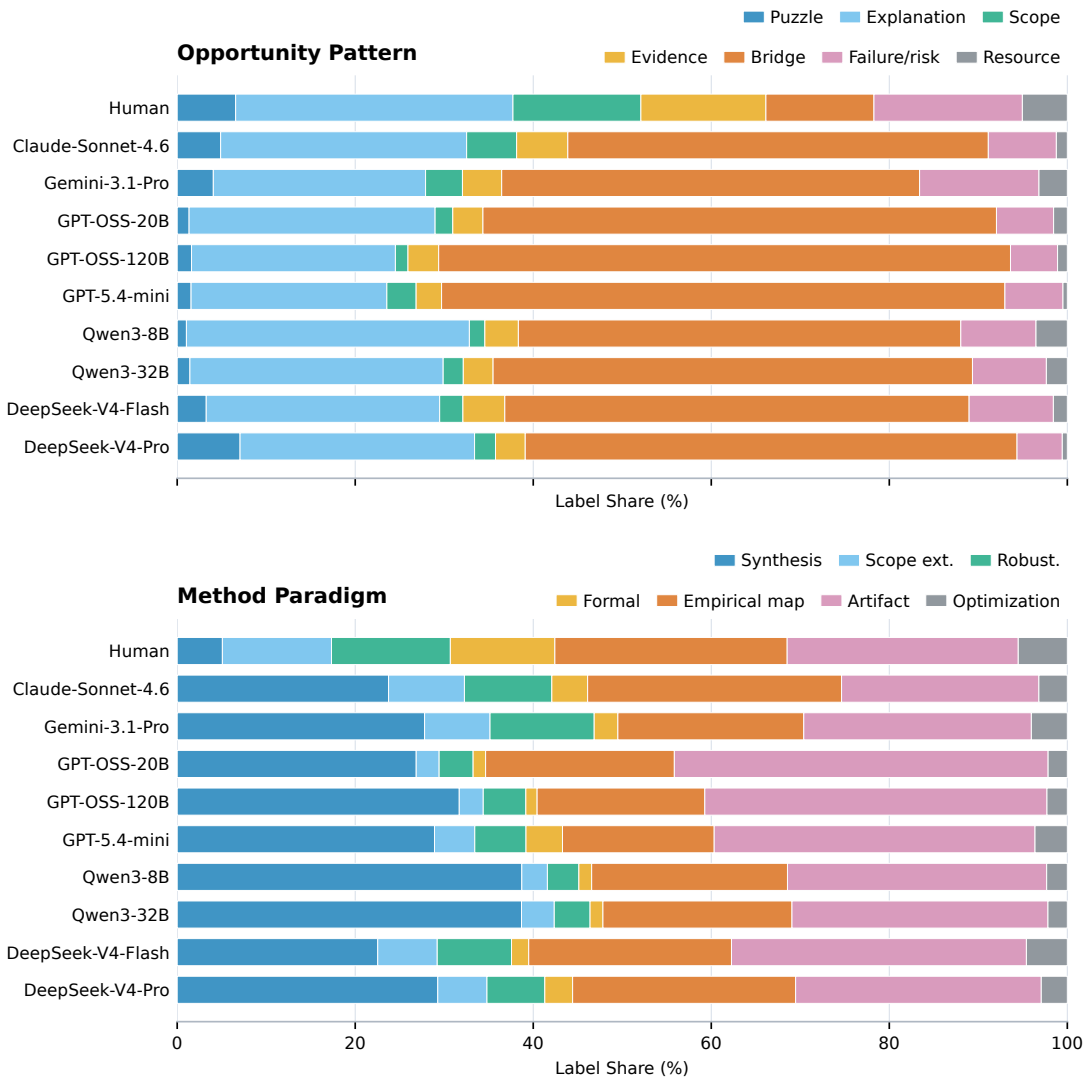

Figure 3: Full label distributions over opportunity patterns and method paradigms. On the opportunity axis, LLMs strongly amplify fragmentation and bridge opportunities (■), where the research gap is framed as disconnected literatures, methods, or evidence streams that should be connected. On the method axis, the largest shift is toward explicit synthesis or unification (■), where the contribution is constructed by integrating or reconciling separate ideas. Thinking-mode outputs are shown for `Qwen3-8B` and `DeepSeek-V4-Flash`.

only a low-level quality issue; even polished and specific model outputs can concentrate on a narrower set of opportunity and method patterns. Surface stitching is a stricter local judgment, while bridge and synthesis labels describe the broader research type. Together, however, they support the same interpretation: LLM ideation often converges on polished combinations of prior work, and for many models those combinations are also less specific. This interpretation is consistent with the archetype analysis in Section 4.5, where model-heavy clusters have lower bottleneck specificity and higher boilerplate than human-heavy clusters built around replacing or decoupling a local mechanism [Gehrmann et al., 2019, Ippolito et al., 2020].

4.4 Does Extended Reasoning Help?

Reasoning is considered a paradigm that can enhance the downstream capabilities of models [Kojima et al., 2022, Wei et al., 2022, Guo et al., 2025b]. However, in our ideation tasks, thinking mode moves the output distribution farther from the human reference for both model settings we test as shown in Table 4. For `Qwen3-8B`, enabling thinking increases bridge opportunities from 49.7% to 71.1% and explicit synthesis from 38.7% to 52.2%. Opportunity entropy drops from 0.658 to 0.481, and TVD from humans increases from 0.382 to 0.590. This adverse effect persists across both weaker and stronger models we tested. The same direction appears for `DeepSeek-V4-Flash`: bridge opportunities rise from 52.2% to 59.1%, synthesis rises from 22.5% to 30.7%, and both opportunity and

