IDEAGENT: AGENTIC *Quality-Diversity* SEARCH FOR RESEARCH IDEA GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) have significantly advanced the automation of scientific discovery. However, existing systems share a core limitation: they optimize ideas independently for either *Quality* or *Diversity*. This often yields closely related ideas or large sets of trivial, unsound, or unclear concepts. We instead argue that research ideation should be framed as a conjunction of *Quality* and *Diversity* objectives, i.e., a *Quality-Diversity* **(QD)** search. Accordingly, we introduce IDEAGENT, a *multi-agent* framework that tracks idea evolution through *lineages*. We jointly drive *Quality* using multi-objective feedback for repair and refinement, while *Diversity* is achieved through *lightweight* sequential memory and explicit comparison against completed ideas, historical variants, and rejected proposals. To evaluate this joint objective, we develop Yield, a metric that computes the largest mutually diverse set of ideas satisfying predefined quality thresholds. Across 32 topics spanning 8 CS domains, IDEAGENT outperforms the best baseline by $3.89\times$ on Yield, while achieving Yield of at least 2 on $8\times$ as many topics. We further analyze quality improvements, showing that repair and refinement improve logical rigor and clarity while preserving non-obviousness.

1 INTRODUCTION

The advent of LLMs has garnered interest in automatic scientific discovery. Recent works have shown the abilities of LLMs in assisting with literature surveys, hypothesis or idea generation, experiment design, draft writing, and now even autonomously executing chunks of a comprehensive research or experiment pipeline (Lu et al., 2024; Gottweis et al., 2026; Yamada et al., 2025). Of the aforementioned, the most fundamental step is *ideation*, where the core *gap* is identified in the background, one or more *research questions* (Sinhahajari et al., 2026) are formulated along with *potential solutions*. In this work, we propose to view the ideation process as a *Quality-Diversity Search* problem. The objective is to discover a *set* of research ideas, where each proposal is *non-obvious*, *sound*, and *clear* (**Quality**) while remaining mutually orthogonal (**Diversity**), allowing researchers to explore multiple independent research directions.

Despite recent progress, LLM-based ideation systems do not treat ideation as a rigorous QD search. Existing methods, from one-shot prompting to multi-agent systems, optimize each generation independently and evaluate ideas in isolation. Consequently, as more ideas are generated, successive generations frequently exhibit a *conceptual collapse* by repeatedly revisiting similar regions of the design space or producing minor, paraphrased variations of earlier concepts. Hence, while these systems may generate strong individual ideas, they offer rapidly diminishing returns (Tang & Yang, 2026) and are often riddled with inconsistencies while pushing for novelty. In real-world scientific workflows, an AI ideation system should generate not one proposal but a diverse set of *independent*, *sound*, *clear*, and *non-obvious* research directions.

Sustained human research naturally operates as a QD search. In each brainstorming session, researchers iterate over a concept—correcting logical flaws, tightening assumptions, and improving its overall *Quality*. This evolution forms a *lineage*, where higher-quality versions replace earlier drafts. Across sessions, diversity relies on recording three types of ideas: *completed* concepts, permanently rejected *dead ends*, and *historical variants* superseded by later refinements. Filtering against these archives forces exploration into new and unexplored directions, maximizing *Diversity*.

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

Motivated by this perspective, we introduce IDEAGENT, a multi-agent framework to navigate the QD search for research ideation. Our framework sequentially develops new ideas against compact memories of prior generations. Each idea is then evaluated by multiple specialized agents, while being compared against completed, historical, and rejected ideas for novelty, each with its dedicated archives. Accepted ideas may optionally be refined, while those that miss by a limited margin are put through repairs targeting exact flaws and inconsistencies. Finally, all comparisons operate over structured summaries for efficient semantic contrast.

Evaluating the QD search framework requires accounting for both objectives jointly. Independently, both *Quality* and *Diversity* metrics can be gamed. For example, a set of novel, but near-clone ideas represents a single contribution, while a diverse set of trivial, invalid, or unclear ideas holds no exploration or scientific potential. To address this loophole, we introduce Yield, a group-level metric that jointly measures quality in terms of non-obviousness, soundness, clarity, and diversity with pairwise distinctness. After scoring each requirement independently, Yield employs hard thresholds on the *quality* axes before extracting the *largest subset* in which all the ideas pairwise satisfy the diversity constraint. Physically, Yield is designed to measures the number of ideas a human researcher could *potentially* pursue from the generated set.

We evaluate IDEAGENT on 32 research topics spanning 8 Computer Science domains, and show that it consistently produces a *higher density of high-quality ideas within a limited ideation budget*. We validate our hypothesis that IDEAGENT achieves a higher average Yield and a greater number of topics with non-zero Yield than all considered baselines. We further analyse improvements across individual quality rubrics using both internal and external evaluators. Although the absolute scores differ between the two judges, they consistently agree on the direction of the improvements, particularly in soundness/rigor, clarity, and non-obviousness.

In short, our contributions are as follows: **(1)** We formulate automated scientific ideation as a QD search and argue that systems must develop multiple high-quality ideas that are sufficiently diverse from one another to maximize the number of *pursuable* research directions. **(2)** We propose IDEAGENT, a multi-agent framework that maintains idea *lineages* using dedicated quality evaluators. We further introduce separate archives for ideas at different life-cycle stages to enable efficient novelty comparison using compact core summaries instead of full textual ideas. Finally, we incorporate targeted repair and refinement routines driven by these evaluations. **(3)** We introduce Yield, a joint metric to assess both the *Quality* and *Diversity* of a set of ideas. We complement this with extensive experiments across 32 research topics, demonstrating the effectiveness of IDEAGENT for the QD search problem by consistently generating a larger set of non-obvious, sound, clear, and mutually distinct ideas.

2 RELATED WORK

Autonomous Frameworks for Scientific Discovery The role of LLMs in scientific discovery has rapidly shifted from drafting assistance to fully autonomous, end-to-end research pipelines (Lu et al., 2024; Gottweis et al., 2026; Schmidgall et al., 2025). To improve hypothesis quality, early frameworks introduced structured multi-agent environments, demonstrating that collaborative debate among distinct agent personas yields more robust proposals than single-turn generation (Su et al., 2025; Ueda et al., 2025). To ground these proposals and prevent superficial suggestions, subsequent architectures incorporated retrieval loops that ground reasoning in the existing literature (Baek et al., 2025; Li et al., 2025; Zhao et al., 2025). More recently, this paradigm has expanded from abstract ideation to active software execution that iteratively implements and optimizes code to solve competitive benchmarks (Yamada et al., 2025; Toledo et al., 2026).

Novelty Bottlenecks in AI-based Ideation Evaluating AI-based ideation pipelines across large corpora of hypotheses reveals a persistent failure of divergent thinking. LLM agents frequently exhibit conceptual narrowness, clustering outputs around the provided seed literature and recombining familiar technical variations rather than introducing genuinely new research directions (Tang & Yang, 2026). While Hu et al. (2025); Radensky et al. (2026) propose methods for algorithmic divergence, achieving true open-ended exploration remains a critical challenge. Finally, empirical studies show that success on complex downstream engineering tasks is fundamentally bounded by the diversity of early-stage ideation (Audran-Reiss et al., 2025).

QD Search in Text Generation Bradley et al. (2024) apply an evolutionary algorithm that uses language models to generate mutations, evaluate candidates for elitism, and operate on a fixed archive grid to map the search space. In contrast, we maintain a more flexible evolutionary process without a fixed search grid or MAP-Elites, while improving idea quality through directed feedback.

3 METHODOLOGY

Task Formulation as QD Search. IDEAGENT formulates scientific ideation as a Quality-Diversity search, i.e., under a fixed discovery budget, the aim is to produce as many high-quality and mutually distinct *ideas* as possible. Formally, given a background corpus $\mathcal{X}$ consisting of user-provided *foundational* research papers targeting an area of interest, we define an *idea* I as a structured research proposal specifying its problem, causal diagnosis, intervention, mechanism, assumptions, expected effect, and evaluation strategy.

Standard QD methods organize the search space using predefined axes and cells (Bradley et al., 2024; Lehman et al., 2024; Lehman & Stanley, 2011; Mouret & Clune, 2015; Pugh et al., 2016; Fontaine & Nikolaidis, 2021; Cully et al., 2015). This is generally too restrictive to express free-form research ideas, which may combine multiple intervention points. We therefore leave idea generation unconstrained and judge diversity from the proposed problem and causal mechanism.

An idea generated in one shot may contain a promising mechanism, but could be inconsistent, obvious, or underspecified. We therefore repair and refine the seed idea while tracking its evolution as a lineage ℓ_I. Close variants of the same mechanism inherit the same lineage and cannot be counted as separate discoveries. A seed and its lineage therefore represent one attempt to discover a mechanism and consume exactly one unit of the discovery budget B.

Agents and Search Memory. The **Ideator** is the primary generator that reads a background corpus $\mathcal{X}$ to generate a seed idea or subsequently refine it based on targeted feedback. Since the initial proposal I is loosely formed, a **Steno** *summarizes* I into a concise quintuple: $\sigma(I) = (p_I, m_I, v_I, a_I, e_I)$, where the fields denote the problem p_I, central mechanism m_I, novel value addition over the background literature v_I, key assumptions a_I, and expected measurable effect e_I. This decomposition provides a compact, interpretable representation for pairwise comparison between ideas.