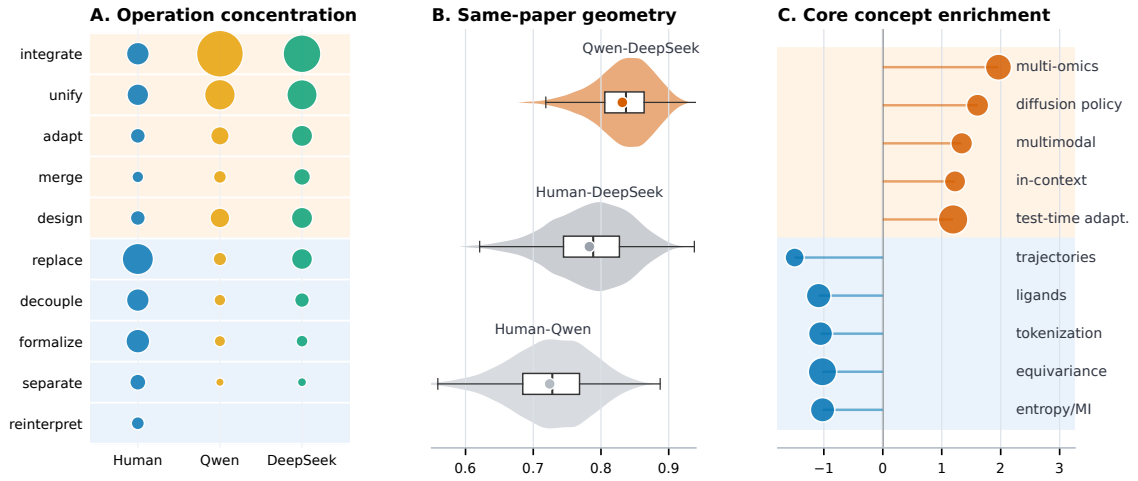

Figure 4: Mechanism analyses for the distributional gap. (A) Circle areas show archetype operation shares by source: model outputs concentrate on reusable template moves such as *integrate* and *unify*, while human ideas retain more local interventions such as *replace*, *decouple*, and *formalize*. (B) Same-paper similarity distributions show that the two model ideas are closer to each other than either is to the human idea. (C) Core concept enrichment ranks clusters by model-vs-human log-odds, with circle area indicating support. Model-heavy concepts are high-frequency technical motifs, whereas human-heavy concepts are narrower mechanism or representation clusters.

method TVD increase. Thinking therefore appears to sharpen the model's preferred ideation template instead of broaden the distribution toward human taste, and further reduces the diversity of generated ideas.

Table 4: Comparison between models with and without reasoning enabled, reporting changes relative to the corresponding base model. Enabling thinking consistently increases bridge and synthesis mass, increases TVD from the human distribution, and lowers entropy, indicating a sharper and less human-like idea distribution.

	Template Mass		Distance to Human				Diagnostics	
Setting	Bridge ↓	Synthesis ↓	Opp. TVD ↓	Opp. Ent. ↑	Meth. TVD ↓	Meth. Ent. ↑	Surface ↓	Boiler. ↓
`Qwen3-8B`	49.7	38.7	0.382	0.658	0.368	0.734	0.58	1.25
`w/ think`	71.1 (+21.4)	52.2 (+13.5)	0.590 (+.208)	0.481 (-.177)	0.472 (+.104)	0.649 (-.085)	0.45 (-.13)	1.11 (-.14)
`DeepSeek-V4-Flash`	52.2	22.5	0.400	0.683	0.246	0.845	0.10	0.92
`w/ think`	59.1 (+6.9)	30.7 (+8.2)	0.470 (+.070)	0.620 (-.063)	0.291 (+.045)	0.823 (-.022)	0.10 (+.00)	0.89 (-.03)

4.5 Mechanism Analyses

The taxonomy tells us *where* model and human distributions differ, but it does not fully explain *why* LLMs repeatedly choose bridge-and-synthesis moves. We therefore run follow-up analyses on human ideas and two representative model sources, `Qwen3-8B` and `DeepSeek-V4-Flash`. Figure 4 suggests that the gap is not simply that models *do more synthesis*. Instead, model ideas concentrate around a recognizable recipe: select a high-frequency technical concept cluster and apply a safe synthesis operation such as integrating, unifying, combining, or adapting it with another nearby concept.

Archetype Clustering. We use `GPT-5.4-mini` to rewrite each proposal into a one-sentence archetype that abstracts away domain-specific details while preserving the high-level idea. We then cluster these valid archetypes using TF-IDF and MiniBatchKMeans with $k = 30$ [Sculley, 2010], and normalize the main verb in each archetype into an operation family. The strongest model-heavy operation is ***integrate***: it appears 7,994 times in model outputs (34.2%) but only 275 times in human ideas (2.35%), giving a model-vs-human log-odds of 3.07. Other overrepresented operations are also synthesis-like, including ***unify*** (8.2% vs. 1.9%, log-odds 1.52), ***design*** (1.5% vs. 0.3%, 1.50), ***merge*** (1.37), and ***adapt*** (1.36).

The human-heavy side looks different. Human ideas more often use local intervention operations such as ***replace***, ***decouple***, and ***formalize***. Specifically, ***replace*** makes up 9.13% of human operations versus just 0.92% for models, and ***decouple*** reaches 2.33% among humans and only 0.21% for models. Two human-majority clusters are largely missing from model outputs: ***replace*** (83.3% human) and ***decouple*** (85.4%). Relative to model-heavy clusters such as integrate / model / learning, they achieve higher bottleneck specificity (2.61, 2.70) and lower boilerplate (0.60,

0.51). Beyond reduced synthesis, human ideas typically revise targeted assumptions, modules and mechanisms in local research settings.

Representation Mechanism. We map proposals, motivations, methods, prior works, archetypes and extracted concept phrases into a shared representation space with 2560-dimensional embeddings with `Qwen3-Embedding-4B`. These representations show strong consensus between model outputs from the same source paper. For a given input paper, the cosine similarity between ideas from `Qwen3-8B` and `DeepSeek-V4-Flash` reaches 0.8316, while the scores stand at 0.7242 for human–Qwen pairs and 0.7829 for human–Deepseek pairs. The model pair exceeds the average human–model similarity, suggesting that distinct model families converge to similar generation patterns even when given the same context and paper-specific prior knowledge as humans.

We quantify how diffusely each proposal is positioned relative to its prior-work set. Given a proposal embedding p and prior-work embeddings $\{w_i\}_{i=1}^{K}$, we compute cosine similarities $s_i = p^\top w_i$, convert them into a similarity distribution, and compute its normalized entropy H. Higher H means that the proposal is comparably close to several prior works, rather than being dominated by a single reference. We also compute $B = p^\top c + H - (s_{(1)} - s_{(2)})$, where c is the normalized centroid of the prior-work embeddings, and $s_{(1)}$ and $s_{(2)}$ are the largest and second-largest similarities B increases when a proposal is close to the overall prior-work centroid, has high entropy over priors, and its top two related-work similarities are not sharply separated. Human proposals have higher values on both quantities ($H = 0.7215$, $B = 1.4662$) than model proposals, with `Qwen3-8B` at $H = 0.6494$ and $B = 1.3345$, and `DeepSeek-V4-Flash` at $H = 0.6745$ and $B = 1.4237$. This suggests that model proposals are less broadly positioned within the available prior-work set.

Concept Enrichment. To connect archetype operations with their semantic content, we rank the representation-analysis concept clusters by model-vs-human log-odds, keeping clusters with at least 30 occurrences. We report both *core* archetype concepts and *all* concepts, which add proposal-level TF-IDF terms. Since the pool contains one human and two model sources, we interpret enrichment by log-odds instead of raw majority.

The strongest model-enriched core clusters are reusable technical motifs: multi-omics integration has 317 occurrences across 180 records and 95.0% model share (log-odds 1.96), followed by diffusion policy (93.1%, 1.61), multimodal generation (91.2%, 1.34), in-context learning (90.2%, 1.23), test-time adaptation / adaptive optimization (89.5%, 1.19), quantization, multi-agent / LLM-agent concepts, and multimodal reasoning (86.6–88.1%). The all-concept view keeps multi-omics and diffusion policy at the top and adds physical constraints, climate models, adversarial robustness, implicit regularization, and cross-modal representations. Many representative phrases already contain *integrate*, *combine*, or *unify*, suggesting that these concepts serve as slots for the synthesis template.

Human-enriched clusters are more local: trajectories and tracking trajectories (63.5% human, log-odds -1.50), ligands and molecular interactions, tokenization and token importance, equivariance / inverse problems / Hamiltonian structure, entropy / mutual information, routing and prototypes, and verification concepts (44.8 to 53.2%, -0.75 to -1.09). In the all-concept view, function vectors / internal representations, policy evaluation, denoising, and geometric concepts remain relatively human-enriched. These clusters name variables, constraints, and mechanisms to intervene on, together with the operation analysis, they suggest that LLMs start from high-frequency technical concepts and wrap them in safe integration moves, while human papers more often modify, separate, or formalize a narrower local mechanism.

The mechanism analyses suggest that the gap between LLMs and human is not random, nor simply a preference for legitimate interdisciplinary work. LLM ideas repeatedly instantiate an archetype-level recipe: choose a salient technical concept cluster, then integrate or unify it with another nearby object. Human ideas are more likely to make a narrower intervention, such as replacing a brittle component, decoupling two confounded mechanisms, or formalizing a local structure. This provides one plausible account of why LLM ideas can seem reasonable while still occupying a more concentrated region of research taste.

5 Conclusion and Discussion

We introduced an evaluation framework for comparing human and LLM research ideas under shared literature-grounded inputs. By extracting human ideas from real papers, prompting LLMs on reconstructed related-work contexts, and labeling ideas with a two-axis research-taste taxonomy, we find that LLMs occupy a much narrower region of research taste than humans. The central pattern is an overproduction of bridge-like opportunities and synthesis-oriented method paradigms, supported by diagnostic and mechanism analyses showing that many model outputs are less specific, more boilerplate-like, and organized around repeated integrate archetypes. LLM ideation should be evaluated as a distributional alignment problem. A useful ideation system should also diversify how it identifies problems, constructs contributions, and departs from familiar synthesis templates. This gives future ideation systems a target: preserve the fluency and scale of LLM generation while shifting proposals toward more specific, mechanism-aware, and less template-bound research patterns.

Limitations

Our corpus is broad but still STEM-centered, so research-taste distributions may differ in social science, humanities, clinical research, or engineering design. Our task reconstructs a local literature context from related works, whereas real researchers draw on tacit expertise, failed attempts, collaborations, reviewer feedback, and long-term research agendas. Although our taxonomy and LLM annotation pipeline are human-validated, they compress nuanced ideas into discrete labels and diagnostic scores, and some proposals naturally mix multiple research moves. Finally, we evaluate a finite set of models, prompts, and one-shot settings; interactive agents, domain-specific systems, retrieval-heavy pipelines, or prompt interventions may reduce some of the observed gaps. Our results should therefore be read as evidence about current LLM ideation under a controlled literature-grounded setting, not as a claim about all possible AI-assisted research workflows.

Acknowledgments

This work was supported in part by the U.S. National Science Foundation under award No. 2541654.

References

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 94003–94092, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ea94957d81b1c1caf87ef5319fa6b467-Paper-Conference.pdf.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. ResearchAgent: Iterative research idea generation over scientific literature with large language models. In *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6709–6738, 2025.

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

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery, 2024. URL https://arxiv.org/abs/2408.06292.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8783–8817, 2024.

Yilun Zhao, Weiyuan Chen, Zhijian Xu, Manasi Patwardhan, Chengye Wang, Yixin Liu, Lovekesh Vig, and Arman Cohan. AbGen: Evaluating large language models in ablation study design and evaluation for scientific research. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12479–12491, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.611. URL https://aclanthology.org/2025.acl-long.611/.

Rosni Vasu, Chandrayee Basu, Bhavana Dalvi Mishra, Cristina Sarasua, Peter Clark, and Abraham Bernstein. HypER: Literature-grounded hypothesis generation and distillation with provenance. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25413–25438, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1292. URL https://aclanthology.org/2025.emnlp-main.1292/.

Yilun Zhao, Kaiyan Zhang, Tiansheng Hu, Sihong Wu, Ronan Le Bras, Yixin Liu, Xiangru Tang, Joseph Chee Chang, Jesse Dodge, Jonathan Bragg, Chen Zhao, Hannaneh Hajishirzi, Doug Downey, and Arman Cohan. Sciarena: An open evaluation platform for non-verifiable scientific literature-grounded tasks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=am6RR85mnc.