Three evaluators assess each draft idea. The **Quality Evaluator** scores non-obviousness (NB_I), mechanism clarity (C_I), and feasibility (F_I) on a 0–100 scale. *Non-obviousness* measures whether the central value addition requires a genuine conceptual leap beyond routine extensions, direct combinations, or ablations. *Mechanism clarity* measures whether the causal steps from intervention to expected effect are well described. *Feasibility* measures whether a plausible implementation and evaluation path exists under realistic research resources. The **Soundness Panel** produces $M = 5$ independently sampled judgments $S_I^1, \ldots, S_I^M \in \{0, \ldots, 9\}$ on the logical and mathematical consistency of the idea's arguments and assumptions. These multiple judgments mitigate rollout-specific biases or oversights (Ho et al., 2026). Their mean forms a single representative score that is rescaled to 0–100 as $S_I = \frac{100}{9M} \sum_{j=1}^{M} S_I^j$. A judgment is *severe* when the score is no more than one-third of the scale, i.e., $S_I^j \leq 3$. At least one severe judgment marks the idea as *disputed*, while three or more make it *invalid*. Invalid ideas are discarded, whereas disputed ideas are retained for possible repairs. The **Diversity Judge** compares core signatures of the active, inactive, and rejected ideas against a freshly-generated idea. Based on the similarity with all previously qualified ideas, a Diversity score $D_I \in [0, 100]$ is issued, along with a list of duplicates identifying the closest active and historical mechanisms for further comparison.

Finally, a **Critic** turns evaluator scores and failure evidence into a focused feedback for repair or refinement of the current idea, rather than proposing a substitute direction.

The controller maintains four stores. The active archive ($\mathcal{A}$) contains currently accepted ideas. The repair queue ($\mathcal{Q}$) temporarily holds a promising near-miss during its immediate repair. The historical/inactive archive ($\mathcal{H}$) retains signatures of previously qualified ideas that have left $\mathcal{A}$, including replaced parents and ideas displaced by higher-quality ones. The rejected-pattern memory ($\mathcal{R}$) stores compact summaries of rejected ideas and recurring failure patterns, rather than full rejected ideas to avoid context bloat from a large number of rejections. The subsequent seed prompts

receive these compact representations and the full parent idea is fed to the Ideator only for repair or refinement.

Search Procedure. For each $b = 1, \ldots, B$, the search completes a bounded exploration of one seed before generating the next: $I_b^{(0)} \longrightarrow [\text{REPAIR}] \longrightarrow [\text{REFINE}] \longrightarrow I_{b+1}^{(0)}$, where the bracketed operations are conditional.

STAGE I – GENERATE AND ASSESS. Each fresh-seed prompt asks the Ideator for a problem or mechanism distinct from all ideas represented in $\mathcal{A}$, $\mathcal{H}$, $\mathcal{Q}$, and $\mathcal{R}$. If the preceding lineage ends as a historical duplicate or matches with a rejected pattern/logic, the prompt is augmented with an explicit instruction to pivot to abandon that failing mechanism or problematic area in its next generation. Once a draft is produced, the Steno and evaluators produce its structured quintuple summary, quality scores, soundness panel, and archive-relative diversity judgment. Based on the evaluator assessments, we define the core qualification gate as: $G(I) = \mathbb{1}\big[(N_I \geq \tau_N \wedge S_I \geq \tau_S \wedge C_I \geq \tau_C) \wedge \neg\,\text{invalid}(I) \wedge \neg\,\text{disputed}(I)\big]$. The primary search uses $\tau_N = \tau_S = \tau_C = 60$ and a diversity floor $\tau_D = 60$. Feasibility is measured, but excluded from the gate, as an ambitious and sound mechanism may be valuable, despite its immediate viability.

STAGE II – ROUTE AND REPAIR. The Diversity Judge first determines the candidate's relation to $\mathcal{A}$ and $\mathcal{H}$, after which the controller applies four cases: **(1)** An idea is a *purely historical match* if it matches only historical mechanisms, but not the active refined successors in their lineage. Such candidates are sub-optimal and rejected. Candidates matching multiple active ideas are also rejected because they cannot be assigned unambiguously to a single lineage. **(2)** A candidate matching exactly one active idea is treated as another version of that lineage rather than a new discovery. If it passes G, it may replace the matched idea based on the aggregated quality score; otherwise, it is rejected. **(3)** A candidate with no duplicate matches, but $D_I < \tau_D$ indicates a combination or reuse of multiple pre-qualified ideas without exactly matching one. Such candidates are rejected for insufficient diversity. Repair is ineffective here because it is instructed to preserve the central mechanism. **(4)** Finally, a candidate with $D_I \geq \tau_D$ and no duplicate matches is admitted to the active set if it passes G. Otherwise, if it falls within $\delta = 20$ points of the non-obviousness, soundness, and clarity thresholds, and is not *invalid* under the Soundness Panel evaluation, it is eligible for *one* repair. A *disputed* candidate may bypass the numerical soundness margin but must satisfy the same diversity and non-obviousness/clarity proximity requirements.

For repair, the Critic receives the full idea, its failed rubrics, quality evidence, soundness objections, and a list of competing or similar mechanisms. The Ideator then returns a corrected child within the same lineage. The child is re-evaluated and must satisfy the same qualification and diversity requirements. If the repair fails, the original seed is retained, and its failure is summarized in $\mathcal{R}$.

Candidates that qualify compete in $\mathcal{A}$ according to $Q_b(I) = 0.7N_I + 0.2S_I + 0.1C_I$. This weighting emphasizes the innate non-triviality of an idea, which is difficult to achieve through refinement or repair. A new mechanism enters $\mathcal{A}$ when capacity is available. If the archive is full, it replaces the weakest active idea only if its Q_b is higher, with ties broken by Feasibility (F_I). When an accepted idea leaves $\mathcal{A}$, its summary is stored in $\mathcal{H}$, not $\mathcal{R}$, since it remains a valid, high-quality idea.

STAGE III – REFINE ACCEPTED IDEAS. An accepted idea is eligible for refinement while it remains below a soft target: $N_I < 90$, $S_I < 80$, or $C_I < 80$. These targets serve only as refinement conditions and are not additional acceptance gates. The Critic addresses the relevant deficit, and the Ideator produces a child within the same lineage. During diversity assessment, its own parent is naturally excluded to avoid trivial lineage comparisons.

The child replaces its parent only if it passes G, clears τ_D relative to foreign mechanisms, and improves Q_b. Further, it must also pass a blinded parent–child non-obviousness comparison performed twice with their positions exchanged to account for position bias. If both comparisons prefer the same candidate, the winner is retained; otherwise, position disagreement is treated as a tie. This safeguard prevents improvements in clarity or soundness from trivializing the central mechanism. A non-improving child is discarded and does not overwrite its accepted parent, while a replaced parent is saved in $\mathcal{H}$. Feasibility remains a soft check throughout the process and is used only to break

ties when Q_b matches. The Ideator provides a plausible implementation and evaluation path, but feasibility neither triggers repair or refinement nor is included in the Critic's directed feedback.

The primary configuration uses $B = 10$ fresh seeds and active capacity $C_{\mathcal{A}} = 10$. Repair and accepted refinement share an allowance of at most $K_{\text{aux}} = 2$ auxiliary drafts per seed, with at most one repair and two accepted refinements. A seed repaired once therefore has room for at most one subsequent refinement, whereas an immediately accepted seed may receive two. The search ends after all B fresh seeds and the final auxiliary chain completes. The resulting $\mathcal{A}$ contains the final set of ideas generated by IDEAGENT, which are used for downstream evaluation.

4 BACKGROUND CORPUS COLLECTION

We follow the data collection procedure of Sinhahajari et al. (2026) to extract background papers from multiple arXiv `CS` domains and build the background corpus $\mathcal{X}$. Our objective is to generate a diverse set of non-trivial, sound research ideas that could *potentially* emerge from a shared set of background papers serving as a seed knowledge bank. To this end, we begin with published papers from each domain and identify the broader set of papers that *might* have influenced them. For each domain, papers published within a predefined time frame are collected and filtered by publication status and ranking. We then independently extract their *influential citations* (ICs) (Sinhahajari et al., 2026) and group papers according to IC overlap. Specifically, two papers belong to the same group if the Intersection-over-Union (IoU) of their ICs exceeds 0.5. The union of all ICs in a group forms its set of *background papers*. Next, we retain only groups of *background papers* supporting at least *three* published papers, ensuring that each knowledge bank has demonstrated the ability to support multiple high-quality research ideas. We define each such collection of background papers as a *topic*, representing a uniquely identifiable sub-domain or task. Topics containing between 5 and 8 background papers are retained and assigned descriptive tags using an LLM. Finally, we set a budget of $\Gamma = 6$ and apply greedy maximum diversity selection (Algorithm 2) to obtain the final collection of topics is provided in Table 16.

5 EVALUATION

In line with our QD formulation of research ideation, we evaluate each method on the returned *set* of N ideas per topic rather than in isolation. A valid set must satisfy two complementary requirements: ideas must individually be high-quality while remaining mutually diverse. To capture these, our evaluation protocol uses an LLM judge to provide independent per-idea *Quality* scores and pairwise *Diversity* metrics, which are then combined as follows[1]:

Quality Axis We evaluate each idea along three primary† and two secondary quality axes:

1. **Non-obviousness**† (NB_I) measures the originality of the core contribution. The judge evaluates the idea according to *Problem-Mechanism pairing*, *Causal Diagnosis*, and *Intervention*, which assess whether the idea would *surprise* a knowledgeable human researcher while ignoring presentation differences.
2. **Soundness**† (S_I) measures whether the proposed mechanism works as claimed and whether the logical flow contains internal contradictions or relies on non-rigorous or untested assumptions. A safe, unoriginal idea can still score low if its reasoning contains gaps, while a novel idea can score high if its logic is internally consistent.
3. **Mechanism Clarity** (C_I) measures how concretely and unambiguously the mechanism is specified, and whether someone else could implement it from the description without inventing missing pieces. This is evaluated independently of whether the mechanism is correct (soundness) or practical to build (feasibility).
4. **Feasibility** measures how practical it would be to build and test the idea under realistic constraints including data, compute, existing tools, and time. It ignores whether the idea is correct or important, as an idea can be highly feasible yet unsound, or vice versa.

[1]The exact scale and descriptions are provided in Appendix D

5. **Significance** measures the potential impact if the idea worked exactly as claimed, both on the target problem and the broader field. The judge assumes the mechanism works as claimed and ignores its feasibility or maturity.

Diversity Axis Given a set of N ideas, *Diversity* is evaluated pairwise. Each idea is independently decomposed into a semantic signature along eight axes: *failure mode, causal diagnosis, intervention, signal source, intervention locus, objective, evaluation regime,* and *assumption class*. Each of the $\binom{N}{2}$ idea pairs is then compared along all eight axes in a single judgment, producing a same/variant/related/distinct relation for each axis. A second call takes these axis-level relations as fixed evidence and produces a single score $D_{ij} \in \{0, \ldots, 9\}$ by weighting disagreement on the *core* axes—*failure mode, causal diagnosis,* and *intervention*—more heavily than the remaining supporting axes. For each topic T, we report the average pairwise diversity as $D(T) = \binom{N}{2}^{-1} \sum_i \sum_{i<j} D_{ij}$

Yield Our primary indicator is $\text{Yield}(\text{NB} \geq k, S \geq l, C \geq m, D \geq \tau)$ (or $Y_{k,l,m,\tau}$), which counts the largest number of *sound, clear, non-trivial*, and *mutually diverse* research ideas in a generated set. We first retain ideas satisfying the thresholds on *non-obviousness* (k), *soundness* (l), and *clarity* (m). We then extract the *largest* subset (maximum clique via branch-and-bound, Algorithm 1) in which every idea pair satisfies $D_{ij} \geq \tau$. Consequently, a set of N highly novel and sound, yet near-identical ideas correctly yields a score of 1 rather than N, penalizing methods that exploit quality by repeatedly paraphrasing a small number of successful concepts.

Successful Topics A topic T is successful at yield level Φ for a method if it generates at least Φ ideas satisfying the Yield criteria. We therefore report the proportion of successful topics as $|\mathcal{T}|^{-1} \sum_{T \in \mathcal{T}} \mathbb{1}[Y^T_{k,l,m,\tau} \geq \Phi]$. An ideal idea generator achieves a ratio of 1.

6 Experimental Setup

All experiments use proprietary large-scale models, with the latest `gpt-5.6-sol` serving as the Ideator. The Critic, Soundness Panel, and Quality Judge are built around `gemini-3.1-pro`, while the Diversity Judge and Steno use `gemini-3.5-flash` and `gemini-2.5-flash`, respectively. During evaluation, we use two external large-scale models, `claude-sonnet-5` and `claude-opus-4.7`, for all judgments. During analysis, we average their scores and treat the result as a *jury verdict* to account for judge bias and sampling stochasticity. Finally, we use the same idea budget ($B = 10$) for all baselines and report aggregated scores across all 32 topics.

Stateless. The simplest generation strategy, in which the Ideator produces B ideas independently without knowledge of one another.

One-Shot ($\mathcal{OS}$). The Ideator generates all B ideas in a single output. With a thinking-enabled Ideator, the model can use its reasoning space as shared memory for developing and diversifying ideas. Furthermore, each idea naturally conditions on all preceding ideas through causal generation, analogous to a rudimentary form of *stateful* generation.

Sequential-Memory ($\mathcal{SM}$). Building on this notion of statefulness, each of the B ideas is generated independently while cumulatively carrying forward the core summaries of previous ideas, allowing the model to retain a lightweight memory of prior generations for diversification. This setup is sequentially similar to IDEAGENT, but omits the quality-improvement subroutines (repair and refinement).

NOVA-inspired $\mathcal{SM}$. Building on Sequential-Memory, we augment it with NOVA's (Hu et al., 2025) iterative *seed-pool*-based germination and replacement strategy for diversification. We discard NOVA's online retrieval module, since our background papers constitute a closed knowledge base. Specifically, we run Sequential-Memory for 3 rounds, each generating B ideas. At the end of every round, an external judge reflects on the generated ideas and selects the 3 most promising as seed concepts, replacing those from the previous round. The next round germinates B new ideas from these seed directions together with the accumulated global memory. Finally, the Critic selects the best B ideas from the complete pool of $3B$ generations, emphasizing distinctness, to form the final

evaluation set. As before, each idea is generated in a single attempt without subsequent repair or refinement.

7 Results and Analysis

All the scores and numbers we present here are aggregated across all 32 topics.

IDEAGENT has a Higher Yield To directly evaluate our core hypothesis that IDEAGENT produces a higher density of high-quality ideas within a diverse set, we compare the Yield of its generated ideas against Sequential-Memory. We define a Yield surface as a 2D grid in which each cell represents the Yield (out of 10) at a fixed diversity and clarity threshold while varying the soundness and non-obviousness thresholds. Figure 1 presents a zoomed-in view of the high-quality region ($NB \geq 6, S \geq 6$), along with the full 10×10 matrix in Figure 2. The full grid reveals a broad Yield plateau for IDEAGENT under relaxed quality thresholds (top-left), where the Yield is 3.6 for IDEAGENT compared to 1.9 for Sequential-Memory, representing an improvement of $1.89\times$, at $(C \geq 6, D_{ij} \geq 7)$. As expected, the surface declines toward the bottom-right as the quality constraints become more stringent, eventually approaching zero for both methods when $NB \geq 8$ and $S \geq 8$.

Table 1: *Jury* averaged results across all 32 topics, with per-judge scores provided in the Supplementary. Bold marks the best method in each row. *Note:* $^{\dagger}p < .05$; $^{\ddagger}p < .001$; $^{\S}p < .0001$ (paired topic-level tests; rubric tests Holm-corrected)

Metric	$\mathcal{SM}$	$\mathcal{OS}$	Stateless	NOVA	IDEAGENT
QD OUTCOMES					
$Y_{7,7,6,7}$	0.188	$0.000^{\dagger}$	0.250	0.281	$\mathbf{1.094}^{\S}$
$Y_{6,7,6,7}$	0.531	$0.000^{\ddagger}$	$0.812^{\dagger}$	$1.156^{\dagger}$	$\mathbf{2.312}^{\S}$
SUCCESSFUL TOPICS					
$Y_{7,7,6,7} \geq 1$	6/32	0/32	7/32	8/32	**27/32**
$Y_{7,7,6,7} \geq 2$	0/32	0/32	1/32	1/32	**8/32**
$Y_{6,7,6,7} \geq 1$	14/32	0/32	24/32	25/32	**31/32**
$Y_{6,7,6,7} \geq 2$	3/32	0/32	2/32	11/32	**23/32**
$Y_{6,7,6,7} \geq 3$	0/32	0/32	0/32	1/32	**15/32**
INDIVIDUAL DIMENSIONS					
NB	6.020	$5.609^{\S}$	6.077	6.142	$\mathbf{6.430}^{\S}$
Soundness	6.534	$5.477^{\S}$	6.589	6.483	$\mathbf{6.720}^{\dagger}$
Clarity	5.416	$4.030^{\S}$	$5.739^{\dagger}$	5.556	$\mathbf{6.341}^{\S}$
Diversity	6.606	$6.739^{\dagger}$	$5.080^{\ddagger}$	$\mathbf{6.757}^{\dagger}$	6.641
Feasibility	4.053	4.034	4.006	**4.150**	3.864
Significance	5.683	$5.316^{\S}$	$\mathbf{5.898}^{\dagger}$	$5.470^{\dagger}$	5.572

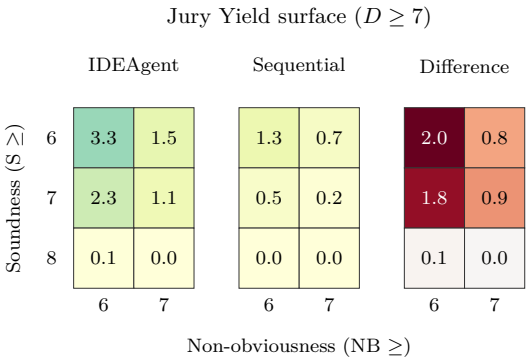


Figure 1: *Jury* averaged Yield surface for IDEAGENT, Sequential-Memory baseline, and their difference.

The primary difference between the two setups becomes evident in the zoomed-in region ($NB \in [6, 7]$, $S \in [6, 8]$). Here, as the quality thresholds become stricter, the Yield decreases from 3.2 (at $NB \geq 6$, $S \geq 6$) to 1.1 (at $NB \geq 7$, $S \geq 7$), before eventually collapsing to zero. At the moderate quality threshold ($NB \geq 6$, $S \geq 6$), IDEAGENT achieves its largest improvement over the baseline, with a gain of $+1.9$. Even when the quality requirements are tightened to ($NB \geq 7$, $S \geq 7$), the gap remains substantial at $+0.9$, demonstrating its ability to generate non-obvious, sound, and mutually diverse research ideas. The advantage is also present under each evaluator separately as IDEAGENT has the highest mean Yield under both `claude-opus-4.7` and `claude-sonnet-5` at $NB \geq 6$ and $NB \geq 7$ (Table 4). Against Sequential Memory, the topic-bootstrap 95% confidence intervals for the jury-averaged mean differences are $[1.375, 2.188]$ at $NB \geq 6$ and $[0.656, 1.156]$ at $NB \geq 7$.