Sihong Wu, Yiling Ma, Yilun Zhao, Tiansheng Hu, Owen Jiang, Manasi Patwardhan, and Arman Cohan. RbtAct: Rebuttal as supervision for actionable review feedback generation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Findings of the Association for Computational Linguistics: ACL 2026*, pages 33965–33992, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. URL https://aclanthology.org/2026.findings-acl.1696/.

Aniketh Garikaparthi, Manasi Patwardhan, Lovekesh Vig, and Arman Cohan. IRIS: Interactive research ideation system for accelerating scientific discovery. In *Proceedings of ACL – System Demonstrations*, pages 592–603, July 2025.

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, Ning Ding, Xuanjing Huang, Yugang Jiang, and Xipeng Qiu. Ai can learn scientific taste, 2026. URL https://arxiv.org/abs/2603.14473.

Lennart Meincke, Ethan R Mollick, and Christian Terwiesch. Prompting diverse ideas: Increasing ai idea variance. *arXiv preprint arXiv:2402.01727*, 2024.

Brandon Smith, Mohamed Reda Bouadjenek, Tahsin Alamgir Kheya, Phillip Dawson, and Sunil Aryal. A comprehensive analysis of large language model outputs: Similarity, diversity, and bias. *arXiv preprint arXiv:2505.09056*, 2025.

Zhivar Sourati, Alireza S Ziabari, and Morteza Dehghani. The homogenizing effect of large language models on human expression and thought. *Trends in Cognitive Sciences*, 2026.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific inspiration machines optimized for novelty. In *Annual Meeting of the Association for Computational Linguistics*, pages 279–299, 2024.

Aishik Sanyal, Samuel Schapiro, Sumuk Shashidhar, Royce Moon, Lav R Varshney, and Dilek Hakkani-Tur. Spark: A system for scientifically creative idea generation. In *International Conference on Computational Creativity*, 2025.

Alireza Ghafarollahi and Markus J. Buehler. Sciagents: Automating scientific discovery through multi-agent intelligent graph reasoning, 2024. URL https://arxiv.org/abs/2409.05556.

Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. Nova: An iterative planning and search approach to enhance novelty and diversity of LLM generated ideas. *arXiv preprint arXiv:2410.14255*, 2024.

Kevin Pu, K. J. Kevin Feng, Tovi Grossman, Tom Hope, Bhavana Dalvi Mishra, Matt Latzke, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. IdeaSynth: Iterative research idea development through evolving and composing idea facets with literature-grounded feedback. In *CHI Conference on Human Factors in Computing Systems*, pages 1–31, 2025.

Tianyang Gu, Jingjin Wang, Zhihao Zhang, and Haohong Li. LLMs can realize combinatorial creativity: Generating creative ideas via LLMs for scientific research. *arXiv preprint arXiv:2412.14141*, 2024.

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. Many heads are better than one: Improved scientific idea generation by a LLM-based multi-agent system. In *Annual Meeting of the Association for Computational Linguistics*, pages 28201–28240, 2025.

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. Evaluating llms' divergent thinking capabilities for scientific idea generation with minimal context, 2026. URL https://arxiv.org/abs/2412.17596.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Myles Kim, Corey M Williams, Stefan Bekiranov, and Aidong Zhang. IdeaBench: Benchmarking large language models for research idea generation. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5888–5899, 2025a.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition, 2026. URL https://arxiv.org/abs/2503.21248.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. GLTR: Statistical detection and visualization of generated text. In Marta R. Costa-jussà and Enrique Alfonseca, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3019. URL https://aclanthology.org/P19-3019/.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1808–1822, 2020.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/mitchell23a.html.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4816–4828. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/260c2432a0eecc28ce03c10dadc078a4-Paper.pdf.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pages 337–371. PMLR, 2023.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International conference on machine learning*, pages 29971–30004. PMLR, 2023.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.20. URL https://aclanthology.org/2025.acl-short.20/.

Hyungyu Shin, Jingyu Tang, Yoonjoo Lee, Nayoung Kim, Hyunseung Lim, Ji Yong Cho, Hwajung Hong, Moontae Lee, and Juho Kim. Mind the blind spots: A focus-level evaluation framework for LLM reviews. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35630–35656, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1805. URL https://aclanthology.org/2025.emnlp-main.1805/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522, 2023.

Seungone Kim, Jay Shin, yejin cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, S Shin, Ryan, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 29927–29962, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/803485352e61e3ebf41221e4776c9fd4-Paper-Conference.pdf.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46, 1960.

Anthropic. Introducing Claude Sonnet 4.6, Feb 2026. URL https://www.anthropic.com/news/claude-sonnet-4-6.

Google. Gemini 3.1 Pro: A smarter model for your most complex tasks, Feb 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

OpenAI. Introducing GPT-5.4, Mar 2026. URL https://openai.com/index/introducing-gpt-5-4/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1): 145–151, 1991.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/f7696a9b362ac5a51c3dc8f098b73923-Paper.pdf.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025b.

David Sculley. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pages 1177–1178, 2010.

A Dataset Details

Each data point starts from a real human paper and contains its extracted idea, represented as a motivation and method, together with a reconstructed local literature context. The literature context consists of proximal prior works, each represented by title and abstract. We collect papers from ICLR, ICML, NeurIPS, and Nature Communications, then merge the human source, all model generations, and all annotations by paper ID. Rows with missing or invalid model outputs or labels are removed. The final matched evaluation set contains 11,683 papers: 5,994 ML papers and 5,689 Nature Communications papers as shown in Table 5.

Table 5: Dataset statistics for the matched evaluation corpus. Papers are rows used in the main distributional analyses, where all evaluated sources have valid outputs and annotations. Avg. priors and median priors report the number of reconstructed prior works per paper.