IDEAGENT vs. Baselines IDEAGENT also outperforms the secondary baseline, NOVA, as shown in Table 1, with per-judge results reported in the Appendix. The improvement in Yield is a clear $2\times$ and $3.89\times$ at $NB \geq 6$ and $NB \geq 7$, respectively. Furthermore, the three primary quality metrics all achieve their highest scores under IDEAGENT. The highest pairwise diversity is naturally achieved by NOVA owing to its final critic-based selection process, which explicitly optimizes for distinctness among the $3B$ candiates.

This is followed by the One-Shot baseline, where all ideas share a common thinking space during generation, unlike the sequential approaches, which rely on a compact memory representation that is inherently susceptible to some information loss. However, over longer runs, compact signatures are substantially more computationally efficient for both comparison and context accumulation. Interestingly, One-Shot achieves a Yield of zero, indicating that none of its generated ideas satisfy the required thresholds. This suggests that maintaining a fully shared memory during generation can also be detrimental, as a poor or low-quality idea may directly influence subsequent generations and propagate its weaknesses. We therefore hypothesize that a lightweight sequential core-memory mechanism is sufficient to induce meaningful diversity while avoiding the risks of excessive cross-contamination between ideas. For the proportion of successful topics, Table 1 shows that IDEAGENT has a clear edge over all baselines, indicating that it can produce qualifying ideas across a broader range of topics and quality–diversity gates. At $D \geq 7$, $S \geq 7$, $C \geq 6$, and $\mathrm{NB} \geq 6$, the best baseline produces at least one qualifying idea on 25 of 32 topics, compared with 31 of 32 for IDEAGENT. When the non-obviousness threshold is increased to $\mathrm{NB} \geq 7$, this difference becomes even sharper: the best baseline succeeds on 8 of 32 topics, whereas IDEAGENT succeeds on 27 of 32. Moreover, Table 1 shows that the baselines attain substantially lower coverage when success requires multiple qualifying ideas. For example, at $Y_{6,7,6,7}$, NOVA produces at least two diverse, high-quality ideas on only 11 of 32 topics, whereas IDEAGENT does so on 23 of 32. The same pattern holds across the other reported quality-diversity gates. Taken together, these results reveal two complementary limitations of the strongest baselines: their qualifying outputs are often limited to a single idea rather than multiple diverse, high-quality ideas for the same topic, and their topic coverage falls sharply under the stricter non-obviousness threshold.

Quality improvements due to repair and refinement The primary distinction between the baselines and IDEAGENT is the inclusion of quality-improvement subroutines in the form of repair and refinement, which we analyze across all 320 lineages. Table 2 illustrates the resulting quality improvements as assessed by the internal judge agents during generation. Specifically, we compare the initial seed idea with its final version after repair or refinement. Of the 320 lineages, only 30 required repair after falling just below the qualification thresholds, of which 28 were successfully improved and ultimately *qualified*. Similarly, 182 of the 320 lineages underwent refinement after qualifying, and in a substantial 83% of cases, the refined version replaced the original idea, showing that the refined version was selected in most cases.

Table 2: Quality improvements from repair and refinement.

		Score		Lineages (%)		
		Seed	Final	↑	=	↓
INTERNAL JUDGES						
Repair	NB	85.3	91.7	64.3	28.6	7.1
	Soundness	76.4	99.6	96.4	3.6	0.0
	Clarity	76.4	93.1	78.6	14.3	7.1
	$Y_{60,60,60,60}$	8.53	9.41		–	
	$Y_{70,60,60,60}$	8.44	9.31		–	
Refinement	NB	83.9	91.3	86.0	13.3	0.7
	Soundness	94.3	99.4	67.8	28.0	4.2
	Clarity	84.5	93.4	80.4	16.8	2.8
	$Y_{60,60,60,60}$	8.53	8.53		–	
	$Y_{70,60,60,60}$	8.44	8.53		–	
EXTERNAL JURY						
Repair	NB	6.44	6.62	35.7	39.3	25.0
	Soundness	6.39	6.88	53.6	35.7	10.7
	Clarity	5.29	6.92	85.7	14.3	0.0
	$Y_{6,7,6,7}$	1.12	1.38		–	
	$Y_{7,7,6,7}$	0.19	0.34		–	
Refinement	NB	6.08	6.48	58.0	31.5	10.5
	Soundness	6.62	6.81	51.0	34.3	14.7
	Clarity	5.70	6.90	83.9	11.9	4.2
	$Y_{6,7,6,7}$	1.12	2.00		–	
	$Y_{7,7,6,7}$	0.19	0.97		–	

Taking a closer look, we analyze improvements across the individual quality rubrics together with the resulting Yield. Since refinement is applied only to ideas that have already qualified, whereas repair is reserved for ideas that narrowly miss the thresholds, refinement candidates naturally begin with higher absolute scores. Soundness benefits the most from both repair and refinement, reaching nearly 100% qualification in both cases, with repair yielding a maximum improvement of $+23.2$. This suggests that targeted corrective feedback is highly effective at resolving logical inconsistencies. Clarity exhibits the second-largest improvement from repair ($+16.7$), while refinement improves it the most relative to the other rubrics, yielding an average gain of $+8.9$. Importantly, these improvements in soundness and clarity do not come at the expense of non-obviousness. Instead, both repair and refinement improve non-obviousness by approximately $+7$ points, indicating that gains across the different quality dimensions are complementary rather than competitive.

An interesting observation is that the final ideas produced by either process converge to remarkably similar scores across all quality rubrics. We hypothesize that this reflects an upper bound on the improvements achievable through targeted textual feedback, with diminishing returns after multiple refinement iterations. Finally, when analyzing Yield at the chosen thresholds, repair provides gains

of $+0.88$ and $+0.87$ at the two non-obviousness thresholds, respectively, whereas refinement yields comparatively smaller gains.

It is also important to verify that the improvements observed by the internal agents during generation are reflected in downstream evaluation. Table 2 also presents the corresponding analysis using the external judges. Although both sets of judges consistently agree on the direction of the improvements, the absolute gains differ substantially because of differences in their scoring scales. Clarity exhibits the largest downstream improvement, with gains of $+1.63$ and $+1.20$ points from repair and refinement, respectively, suggesting that improvements in clarity are readily recognized by independent evaluators. In contrast to the internal assessments during generation, the external judges reverse the relative effectiveness of repair and refinement. Refinement contributes the largest gains in Yield, improving it by $+0.88$ and $+0.78$ at the NB thresholds of 6 and 7, respectively. Finally, we refrain from directly correlating score deltas between the internal and external evaluations because the two systems use different rubrics and scoring scales. Instead, we emphasize their qualitative agreement regarding the consistent persistence of the observed improvements.

Inter-Annotator Agreement To establish the reliability of our evaluation framework, we analyze the inter-annotator agreement between the two judges. As shown in Table 3, agreement per rubric is computed using *linear-weighted* Cohen's Kappa on the pooled evaluations across all topics and runs. Across the primary rubrics, soundness shows the lowest agreement (0.26), demonstrating that judging logical and mathematical consistency of mechanisms and assumptions remains subjective (Ho et al., 2026). Clarity achieves the highest agreement (0.60), indicating that clear ideas are readily identified by the models. The remaining core rubrics, Non-obviousness and Diversity, exhibit moderate agreement of ~ 0.43, providing a reasonable indicator of the jury's internal alignment. Taken together, agreement is consistently higher for rubrics with concrete, checkable referents in an idea's text (clarity, non-obviousness, diversity) than for those requiring more holistic or forward-looking judgment (feasibility, significance, and to a lesser extent, soundness), suggesting that judge disagreement reflects each rubric's inherent ambiguity rather than a general unreliability of the evaluation framework. The Spearman column in Table 3 reports whether the two evaluators rank the same ideas relatively higher or lower. Rank correlations are positive across all rubrics, ranging from 0.409 for soundness to 0.761 for feasibility. According to Schober et al. (2018), these values indicate moderate rank consistency for non-obviousness, soundness, mechanism clarity, significance, and pairwise diversity, and strong rank consistency for feasibility. Interestingly, feasibility has low absolute-score agreement ($\kappa = 0.210$) but strong rank consistency ($\rho = 0.761$), indicating that the judges use different score levels while largely agreeing on which ideas are relatively more or less feasible.

Table 3: Agreement b/w `claude-opus-4.7` and `claude-sonnet-5` across all methods and topics.

Rubric	Linear-weighted κ	Median Spearman ρ
NB	0.455	0.610
Soundness	0.268	0.409
Clarity	**0.604**	0.600
Feasibility	0.210	**0.761**
Significance	0.352	0.504
Diversity	0.424	0.679

8 Conclusion

In this work, we presented IdeaAgent, a multi-agent framework that reframes scientific ideation as a QD search to build *portfolios of ideas* with high quality density. Through sequential generation and multi-agent evaluation centered on the core requirements of QD search, we introduced a quality-improvement subroutine that repairs or optionally refines ideas based on identified *opportunities*. We further demonstrated lineage-inspired idea management using dedicated archives for completed, rejected, and historical ideas, efficiently enabled by core signature comparison. To jointly evaluate quality and diversity, we proposed Yield, which filters ideas using fixed thresholds on its primary quality indicators before selecting the largest subset whose ideas satisfy a pairwise diversity threshold. Finally, we evaluated IdeaAgent across 32 Computer Science research topics and demonstrated the contributions of both the quality-improvement loop and compact stateful memory. We hope our problem formulation provides a new perspective on AI-driven scientific ideation.

AI USE STATEMENT

In this work, we have used AI for code writing and repository management. All the code written by AI was checked and verified to run correctly and as required by two authors of this paper. Additionally, we used AI for initial brainstorming, i.e., for quick idea testing and verification, and not full research problem generation. Final draft grammatical polishing and rephrasing were done with the help of AI, but the main content was still written by the authors. All the related work and research gaps were identified by the authors themselves without the use of AI anywhere. Finally, we take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

In this work, we study and develop an agentic framework for automated idea generation and scientific discovery. However, the ideas generated might be incomplete, unsafe, illegal, impractical, or overlapping with existing works. Hence, we strongly argue and resonate with the fact that all the AI-generated ideas should be proofread and verified by humans before working on them further in a **practical deployment**.

REPRODUCIBILITY STATEMENT

Our entire codebase was developed **exactly** around the setup described in the paper. Further, we plan to completely open-source it, along with the background data used for idea generation. All the papers used from the background data collection were publicly available on the web, and directly scraped with no private modifications. The Agents used in the paper are also publicly available as APIs from the respective providers. However, the exact generation outputs might vary slightly depending on the reasoning-effort and sampling strategies used by them internally, and we do not have any control over it. All the data processing and main methodology pieces are well-described in the main paper, and no significant section is withheld. The exact algorithms used for topic selection and Yield are also described as is in the appendix.

REFERENCES

Alexis Audran-Reiss, Jordi Armengol-Estapé, Karen Hambardzumyan, Amar Budhiraja, Martin Josifoski, Edan Toledo, Rishi Hazra, Despoina Magka, Michael Shvartsman, Parth Pathak, Justine T Kao, Lucia Cipolina-Kun, Bhavul Gauri, Jean-Christophe Gagnon-Audet, Emanuel Tewolde, Jenny Zhang, Taco Cohen, Yossi Adi, Tatiana Shavrina, and Yoram Bachrach. What does it take to be a good ai research agent? studying the role of ideation diversity, 2025. URL https://arxiv.org/abs/2511.15593.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. ResearchAgent: Iterative research idea generation over scientific literature with large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 6709–6738, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.342. URL https://aclanthology.org/2025.naacl-long.342/.

Herbie Bradley, Andrew Dai, Hannah Benita Teufel, Jenny Zhang, Koen Oostermeijer, Marco Bellagente, Jeff Clune, Kenneth Stanley, Gregory Schott, and Joel Lehman. Quality-diversity through AI feedback. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=owokKCrGYr.

Antoine Cully, Jeff Clune, Danesh Tarapore, and Jean-Baptiste Mouret. Robots that can adapt like animals. *Nature*, 521(7553):503–507, May 2015. ISSN 1476-4687. doi: 10.1038/nature14422. URL https://doi.org/10.1038/nature14422.

Matthew Fontaine and Stefanos Nikolaidis. Differentiable quality diversity. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances*

in Neural Information Processing Systems, volume 34, pp. 10040–10052. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/532923f11ac97d3e7cb0130315b067dc-Paper.pdf.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomašev, Dina Zverinski, Ivor Rendulic, Elahe Vedadi, Florian Hasler, Luka Rimanic, Marina Boia, Ivan Budiselic, Ben Feinstein, Mathias Bellaiche, Tom Sheffer, Jan Freyberg, Jeremy Ratcliff, Ottavia Bertolli, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, José R. Penadés, Gary Peltz, Yossi Matias, James Manyika, Demis Hassabis, Yunhan Xu, Pushmeet Kohli, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Accelerating scientific discovery with Co-Scientist. *Nature*, 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10644-y. URL https://doi.org/10.1038/s41586-026-10644-y.

Sy-Tuyen Ho, Minghui Liu, Huy Nghiem, and Furong Huang. Soundnessbench: Can your ai scientist really tell good research ideas from bad ones?, 2026. URL https://arxiv.org/abs/2605.30329.

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 framework for enhancing scientific innovation with large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21330–21359, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1099. URL https://aclanthology.org/2025.findings-acl.1099/.

Joel Lehman and Kenneth O. Stanley. Evolving a diversity of virtual creatures through novelty search and local competition. In *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation*, GECCO '11, pp. 211–218, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450305570. doi: 10.1145/2001576.2001606. URL https://doi.org/10.1145/2001576.2001606.

Joel Lehman, Jonathan Gordon, Shawn Jain, Kamal Ndousse, Cathy Yeh, and Kenneth O. Stanley. *Evolution Through Large Models*, pp. 331–366. Springer Nature Singapore, Singapore, 2024. ISBN 978-981-99-3814-8. doi: 10.1007/978-981-99-3814-8_11. URL https://doi.org/10.1007/978-981-99-3814-8_11.

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. Chain of ideas: Revolutionizing research via novel idea development with LLM agents. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 8971–9004, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.477. URL https://aclanthology.org/2025.findings-emnlp.477/.

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.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites, 2015. URL https://arxiv.org/abs/1504.04909.

Justin K. Pugh, Lisa B. Soros, and Kenneth O. Stanley. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, Volume 3 - 2016, 2016. ISSN 2296-9144. doi: 10.3389/frobt.2016.00040. URL https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2016.00040.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S. Weld. Scideator: Human-llm compound system for scientific ideation through facet recombination

and novelty evaluation. In *Proceedings of the ACM Conference on AI and Agentic Systems*, CAIS '26, pp. 348–374, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400724152. doi: 10.1145/3786335.3813161. URL https://doi.org/10.1145/3786335.3813161.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using LLM agents as research assistants. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5977–6043, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.320. URL https://aclanthology.org/2025.findings-emnlp.320/.

Patrick Schober, Christa Boer, and Lothar A. Schwarte. Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768, 2018. doi: 10.1213/ANE.0000000000002864.

Soumitra Sinhahajari, Navonil Majumder, and Soujanya Poria. On the limits of llm-as-judge for scientific novelty assessment, 2026. URL https://arxiv.org/abs/2606.12071.

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 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28201–28240, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1368. URL https://aclanthology.org/2025.acl-long.1368/.

Yixuan Tang and Yi Yang. Ai research agents narrow scientific exploration, 2026. URL https://arxiv.org/abs/2605.27905.

Edan Toledo, Karen Hambardzumyan, Martin Josifoski, RISHI HAZRA, Nicolas Baldwin, Alexis Audran-Reiss, Michael Kuchnik, Despoina Magka, Minqi Jiang, Alisia Maria Lupidi, Andrei Lupu, Roberta Raileanu, Tatiana Shavrina, Kelvin Niu, Jean-Christophe Gagnon-Audet, Michael Shvartsman, Shagun Sodhani, Alexander H Miller, Abhishek Charnalia, Derek Dunfield, Carole-Jean Wu, Pontus Stenetorp, Nicola Cancedda, Jakob Nicolaus Foerster, and Yoram Bachrach. AI research agents for machine learning: Search, exploration, and generalization in MLE-bench. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=RwfrdKSgCE.

Keisuke Ueda, Wataru Hirota, Takuto Asakura, Takahiro Omi, Kosuke Takahashi, Kosuke Arima, and Tatsuya Ishigaki. Exploring the design of multi-agent LLM dialogues for research ideation. In Frédéric Béchet, Fabrice Lefèvre, Nicholas Asher, Seokhwan Kim, and Teva Merlin (eds.), *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 322–337, Avignon, France, August 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.sigdial-1.26/.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search, 2025. URL https://arxiv.org/abs/2504.08066.

Keyu Zhao, Weiquan Lin, Qirui Zheng, Fengli Xu, and Yong Li. Deep ideation: Designing llm agents to generate novel research ideas on scientific concept network, 2025. URL https://arxiv.org/abs/2511.02238.

A Limitations and Scope of Work

Our work has several limitations that warrant discussion.