Corpus	Source years	Papers	Avg. priors	Median priors
ML Conference	2023–2026	5,994	6.52	7
Nature Communications	2023–2025	5,689	5.89	6
Mixed	2023–2026	11,683	6.21	6

B Taxonomy Design

The taxonomy was designed to compare research taste rather than topic, field, or technical substrate. Before coding examples, we reviewed proposal-writing and research-gap guidance from DARPA's Heilmeier Catechism,[2] NIH application guidance,[3] the NSF PAPPG project-description instructions,[4] and AHRQ guidance on identifying research gaps.[5] These documents shaped the distinction between identifying why current knowledge is insufficient and explaining what new contribution will address that insufficiency. We then iteratively refined labels on sampled human and LLM ideas, removing categories tied to domain or method family.

Opportunity Pattern.

- *Puzzle / Contradiction*: a paradox, tradeoff, surprising failure, or conflicting evidence.
- *Explanation Gap*: a missing causal, mechanistic, theoretical, or explanatory account.
- *Scope Mismatch*: unrealistic assumptions, narrow regimes, unclear transferability, or boundary conditions.
- *Evidence Gap*: missing ways to observe, measure, benchmark, audit, diagnose, validate, compare, or accumulate evidence.
- *Bridge Opportunity*: disconnected literatures, theories, evidence streams, communities, or methods that could be connected.
- *Failure / Risk Gap*: brittleness, unreliability, bias, uncertainty, safety/privacy/security risk, or reproducibility concerns.
- *Resource Bottleneck*: cost, compute, time, data access, sample scarcity, experimental burden, deployment friction, usability, or scalability constraints.

Method Paradigm.

- *Synthesis / Unification*: bridges, integrates, reconciles, or unifies separate literatures, theories, evidence streams, mechanisms, or methods.
- *Relax / Extend Scope*: makes prior work function under weaker assumptions, broader scope, new regimes, noisier conditions, or more realistic settings.
- *Robustification*: reduces failures, brittleness, risk, uncertainty, bias, unreliability, or trustworthiness problems.
- *Formal Derivation*: introduces a formal model, theorem, bound, objective, derivation, proof, ontology, taxonomy, conceptual distinction, or explanatory formulation.

[2] https://www.darpa.mil/about/heilmeier-catechism
[3] https://www.grants.nih.gov/grants-process/write-application/advice-on-application-sections
[4] https://www.nsf.gov/policies/pappg/24-1/ch-2-proposal-preparation
[5] https://www.ncbi.nlm.nih.gov/books/NBK62480/

Table 6: Guideline primitives used to ground the research-taste taxonomy. We report the related items extracted from each source and how they informed our taxonomy.

Source	Extracted Information	For Our Taxonomy
DARPA Heilmeier Catechism	Limits of current practice; novelty of the approach; risk; cost and timeline; mid-term and final checks for success.	Separates the motivating limitation from the proposed approach, and directly motivates labels for scope limits, risk/failure, resource bottlenecks, and success-oriented evidence.
NIH Application Guidance	Specific Aims; hypothesis-based goals; scientific question; Significance; Innovation; Approach; major experiments; rigor, reproducibility, transparency; facilities and resources.	Supports the motivation–method split through Significance/Innovation versus Approach, and motivates labels involving explanation, empirical work, robustness/reproducibility, and resources.
NSF PAPPG Project Description	Objectives and methods; potential to advance knowledge; relationship to the present state of knowledge; work plan and experimental procedures; how success will be known; research products such as data, software, models, samples, and equipment; interdisciplinary work and integration or transfer of knowledge; facilities and equipment.	Gives the statement of the proposal pattern: what, why, how, success criteria, and benefits. It supports evidence, artifact/system, resource, bridge/synthesis, and scope-oriented categories.
AHRQ Research-Gap Framework	Why a research gap exists; where evidence falls short; insufficient or imprecise information; biased information; inconsistent or unknown consistency; not-the-right information; population, intervention, comparison, outcome, and setting elements; translation from gaps to research questions.	Provides the grounding for the opportunity axis, especially evidence gaps, contradictions, scope mismatch, risk/bias, empirical characterization, and scope extension.

- *Empirical Mapping*: builds or applies systematic measurement, benchmarks, diagnostics, datasets, empirical maps, comparative studies, or pattern analyses.
- *Artifact / System*: builds a concrete artifact, software system, platform, device, material, prototype, or deployment workflow as the central contribution.
- *Optimization / Search*: uses optimization, search, screening, tuning, active/adaptive design, scaling, resource allocation, or efficiency strategies to discover or improve a solution.

C Experimental Details and Configurations

Artifacts and intended use. This work uses public scholarly artifacts: paper metadata, titles, abstracts, extracted prior-work contexts, model-generated ideas, taxonomy labels, and annotation outputs. Source papers are drawn from ML conference proceedings and Nature Communications; abstracts are retrieved from public scholarly metadata providers when available. We use these artifacts only for research evaluation. Released dataset will preserve source attribution.

Privacy and content. The corpus does not contain private user-generated text, recruited participant records, or human-subject data. It consists of public scholarly text and derived research-idea summaries.

Model and compute setup. Each evaluated model generates one idea per input context. The main proposal-generation prompt uses the detailed profile in Figure 6. Local open-weight runs use temperature 0.6, top-$p = 0.95$, top-$k = 20$, maximum 2,048 new tokens, and one output per input; the GPT API run uses temperature 1.0 and a JSON-schema output constraint. Thinking-mode runs differ only in the model-side thinking flag or reasoning setting. API models are run through provider APIs; local model and embedding jobs are run on cluster GPUs. The embedding analysis uses `Qwen3-Embedding-4B` with 2,560-dimensional embeddings, maximum length 512, batch size 12, bfloat16, and last-token pooling on CUDA.

Analysis parameters. No hyperparameter search is performed for the reported distributional metrics. Archetype clustering uses TF-IDF with lowercasing, English stop words, 1–2 grams, $\min\_df = 2$, $\max\_df = 0.85$, sublinear TF, and MiniBatchKMeans with $k = 30$, batch size 512, seed 13, and `n_init=auto`. Concept extraction uses TF-IDF over proposal text with 1–3 grams and custom stop words, while concept clustering uses MiniBatchKMeans over Qwen3 embeddings with batch size 512 and seed 13. Linear probes use 5-fold stratified cross-validation with balanced logistic regression, $C = 1.0$, `liblinear`, and fixed random seed 13.