1. All internal and external evaluations rely primarily on LLM judges to assess non-obviousness, soundness, clarity, feasibility, and diversity. To mitigate evaluator-specific bias, we employ two independent large-scale judge models and average their scores. Likewise, to calibrate the crucial soundness evaluation, we aggregate five independently sampled scores to mitigate sampling bias. Moreover, our primary metric, Yield, is model-agnostic and can be paired with alternative evaluation protocols, including human judgments.
2. We restrict this work to generating portfolios of diverse, high-quality research ideas and evaluating them accordingly. Consequently, we do not evaluate **real-world** feasibility, exact correctness, or practical effectiveness of the generated ideas, which would require implementation, debugging, end-to-end experimentation, and expert validation. Likewise, we **do not** claim that the generated ideas in our experiments are necessarily immediately publishable or absent from the existing literature. Verifying such claims would require a comprehensive retrieval over both published and emerging research to identify related prior work, along with a significantly increased monetary investment beyond our scope. However, we do show that our ideas exhibit strong soundness and logical consistency across multiple rigorous evaluations.
3. Each idea is given one repair opportunity and at most two refinements, if one is not consumed by repair. Both repair and refinement are computationally expensive because they require all agents (Quality, Diversity, Soundness, Ideator, and Critic). Since our setup uses proprietary LLMs, additional rounds would substantially increase monetary cost. More generally, we do not rigorously optimize hyperparameters (rubric thresholds, iteration counts, and API generation settings) because of computational, monetary, and time constraints. Consequently, the reported results may be under-optimized. Nevertheless, our goal is to demonstrate the value of repair and refinement for improving ideas along the Quality axis.
4. All experiments use proprietary LLMs accessed through their APIs because we found that **current** smaller open-source models lacked sufficient knowledge, idea formulation/refinement, and evaluation capability. However, our framework is model-agnostic, and we expect the underlying principles of sequential generation, targeted multi-dimensional evaluation, and feedback to transfer across future models. In preliminary experiments, we evaluated `Gemma-4-31B`, `Qwen-3.6-27B`, `GPT-OSS-120B`, `gemini-3.1-pro`, `claude-sonnet-5`, and `claude-opus-4.8`. These models generally produced non-obvious, diverse ideas with clearly articulated mechanisms; however, the evaluators *frequently* identified serious soundness flaws, substantially reducing overall quality. `claude-fable-5` generated moderately sound, highly non-obvious, and diverse ideas, but budget constraints precluded its inclusion in the main experiments. In a pilot run on `cs.AI-001`, the internal judge assigned its ideas a mean soundness score of 66.4, compared with 91.8 for `gpt-5.6-sol`, triggering several repairs of the initial seed drafts. Relative to `gpt-5.6-sol`, `claude-fable-5` achieved a 9-point advantage in non-obviousness but scored three points lower in diversity. Thus, despite its novelty advantage, its lower initial soundness increased the required repair effort and, together with the budget constraints, motivated our decision not to use it as the primary Ideator.
5. Due to the authors belonging to the CS and AI domain, we restrict the data collection and experiments to them. But we strongly suggest that the proposed method is generic and can also be transposed to other sciences such as Chemistry, Biology, or Physics.
6. Most importantly, the motivation of this work is **not** to claim SoTA in any manner, but just present a **proof-of-concept** for the new QD formulation of the ideation process.

B Additional Results and Discussions

Human evaluation. As a small comparative check against human ratings, two authors scored 25 topic-matched idea pairs on the same three quality rubrics used by the Claude jury, blind to which system produced each idea. Table 5 shows that human raters independently reproduce the automatic ranking. IDEAGENT is preferred on 69.6% of topic-matched pairs ($p < .05$), and leads on all three rubrics, and on no rubric does NOVA lead under either jury. The magnitude of the advantage is also consistent across juries: the human mean margin ($+0.47$) closely matches the Claude jury's ($+0.52$), despite the two juries operating on different absolute scales, with human raters scoring roughly 0.7 points lower overall.

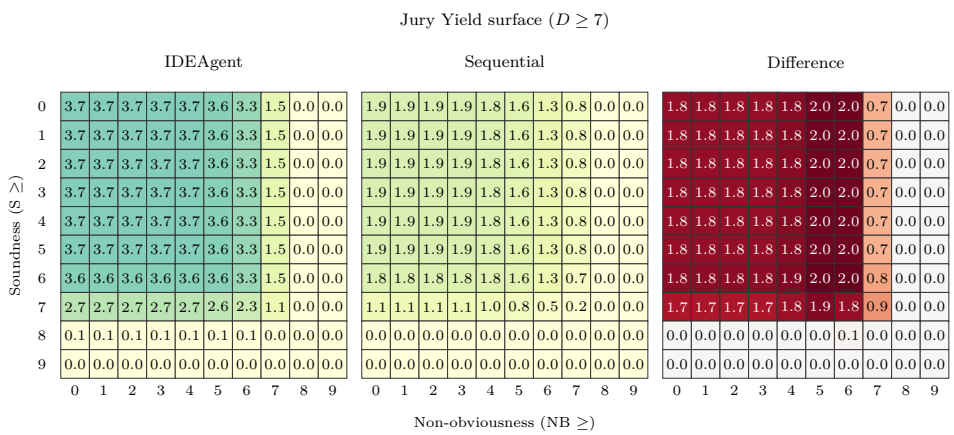


Figure 2: Full Yield surfaces for IDEAGENT, Sequential-Memory baseline, and their difference.

Table 4: Results across 32 topics as evaluated by the two external judges. Bold marks the best method in each row. *Note.* $^{\dagger}p < .05$; $^{\ddagger}p < .001$; $^{\S}p < .0001$ (paired topic-level tests; rubric tests Holm-corrected).

Metric	$\mathcal{SM}$	$\mathcal{OS}$	Stateless	NOVA	IDEAGENT
QD OUTCOMES					
$Y_{7,7,6,7}$	1.312	$0.031^{\S}$	1.312	1.562	$\mathbf{3.531}^{\S}$
$Y_{6,7,6,7}$	2.375	$0.031^{\S}$	$1.719^{\dagger}$	2.844	$\mathbf{4.656}^{\S}$
SUCCESSFUL TOPICS					
$Y_{7,7,6,7} \geq 1$	22/32	1/32	28/32	23/32	**32/32**
$Y_{7,7,6,7} \geq 2$	14/32	0/32	12/32	17/32	**31/32**
$Y_{7,7,6,7} \geq 3$	6/32	0/32	2/32	7/32	**26/32**
$Y_{6,7,6,7} \geq 1$	27/32	1/32	31/32	31/32	**32/32**
$Y_{6,7,6,7} \geq 2$	21/32	0/32	19/32	27/32	**31/32**
$Y_{6,7,6,7} \geq 3$	17/32	0/32	5/32	18/32	**30/32**
INDIVIDUAL DIMENSIONS					
NB	6.175	$5.609^{\S}$	6.206	6.331	$\mathbf{6.612}^{\S}$
Soundness	6.862	$5.794^{\S}$	6.891	6.834	**6.978**
Clarity	5.447	$3.878^{\S}$	5.728	5.597	$\mathbf{6.353}^{\S}$
Diversity	7.062	$7.190^{\dagger}$	$5.309^{\S}$	$\mathbf{7.268}^{\dagger}$	7.047
Feasibility	4.862	4.772	4.847	**5.006**	4.637
Significance	5.975	$5.534^{\S}$	**6.137**	$5.775^{\dagger}$	$5.822^{\dagger}$

(a) `claude-opus-4.7`

Metric	$\mathcal{SM}$	$\mathcal{OS}$	Stateless	NOVA	IDEAGENT
QD OUTCOMES					
$Y_{7,7,6,7}$	0.406	$0.000^{\ddagger}$	0.375	0.500	$\mathbf{1.219}^{\S}$
$Y_{6,7,6,7}$	0.844	$0.031^{\S}$	1.000	$1.250^{\dagger}$	$\mathbf{2.406}^{\S}$
SUCCESSFUL TOPICS					
$Y_{7,7,6,7} \geq 1$	12/32	0/32	11/32	14/32	**28/32**
$Y_{7,7,6,7} \geq 2$	1/32	0/32	1/32	2/32	**10/32**
$Y_{6,7,6,7} \geq 1$	20/32	1/32	28/32	28/32	**31/32**
$Y_{6,7,6,7} \geq 2$	6/32	0/32	4/32	11/32	**26/32**
$Y_{6,7,6,7} \geq 3$	1/32	0/32	0/32	1/32	**15/32**
INDIVIDUAL DIMENSIONS					
NB	5.866	$5.609^{\dagger}$	5.947	5.953	$\mathbf{6.247}^{\S}$
Soundness	6.206	$5.159^{\S}$	6.287	6.131	$\mathbf{6.463}^{\ddagger}$
Clarity	5.384	$4.181^{\S}$	$5.750^{\dagger}$	5.516	$\mathbf{6.328}^{\S}$
Diversity	6.150	**6.288**	$4.850^{\S}$	6.246	6.234
Feasibility	3.244	**3.297**	3.166	3.294	3.091
Significance	5.391	$5.097^{\ddagger}$	$\mathbf{5.659}^{\dagger}$	$5.166^{\dagger}$	5.322

(b) `claude-sonnet-5`

Table 6 reports the joint distribution of per-topic verdicts. The juries jointly favour IDEAGENT on 12 of 25 topics and jointly favour NOVA on only one. Where they differ, the disagreement is clearly one-sided, i.e., on 6 topics the human jury awards the pair to NOVA while the Claude jury awards it to IDEAGENT, against 2 in the opposite direction. Hence, the human jury is the more conservative of the two, returning a decisive verdict for IDEAGENT on 16 topics against the Claude jury's 20, rather than being systematically opposed to it. Both juries converge on the *same overall conclusion* from independent evidence, with the human study providing the stricter test.

We note that this study uses two raters over 25 pairs and is aimed to test the overall ranking rather than per-rubric differences. The individual rubric margins are directionally consistent but do not reach significance under the human jury.

Table 5: Human evaluation of IDEAGENT against Nova, the strongest baseline, alongside the Claude jury scoring the same ideas. Two expert raters scored 25 topic-matched idea pairs on a 0–9 rubric, blind to source. Win rate is the share of pairs on which IDEAGENT scores higher, excluding ties. Human values average both raters; Claude values average Claude Opus 4.7 and Claude Sonnet 5; Overall is the equal-weight mean of the three rubrics. $^{\dagger}p < .05$, $^{\ddagger}p < .001$ (Wilcoxon signed-rank over per-topic score differences).

	Human jury			Claude jury		
Rubric	IDEAGENT	NOVA	Win rate	IDEAGENT	NOVA	Win rate
NB	5.48	5.10	61.9%	6.44	6.18	63.2%
Soundness	5.68	5.18	62.5%	6.80	6.32	93.8%†
Mechanism clarity	6.38	5.86	70.0%	6.46	5.64	81.0%†
Overall	**5.85**	5.38	**69.6%**†	**6.57**	6.05	**87.0%**‡

Table 6: Joint distribution of per-topic verdicts. Each jury returns one verdict per topic—IDEAGENT wins, tie, or NOVA wins—given by the sign of its mean IDEAGENT minus NOVA score difference over the three rubrics. Both juries independently favor IDEAGENT on the large majority of topics ($20/25$ and $16/25$), and jointly favor NOVA on a single topic.