Human audit instructions. Two authors audited a stratified sample of 150 annotations. For each item, they were shown the prior-work titles, motivation, method, LLM-assigned labels, diagnostic scores, and the taxonomy definitions in Appendix B. They independently judged whether the primary Opportunity Pattern, primary Method Paradigm, and diagnostic-score profile were acceptable under the codebook, then disagreements were inspected through confusion matrices. No external annotators were recruited or paid.

AI-assistant use. LLMs are part of the research pipeline: they assist with idea extraction, related-work reconstruction, proposal generation, annotation, and archetype rewriting. The manuscript reports these uses in the method and prompt appendices. We use LLMs for writing polishing and grammar checking.

D Prompts for Idea Generation and Annotation

We show prompt templates used in the pipeline. Bracketed fields are filled separately for each paper.

Prompt: Prior Work Extraction

You are an expert AI research analyst. Given a paper, identify the **proximal prior works** that most directly shaped its core idea. Focus on papers the authors likely engaged with when forming the contribution, not on general background knowledge.

Step 1: Identify the core idea. Briefly determine the paper's main innovation, the prior limitation or gap it responds to, and the specific insight that makes the contribution non-obvious.

Step 2: Select proximal prior works. Choose **5–7** papers from the related-work context that are most useful for reconstructing how the current idea could have emerged. Prefer recent and specific predecessors over classic, foundational, or textbook-style works.

For each candidate prior work, apply three checks:

1. **Counterfactual check**: Would the authors likely still have produced this specific idea without reading this paper? If yes, exclude it.
2. **Specificity check**: What concrete method, finding, limitation, dataset, or insight from this prior work informed the current paper?
3. **Proximity check**: Is this the most recent or direct source for that influence? Prefer the newer and more proximal paper.

Do not include:

- classic foundational works cited only as background;
- generic tools, datasets, libraries, or infrastructure;
- papers used only as baselines or mentioned only in passing;
- papers that share the topic but do not shape the core idea.

Output (JSON): return 5–7 prior works. For each, provide `cite_id`, title, authors, year, and a one-sentence explanation of what the authors learned from it and how it informed the current paper.

Figure 5: Prompt used to reconstruct the proximal prior works for each human paper.

Prompt: LLM Idea Generation

You are a research scientist skilled at synthesizing ideas from existing literature into novel research proposals. You are given a set of related research papers with titles and abstracts. Analyze these papers, identify research gaps and opportunities, and propose one coherent novel research idea.

Input format. The user message lists prior works as blocks of `# Title (cite_id)` followed by `## Abstract: ...`.

Output requirements. Based only on the papers above, return a valid JSON object with exactly two string fields:

```
{
  "motivation": "...",
  "method": "..."
}
```

The motivation should synthesize the research gap, why it matters, and why the listed works leave room for the proposed idea. The method should describe a concrete, feasible high-level approach and explain how it addresses the gap. Do not include citations outside the provided papers, markdown fences, or any text before or after the JSON object.

Figure 6: Prompt template used to generate LLM research ideas from the reconstructed prior-work context.

Prompt: Research-Taste Annotation

You are an expert annotator of research taste. Label the proposal using high-level categories that apply across ML/AI, natural science, medicine, engineering, and social or behavioral science. Do not classify by topic, domain, or technical substrate. Classify the proposal's **problem-finding pattern** and **idea-construction paradigm**. The two axes use disjoint labels; never copy a method-paradigm label into the opportunity axis, or vice versa.

Opportunity Pattern labels.

- **Puzzle / Contradiction**: the gap comes from a paradox, tradeoff, surprising failure, or conflicting evidence.
- **Explanation Gap**: the proposal asks why something works, fails, varies, or appears.
- **Scope Mismatch**: prior work relies on narrow, unrealistic, or poorly transferable assumptions.
- **Evidence Gap**: the field lacks measurement, benchmarking, auditing, diagnosis, validation, or comparison.
- **Bridge Opportunity**: disconnected literatures, methods, theories, or evidence streams need to be connected.
- **Failure / Risk Gap**: existing approaches raise reliability, robustness, safety, bias, uncertainty, or reproducibility concerns.
- **Resource Bottleneck**: progress is limited by cost, compute, data, time, deployment, usability, or scalability.

Method Paradigm labels.

- **Synthesis / Unification**: integrates or reconciles separate literatures, mechanisms, theories, evidence streams, or methods.
- **Relax / Extend Scope**: makes prior work apply under broader, noisier, weaker-assumption, or more realistic settings.
- **Robustification**: reduces failure, brittleness, risk, uncertainty, bias, unreliability, or trustworthiness problems.
- **Formal Derivation**: introduces a theory, theorem, bound, objective, proof, taxonomy, or conceptual formulation.
- **Empirical Mapping**: builds or applies measurements, benchmarks, datasets, diagnostics, empirical maps, or comparative studies.
- **Artifact / System**: constructs a concrete tool, platform, software system, device, material, prototype, or workflow.
- **Optimization / Search**: uses optimization, search, tuning, screening, selection, allocation, scaling, or efficiency strategy.

Input fields. Paper ID; prior-work titles for context; proposal motivation; proposal method.

Decision guidance. The opportunity axis asks how the gap is found. The method-paradigm axis asks what kind of research move constructs the paper. Use Synthesis / Unification only when bridging or reconciling separate lines of work is central, not merely when a method has multiple components. Use Empirical Mapping for estimating, auditing, diagnosing, quantifying, or characterizing a phenomenon. Use Artifact / System only when a concrete artifact, system, tool, platform, material, or prototype is the central deliverable. Use Optimization / Search when the central move is efficiency, scaling, search, tuning, selection, allocation, or resource-aware design.