	Claude jury			
Human jury	IDEAGENT	Tie	NOVA	Total
IDEAGENT wins	**12**	2	2	16
Tie	2	0	0	2
NOVA wins	6	0	**1**	7
Total	20	2	3	25

C COST AND FUTHER SCALING

Cost / Successful Topic A natural question in research ideation is whether it is better to sample more ideas using cheaper methods or fewer ideas using more capable but expensive ones. To answer this, we report the Cost per Successful Topic (CST) of a method M, defined as the total generation cost across all attempted topics divided by the number of successful topics:

$$\mathrm{CST}_M(k, l, m, \tau, \Phi) = \frac{\sum_{T \in \text{Topics}} C_{M,T}}{\sum_{T \in \text{Topics}} \mathbb{1}\left[Y^{M,T}_{k,l,m,\tau} \geq \Phi\right]} \tag{1}$$

where $C_{M,T}$ is the generation cost of method M on topic T, and Φ is the required yield. $n(\text{Topics}) = 32$, $(l, m, \tau) = (7, 6, 7)$, $k \in \{6, 7\}$, and $\Phi \in \{1, 2, 3\}$. If no topic reaches the required yield, CST is reported as *NA*.

The objective of QD search is to generate diverse, high-quality ideas for each topic. The pairwise diversity constraint is meaningful only when at least two ideas qualify; consequently, $\Phi = 2$ is our primary diverse-portfolio criterion.

Stateless is the most cost-effective method when success requires only one qualifying idea, although this criterion does not exercise pairwise diversity. When success requires at least two diverse, high-quality ideas for the same topic, IDEAGENT becomes the most cost-effective method at both non-obviousness thresholds. Compared with the strongest baseline, it is approximately $1.6\times$ cheaper per successful topic while succeeding on $2.1\times$ as many topics at $\mathrm{NB} \geq 6$ and $8\times$ as many topics at $\mathrm{NB} \geq 7$. At the stricter requirement of three qualifying ideas, IDEAGENT is approximately $11\times$ cheaper than the only baseline that succeeds.

Scaling the Baselines We scaled the Stateless and Sequential Memory baselines to approximately match IDEAGENT's generation budget. Stateless was scaled from 10 to 50 fresh generations across all 32 topics, while Sequential Memory was scaled from 10 to 30 and 45 generations on a matched

Table 7: Cost (USD) per successful topic across 32 topics, evaluated by two external judges. Bold marks the cheapest method in each row; *NA* indicates the method never reached that yield threshold within budget.

Metric	$\mathcal{SM}$	$\mathcal{OS}$	Stateless	NOVA	IDEAGENT
COST / SUCCESSFUL TOPIC					
$Y_{7,7,6,7} \geq 1$	\$6.03	*NA*	**\$5.07**	\$16.32	\$6.52
$Y_{7,7,6,7} \geq 2$	*NA*	*NA*	\$35.52	\$130.56	**\$22.00**
$Y_{6,7,6,7} \geq 1$	\$2.58	*NA*	**\$1.48**	\$5.22	\$5.68
$Y_{6,7,6,7} \geq 2$	\$12.05	*NA*	\$17.76	\$11.87	**\$7.65**
$Y_{6,7,6,7} \geq 3$	*NA*	*NA*	*NA*	\$130.56	**\$11.73**

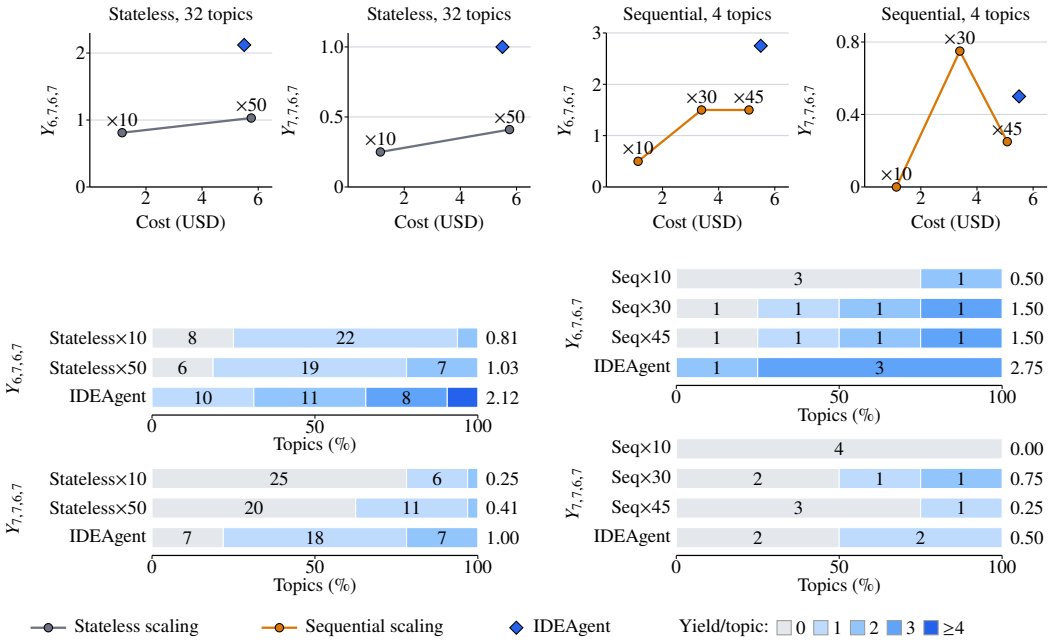


Figure 3: Scaling fresh generation does not reproduce IDEAGENT's quality–diversity performance. **Top:** cost–Yield at both quality gates, for Stateless on 32 matched topics and Sequential-Memory on four. **Bottom:** distribution of per-topic Yield; the number beside each bar is that arm's mean.

subset of four topics to examine its scaling behaviour. From each generation pool, up to 10 ideas were retained by `gemini-3.1-pro` using the selection criteria described above.

As shown in Figure 3, scaling Stateless from 10 to 50 generations increases mean Yield by approximately 27% at the NB ≥ 6 gate and 62% at the NB ≥ 7 gate. Nevertheless, IDEAGENT achieves approximately $2.06\times$ and $2.46\times$ higher Yield than the scaled Stateless baseline at the respective gates. The per-topic Yield distributions show the same trend: additional fresh generations improve Stateless but do not match IDEAGENT's ability to produce multiple qualifying ideas per topic.

On the four-topic subset, Sequential Memory improves threefold at NB ≥ 6 when scaled from 10 to 30 generations, but shows no further gain at 45 generations. At NB ≥ 7, performance is non-monotonic, increasing at 30 generations before declining at 45. Although this smaller experiment is exploratory, its per-topic Yield distribution likewise indicates that additional sampling alone does not reproduce IDEAGENT's performance. Together, these experiments suggest that IDEAGENT's gains cannot be explained solely by a larger generation budget.

D Evaluation Rubrics and Algorithms

Table 8: The eight diversity axes. The core axes are marked with †.

Axis	**What it asks**
Failure Mode†	What problem is solved?
Causal Diagnosis†	Why does the problem occur?
Intervention†	What is the central mechanism?
Signal Source	What measurable signal is used?
Intervention Locus	Where in the system does it act?
Objective	What tradeoff is optimised?
Evaluation Regime	Where would it be tested?
Assumption Class	What hidden assumption must hold?

Table 9: Three non-obviousness axes.

Axis	**What it asks**
Problem–Mechanism Pairing	Is the mechanism chosen the obvious/default fix for the stated problem, or a non-default one?
Causal Diagnosis	Is the underlying cause of the failure being addressed a standard, expected one, or a novel diagnosis?
Intervention	Is the point in the pipeline/system where the fix is applied a conventional one, or an unusual one reviewers rarely consider?

Table 10: Non-obviousness rubric. The accepted region used in our study is shown in green.

NB	Meaning
0	*Obvious*: a standard fix; the first thing a reviewer would propose.
1	Near-trivial variant of a known fix; no new reasoning.
2	*Straightforward extension*: familiar components combined in the expected way.
3	Predictable combination of two known ideas.
4	*Mildly non-obvious*: an interesting framing, but the core move is still the default.
5	*Moderately non-obvious*: a real failure mode paired with a non-default mechanism, still reachable by careful reasoning.
6	*Solidly non-obvious*: a meaningful failure mode and a mechanism that is not the obvious one.
7	*Notably non-obvious*: a surprising problem–mechanism pairing or an unusual signal source.
8	*Highly non-obvious*: a novel causal diagnosis, or a mechanism acting where reviewers rarely look.
9	*Exceptional*: a reframing of the problem itself.

E Background Corpus Collection

Table 11: Soundness rubric. The *severe* cases are shown in red, and the accepted region is in green.

S	Meaning
0	*Fatally flawed*: the mechanism contradicts itself or relies on something unavailable in its own setting.
1	*Mostly unsound*: the core causal claim doesn't follow from its own premises without unstated extra leaps.
2	*Significant gap*: the mechanism is plausible in isolation but rests on an assumption very likely false or untested.
3	*Shaky*: reasoning holds in the common case but ignores an obvious failure mode of its own mechanism.
4	*Mostly sound with a caveat*: the core logic holds, but at least one non-trivial edge case or confound is unaddressed.
5	*Sound with minor gaps*: the mechanism is logically coherent; a few secondary assumptions are untested but plausible.
6	*Solid*: the causal chain from mechanism to claimed effect is clear and internally consistent, with reasonable, explicitly stated assumptions.
7	*Rigorous*: the reasoning anticipates and addresses its own likely failure modes or confounds.
8	*Highly rigorous*: plausible alternative explanations for the claimed effect are explicitly considered and ruled out by the mechanism's own design.
9	*Airtight*: the mechanism follows from well-justified premises with no exploitable logical or technical gap; a skeptical reviewer finds no wedge to attack the core reasoning.