Output (JSON): return primary and secondary labels for each axis, confidence scores, diagnostic scores for `surface_stitching`, `bottleneck_specificity`, and `boilerplate`, and one concise rationale sentence.

```
{
  "labels": {
    "opportunity_pattern": {"primary": "<one opportunity_pattern label>", "secondary": "<one
opportunity_pattern label or none>"},
    "research_idea_paradigm": {"primary": "<one research_idea_paradigm label>", "secondary":
"<one research_idea_paradigm label or none>"}
  },
  "confidence": {"opportunity_pattern": <0.0-1.0>, "research_idea_paradigm": <0.0-1.0>},
  "diagnostics": {
    "surface_stitching": <true or false>,
    "surface_stitching_score": <integer 0-3, where 3 is clearly superficial A+B stitching>,
    "bottleneck_specificity": <integer 0-3, where 3 identifies a precise bottleneck, mechanism,
or limiting factor>,
    "boilerplate_score": <integer 0-3, where 3 is highly generic or boilerplate>
  },
  "rationale": "<one concise sentence>"
}
```

Figure 7: Annotation prompt. The implementation uses the same label set and maps the human-readable labels shown here to internal JSON keys.

E Additional Distributional Analyses

E.1 Domain-Specific Results

We present the main distributional comparison separately for the Machine Learning corpus and the Nature Communications corpus in Table 7 and Table 8.

Table 7: ML distributional distances against the human distribution. Ent. is normalized entropy. Header arrows indicate the direction closer to the human distribution.

	Opportunity Pattern			Method Paradigm		
Source	TVD ↓	JSD ↓	Ent. ↑	TVD ↓	JSD ↓	Ent. ↑
Human	—	—	0.952	—	—	0.968
`Claude-Sonnet-4.6`	**0.447**	**0.172**	**0.699**	**0.288**	**0.105**	**0.896**
`Gemini-3.1-Pro`	0.448	0.188	0.676	0.350	0.156	0.835
`GPT-OSS-20B`	0.643	0.338	0.438	0.532	0.254	0.683
`GPT-OSS-120B`	0.683	0.379	0.383	0.523	0.266	0.679
`GPT-5.4-mini`	0.579	0.276	0.525	0.397	0.155	0.819
`Qwen3-8B`	0.598	0.302	0.505	0.533	0.301	0.641
`Qwen3-8B-Think`	0.719	0.428	0.321	0.640	0.393	0.518
`Qwen3-32B`	0.560	0.266	0.553	0.512	0.280	0.666
`DeepSeek-V4-Flash`	0.522	0.231	0.605	0.326	0.135	0.855
`DeepSeek-V4-Flash-Think`	0.593	0.291	0.511	0.402	0.194	0.785
`DeepSeek-V4-Pro`	0.586	0.296	0.542	0.399	0.175	0.809

Table 8: Nature Communications distributional distances against the human distribution. Ent. is normalized entropy. Header arrows indicate the direction closer to the human distribution.

	Opportunity Pattern			Method Paradigm		
Source	TVD ↓	JSD ↓	Ent. ↑	TVD ↓	JSD ↓	Ent. ↑
Human	—	—	0.822	—	—	0.790
`Claude-Sonnet-4.6`	0.250	0.093	0.670	0.177	**0.045**	0.698
`Gemini-3.1-Pro`	**0.243**	**0.079**	**0.728**	**0.140**	0.047	**0.741**
`GPT-OSS-20B`	0.298	0.124	0.591	0.232	0.096	0.565
`GPT-OSS-120B`	0.351	0.159	0.585	0.252	0.094	0.608
`GPT-5.4-mini`	0.441	0.216	0.528	0.278	0.098	0.701
`Qwen3-8B`	0.250	0.085	0.617	0.221	0.100	0.635
`Qwen3-8B-Think`	0.454	0.221	0.540	0.314	0.173	0.657
`Qwen3-32B`	0.299	0.127	0.593	0.215	0.105	0.656
`DeepSeek-V4-Flash`	0.272	0.111	0.641	0.181	0.054	0.652
`DeepSeek-V4-Flash-Think`	0.340	0.144	0.617	0.184	0.066	0.680
`DeepSeek-V4-Pro`	0.279	0.128	0.626	0.182	0.054	0.665

Table 9: Domain-specific percentages for the two labels most emphasized in the main text: Bridge Opportunity on the opportunity axis and Synthesis / Unification on the method-paradigm axis. Header arrows mark the direction closer to the human distribution.

Source	ML Bridge ↓	ML Synthesis ↓	NC Bridge ↓	NC Synthesis ↓
Human	14.0	6.6	10.2	3.4
`Claude-Sonnet-4.6`	58.7	35.5	35.2	11.4
`Gemini-3.1-Pro`	58.8	41.6	34.5	13.2
`GPT-OSS-20B`	78.4	44.0	35.9	8.8
`GPT-OSS-120B`	82.3	51.1	45.2	11.3
`GPT-5.4-mini`	71.9	37.8	54.2	19.5
`Qwen3-8B`	73.8	59.9	24.3	16.4
`Qwen3-8B-Think`	85.9	70.7	55.5	32.8
`Qwen3-32B`	70.0	57.1	36.9	19.2
`DeepSeek-V4-Flash`	66.2	35.5	37.4	8.8
`DeepSeek-V4-Flash-Think`	73.3	46.7	44.2	13.7
`DeepSeek-V4-Pro`	71.5	46.5	38.1	11.1

E.2 Full-Paper Context Ablation

The main experiments represent each proximal prior work with its title and abstract. To test whether the bridge-and-synthesis pattern is an artifact of this compressed context, we run a full-paper context ablation on 1,000 inputs: 500 sampled from the Machine Learning corpus and 500 sampled from the Nature Communications corpus. For each related-work paper, the model first reads the full text and produces a compact summary with detailed motivation, method and insight. We then use these model-generated full-paper summaries in place of abstracts as the context for idea generation, keeping the downstream annotation pipeline unchanged.

Table 10 compares the original abstract-context condition with the full-paper-summary condition on the same subset for `Qwen3-8B` and `DeepSeek-V4-Flash`. The richer context does not remove the bridge-heavy opportunity pattern. For `Qwen3-8B`, bridge opportunities increase from 456 to 551 cases; for `DeepSeek-V4-Flash`, they increase from 489 to 521 cases. The listed method-paradigm categories also do not show a compensating shift toward formal derivation, empirical mapping, artifact construction, or optimization.

Table 10: Full-paper context ablation on a 1,000 paper subset. Original uses title and abstract context; full context replaces abstracts with model-generated summaries of the full papers. All taxonomy labels are shown; parentheses show count changes from the original condition. Metric rows report TVD/JSD against the same human reference distribution as Table 1; Ent. is normalized entropy.

	`Qwen3-8B`		`DeepSeek-V4-Flash`	
Label	Original	Full context	Original	Full context
Opportunity Pattern				
Puzzle / Contradiction	8	5 (-3)	39	36 (-3)
Explanation Gap	353	284 (-69)	281	246 (-35)
Scope Mismatch	17	21 (+4)	21	27 (+6)
Evidence Gap	43	41 (-2)	58	55 (-3)
Bridge Opportunity	456	551 (+95)	489	521 (+32)
Failure / Risk Gap	85	69 (-16)	94	95 (+1)
Resource Bottleneck	38	29 (-9)	18	20 (+2)
Method Paradigm				
Synthesis / Unification	389	451 (+62)	233	260 (+27)
Relax / Extend Scope	24	26 (+2)	60	47 (-13)
Robustification	36	32 (-4)	82	95 (+13)
Formal Derivation	23	9 (-14)	29	28 (-1)
Empirical Mapping	255	229 (-26)	251	229 (-22)
Artifact / System	242	234 (-8)	286	286 (0)
Optimization / Search	31	19 (-12)	59	55 (-4)
Opportunity Pattern Metrics				
TVD ↓	0.376	0.430 (+.054)	0.368	0.400 (+.032)
JSD ↓	0.165	0.205 (+.040)	0.152	0.160 (+.008)
Ent. ↑	0.669	0.623 (-.046)	0.706	0.701 (-.005)
Method Paradigm Metrics				
TVD ↓	0.338	0.400 (+.062)	0.213	0.236 (+.023)
JSD ↓	0.182	0.229 (+.047)	0.079	0.093 (+.014)
Ent. ↑	0.752	0.699 (-.053)	0.867	0.860 (-.007)

E.3 Prompt Ablation

The main generation prompt asks models to analyze a set of prior papers, identify research gaps and opportunities, and propose one coherent idea. This wording may itself encourage models to connect papers into a synthesis-style proposal. We therefore run a prompt ablation that relaxes this constraint. We use more neutral terms like *generate* and *describe* instead of expressions that might affect the paradigm of the model ideation.

Prompt: Relaxed LLM Idea Generation

You are a research scientist skilled at generating ideas from existing literature into novel research proposals. You are given a set of related research papers with titles and abstracts. Analyze these papers, identify research gaps and opportunities, and propose one coherent novel research idea.

Input format. The user message lists prior works as blocks of `# Title (cite_id)` followed by `## Abstract: ...`.

Output requirements. Based only on the papers above, return a valid JSON object with exactly two string fields:

```
{
  "motivation": "...",
  "method": "..."
}
```

The motivation should describe the research gap, why it matters, and why the listed works leave room for the proposed idea. The method should describe a concrete, feasible high-level approach and explain how it addresses the gap. Do not include citations outside the provided papers, markdown fences, or any text before or after the JSON object.

Figure 8: Relaxed generation prompt used for the prompt ablation.

Table 11 reports the resulting count changes on the full matched evaluation set. The ablation changes the distribution, but it does not eliminate the main qualitative pattern. For `Qwen3-8B`, bridge opportunities decrease from 5,807 to 5,247 cases, while failure / risk gaps increase from 987 to 1,238 cases; nevertheless, bridge opportunities remain the largest listed opportunity category. For `DeepSeek-V4-Flash`, bridge opportunities increase from 6,094 to 6,368 cases under the relaxed prompt. On the method axis, the relaxed prompt moves some mass toward artifact / system and optimization / search categories, especially for `Qwen3-8B`, but the shifts are modest relative to the overall corpus size. These results suggest that prompt wording affects the exact label mix, while the tendency to organize generated ideas around bridge-like opportunities remains stable across prompt variants.

Table 11: Prompt ablation on the full matched evaluation set of 11,683 papers. Original uses the main generation prompt in Figure 6; relaxed uses the prompt in Figure 8. All taxonomy labels are shown; parentheses show count changes from the original condition. Metric rows report TVD/JSD against the human reference distribution; Ent. is normalized entropy.

	`Qwen3-8B`		`DeepSeek-V4-Flash`	
Label	Original	Relaxed	Original	Relaxed
Opportunity Pattern				
Puzzle / Contradiction	121	86 (-35)	381	348 (-33)
Explanation Gap	3,714	3,917 (+203)	3,064	3,017 (-47)
Scope Mismatch	203	280 (+77)	305	333 (+28)
Evidence Gap	440	482 (+42)	550	478 (-72)
Bridge Opportunity	5,807	5,247 (-560)	6,094	6,368 (+274)
Failure / Risk Gap	987	1,238 (+251)	1,106	957 (-149)
Resource Bottleneck	411	433 (+22)	183	182 (-1)
Method Paradigm				
Synthesis / Unification	4,523	3,938 (-585)	2,632	2,634 (+2)
Relax / Extend Scope	336	459 (+123)	781	760 (-21)
Robustification	413	531 (+118)	975	929 (-46)
Formal Derivation	168	204 (+36)	226	223 (-3)
Empirical Mapping	2,570	2,612 (+42)	2,661	2,549 (-112)
Artifact / System	3,402	3,597 (+195)	3,871	4,031 (+160)
Optimization / Search	271	342 (+71)	537	557 (+20)
Opportunity Pattern Metrics				
TVD ↓	0.382	0.351 (-.031)	0.400	0.424 (+.024)
JSD ↓	0.179	0.152 (-.027)	0.167	0.181 (+.014)
Ent. ↑	0.658	0.690 (+.032)	0.683	0.661 (-.022)
Method Paradigm Metrics				
TVD ↓	0.368	0.335 (-.033)	0.246	0.260 (+.014)
JSD ↓	0.190	0.153 (-.037)	0.086	0.089 (+.003)
Ent. ↑	0.734	0.774 (+.040)	0.845	0.840 (-.005)