Table 12: Mechanism clarity rubric. The accepted region is shown in green.

C	Meaning
0	*Empty / hand-wavy*: no actual mechanism given, just a goal restated as a method.
1	*Vague direction*: names a technique family without saying what changes or how.
2	*Underspecified*: identifies a component to modify but not what signal, rule, or update it uses.
3	*Sketch-level*: the mechanism's general shape is stated but key steps are asserted rather than defined.
4	*Partially specified*: most of the mechanism is concrete; one central step remains hand-wavy or unresolved.
5	*Adequately specified*: the mechanism's main steps are all named and connected, though exact operationalization (thresholds, formulas) is left open.
6	*Well specified*: each step is concrete enough that an implementer would only need to choose minor free parameters.
7	*Precisely specified*: inputs, the transformation applied, and outputs are all stated; no structural ambiguity remains.
8	*Fully operational*: could be implemented directly from the description; only routine engineering choices are left.
9	*Unambiguous and complete*: every step, signal, and decision point is explicit enough that two independent implementers would build functionally the same thing.

Table 13: Feasibility rubric.

ϕ	Meaning
0	*Practically impossible*: requires resources, data, or capabilities that don't exist and aren't obtainable.
1	*Extremely demanding*: would require large, dedicated infrastructure or data collection beyond a typical project's scope.
2	*Very demanding*: feasible only with substantial new tooling, data collection, or compute beyond standard academic resources.
3	*Demanding*: needs meaningful new infrastructure (e.g. a new training pipeline or annotated dataset) not implied by the idea itself.
4	*Moderately demanding*: buildable with standard resources but requires nontrivial engineering effort or a moderately large compute budget.
5	*Feasible with effort*: implementable with existing tools/data with a reasonable, if nontrivial, engineering effort.
6	*Fairly feasible*: mostly reuses existing infrastructure/data; one or two components need custom work.
7	*Feasible*: could be built and tested by a small team using standard, available tools and data with modest effort.
8	*Highly feasible*: implementable quickly by adapting existing pipelines with minor modification.
9	*Immediately actionable*: testable essentially off-the-shelf, with data/tools/compute already at hand.

Table 14: Significance rubric.

ζ	Meaning
0	*Negligible*: even complete success would not meaningfully change anything a practitioner or researcher cares about.
1	*Marginal*: a small, local improvement noticeable only in narrow or synthetic settings.
2	*Minor*: a modest improvement on a narrow problem with little downstream consequence.
3	*Limited*: addresses a real but narrow problem; success would matter only to a small niche.
4	*Moderate*: meaningfully improves a recognized problem, though the problem itself is not central to the field.
5	*Notable*: success would meaningfully move a problem that a nontrivial part of the field actively cares about.
6	*Significant*: success would change how a meaningful subset of practitioners approach a known, important problem.
7	*Important*: success would resolve or substantially advance a problem widely recognized as a major open issue.
8	*Major*: success would shift how the field thinks about the underlying failure mode itself, beyond just improving numbers.
9	*Field-defining*: success would open or redefine a research direction, changing what subsequent work in the area targets.

Table 15: Holistic pairwise distinctness rubric (D_{ij}). The accepted region is shown in green.

D	Meaning
0	*Semantic duplicate*: same research claim and substantive mechanism.
1	*Cosmetic rewrite*: only wording, presentation, or naming differs.
2	*Minor variant*: same core mechanism and causal story with a small operational change.
3	*Meaningful variant*: meaningful modification, but still the same research direction.
4	*Sibling direction*: one core component changes; recognizable common direction remains.
5	*Partially distinct*: substantial change, but mechanism-family/core overlap remains.
6	*Distinct direction*: separately actionable direction with a genuinely different central mechanism or causal route; contextual differences alone never qualify.
7	*Clearly distinct*: mechanism, diagnosis, and evidence path are clearly separated.
8	*Highly distinct*: only the broad problem area or domain substantially connects the ideas.
9	*Orthogonal*: essentially all substantive components differ; reserve this for rare cases.

Algorithm 1 Exact Novel-Sound-Clear Count: $Y_{k,l,m,\tau}$

Require: Idea set $\mathcal{I}$; non-obviousness scores $\{\mathrm{NB}_i\}$; soundness scores $\{S_i\}$; pairwise distinctness scores $\{D_{ij}\}$; thresholds: k (non-obviousness), l (soundness), m (clarity), τ (distinctness)
Ensure: $\mathcal{K}^\star$, a *maximum-cardinality mutually distinct* eligible set

$\mathcal{N} \leftarrow \{\, i \in \mathcal{I} : \mathrm{NB}_i \geq k \ \wedge \ S_i \geq l \wedge C_i \geq m \}$
$E \leftarrow \{\, (i,j) \in \mathcal{N} \times \mathcal{N} : D_{ij} \geq \tau \,\}$
$\mathcal{K}^\star \leftarrow \emptyset$
function EXPAND($\mathcal{K}, \mathcal{C}$)
    if $|\mathcal{K}| > |\mathcal{K}^\star|$ **then**
        $\mathcal{K}^\star \leftarrow \mathcal{K}$
    end if
    if $|\mathcal{K}| + |\mathcal{C}| \leq |\mathcal{K}^\star|$ **then**
        return
    end if
    for $v \in \mathcal{C}$, in order **do**
        if $|\mathcal{K}| + |\mathcal{C}| \leq |\mathcal{K}^\star|$ **then**
            break
        end if
        $\mathcal{C}' \leftarrow \{\, u \in \mathcal{C} : u \neq v, \ (v,u) \in E \,\}$
        EXPAND($\mathcal{K} \cup \{v\}, \ \mathcal{C}'$)
        $\mathcal{C} \leftarrow \mathcal{C} \setminus \{v\}$
    end for
end function
EXPAND($\emptyset, \ \mathcal{N}$)
return $\mathcal{K}^\star$

Algorithm 2 Greedy Max-Min Topic Diversity Selection

Require: Topic names $\{t_1, \ldots, t_N\}$, sentence encoder $\mathcal{F}_\theta$, budget Γ
Ensure: Selected set $\mathcal{S}$, $|\mathcal{S}| = \Gamma$

$\hat{e}_i \leftarrow \mathcal{F}_\theta(t_i) \ \forall i$ ▷ encode topic names (normalized vectors)
$S \leftarrow \hat{E}\hat{E}^\top$ where $\hat{E} = [\hat{e}_1, \ldots, \hat{e}_N]^\top$ ▷ cosine similarity matrix
$s^* \leftarrow \arg\min_i \frac{1}{N-1} \sum_{j \neq i} S_{ij}$ ▷ most isolated topic as seed
$\mathcal{S} \leftarrow \{s^*\}, \quad \mathcal{R} \leftarrow \{1, \ldots, N\} \setminus \{s^*\}$ ▷ initialise selected and remaining sets
while $|\mathcal{S}| < \Gamma$ **do**
    $i^* \leftarrow \arg\min_{i \in \mathcal{R}} \max_{j \in \mathcal{S}} S_{ij}$ ▷ pick topic least similar to any selected
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{i^*\}, \quad \mathcal{R} \leftarrow \mathcal{R} \setminus \{i^*\}$ ▷ update selected and remaining sets
end while
return $\mathcal{S}$

Table 16: List of domains and topics included in our dataset.

Domain	# BKDG	Topic Name
cs.AI	8	Scaling LLM Reasoning via Reinforcement Learning
	8	Scaling LLM Reasoning via Reinforcement Learning
	5	Reasoning in Large Language Models
	6	Reinforcement Learning for Advanced AI Reasoning
cs.CL	8	Deep Learning Representation and Analysis
	8	Scaling Laws for Neural Networks
	7	Knowledge Localization and Model Editing
	8	Efficient Attention Mechanisms and KV Cache Optimization
	6	Multilingual Multitask Language Understanding Evaluation
	6	Adversarial Robustness of LLM Safety Alignment
cs.CV	8	Efficient Long-Range Dependency Modeling
	8	Embodied AI and Autonomous Systems
	7	Controllable Text-to-Image Synthesis and Editing
	8	Adaptive and Personalized Generative Modeling
	6	Deep Learning for Image Anomaly Detection
	6	Multi-camera Bird's-Eye-View 3D Object Detection
cs.DC	7	Efficient Large Language Model Serving Systems
	7	Scalable LLM Training and Inference Systems
cs.IR	7	Autoregressive Generative Modeling Across Modalities
cs.LG	8	Large-Scale Model Parameter Optimization
	7	Offline Reinforcement Learning and Generative Modeling
	7	Geometry and Linearity of Weight Space
	6	Unifying Diffusion and Flow-Based Generative Models
	6	Efficient LLM Compression and Adaptation
	5	Foundation Models for Time Series Forecasting
cs.RO	8	Diffusion Models for End-to-End Autonomous Driving
	8	Robot Manipulation Policy Learning
	7	End-to-End Autonomous Driving
	7	Imitation Learning for Robotic Manipulation
	7	Visuomotor Learning for Robot Manipulation
	5	Diffusion-based Generative Modeling
cs.SE	7	RL and Search for LLM Reasoning