NANYANG TECHNOLOGICAL UNIVERSITY
DECLARE LAB | RESEARCH PAPER | UNDER REVIEW

# IDEAgent: Agentic Quality-Diversity Search for Research Idea Generation

**Varun Gumma**[1], **Navonil Majumder**[1], **Soumitra Sinhahajari**[1], **Soujanya Poria**[1]

[1]DeCLaRe Lab, Nanyang Technological University, Singapore



**ABSTRACT**

Large Language Models (LLMs) have significantly automated the process of scientific discovery over the past few years. However, existing systems share one core limitation: they generate and optimize ideas independently for either *Quality* or *Diversity*. This often leads to the generation of ideas in close proximity to one another or to a large set of trivial, unsound, or unclear concepts. In this work, we instead argue that research ideation should be treated as a conjunction of both objectives and framed as a *Quality-Diversity* (QD) search. In line with this perspective, we introduce IDEAgent, a *multi-agent* framework that manages the evolution of ideas through *lineages*. We jointly drive *Quality* using multi-objective feedback for dedicated repair and refinement, while *Diversity* is achieved through *lightweight* sequential memory and explicit comparison against completed ideas, their historical ancestors, and rejected proposals. To systematically evaluate this QD conjunction, we develop Yield, a joint metric that computes the *largest* set of *mutually diverse* ideas that satisfy a predetermined quality threshold. Finally, through evaluations across 32 topics spanning 8 domains of Computer Science, we show that IDEAgent outperforms the best baseline by 3.89× on Yield, while achieving non-zero Yield on 8× more topics. We further corroborate these findings through an analysis of quality improvements, showing that repair and refinement are crucial for building logical rigor and clarity while preserving non-obviousness. To encourage future research on QD-search-based ideation, we open-source IDEAgent at https://github.com/declare-lab/IDEAgent.

**CODE** GitHub

**CORRESPONDENCE** Soujanya Poria (soujanya.poria@ntu.edu.sg)

**DATE** July 27, 2026

## 1 Introduction

The advent of Large Language Models (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* are formulated along with *potential solutions* (the ideas) (Sinhahajari et al., 2026). In this work, we propose to view the ideation process as a *Quality-Diversity (QD) Search* problem. The objective is to discover a *set* of research ideas, where each proposal is *non-obvious*, *sound* and *clear* (**Quality axis**) while remaining mutually orthogonal and non-overlapping with the rest (**Diversity axis**), allowing researchers to explore multiple distinct directions of inquiry rather than repeatedly refining the same or a narrow set of ideas.

Despite recent work, LLM-based ideation systems fail to view this task as a rigorous QD search, but instead develop ideas independently. Current ideation systems treat each generation, either with one-shot prompting or multi-agent systems, as an isolated optimization problem for quality or diversity tasks independently, and also evaluate accordingly. Consequently, as the system attempts to expand to further ideas, 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 individually *plausible* ideas, they offer rapidly diminishing returns (Tang and Yang, 2026) and are often riddled with inconsistencies while pushing for novelty. We argue that in a real-world scientific workflow, the value of an AI-ideation system is defined by its ability to not only generate a single proposal, but a diverse set of *independent*, *sound*, *clear* and *non-obvious* research directions that offer richer alternatives to pursue.

Sustained human research naturally operates as a *Quality-Diversity* search. In each individual brainstorming session, researchers iterate over a concept to build a non-obvious or *novel* mechanism—correcting logical flaws, tightening assumptions, and polishing its overall *Quality*. Such a natural *evolution* of an idea can be represented as a *lineage*, where new higher-quality versions replace initial parent drafts. Across sessions, maintaining diversity relies on a structured record of past explorations where new ideas usually are cross-referenced across three distinct sets: *finalized and completed* high-quality concepts, *dead ends and failed hypotheses* that were permanently discarded, and the any older/historical ideas of the other *lineages* that were conceptually sound but eventually superseded by more refined variants. By continuously filtering against these, subsequent thinking is forced into new and unexplored directions, maximizing *Diversity*.

Motivated by this perspective, we introduce IDEAgent, a multi-agent framework to navigate the QD search for research ideation. Our framework first sequentially develops new ideas against compact memories of prior generations for conceptual diversification. Later, every idea is later rigorously evaluated for quality by multiple specialized agents, while being actively compared against completed ideas, their historical variants, and any rejected directions for novelty and rediscoveries, each with their dedicated archives. Ideas which pass the quality assessment are eligible to acceptance with optional refinement, while those that miss by a limited margin are put through repairs targeting exact flaws and inconsistencies. Finally, all these idea comparisons in our system are efficiently enabled via structured summaries for a more interpretable and point-wise semantic contrast.

Validating the QD search framework naturally requires an evaluation logic that accounts for the joint objective. Independently, both *Quality* and *Diversity* metrics can be gamed and inflated. 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 and 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 exactly measures our proposed requirement – the number of *viable* ideas a human researcher could pursue from the generated set.

Finally, 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*. Through a comprehensive analysis, 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 highlight the gains brought by our improvement pipeline across individual quality rubrics according to both internal and external evaluators. Although the absolute scores differ between the two judges, they consistently agree on the direction and persistence of the improvements, particularly in soundness/rigor, clarity, and the preservation of non-obviousness.

Our contributions are as follows:

1. We formulate automated scientific ideation as a Quality-Diversity (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 and open-source IDEAgent, a multi-agent framework that maintains idea *lineages* throughout the search process using dedicated multi-dimensional internal evaluations for quality

assessment. We further introduce separate archives for ideas at different stages of their life cycles to enable efficient novelty comparison using compact core summaries instead of full textual ideas. Finally, we incorporate *opportunity-based* quality improvement subroutines that build upon these multi-agent evaluations to provide targeted feedback in the form of *repairs* and *refinements*.

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 as a *potential solution* to 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 simple drafting assistance to fully autonomous, end-to-end research pipelines (Lu et al., 2024; Gottweis et al., 2026; Schmidgall et al., 2025). To improve the quality of generated hypotheses, 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 thoughts and prevent superficial suggestions, subsequent architectures incorporated retrieval loops designed to ground the model's reasoning within existing literature (Baek et al., 2025; Li et al., 2025; Zhao et al., 2025). More recently, this paradigm has expanded from abstract ideation into active software execution to iteratively implement and optimize 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 failure in divergent thinking. LLM agents frequently exhibit conceptual narrowness, clustering their outputs tightly around the provided seed literature and recombining familiar technical variations rather than introducing genuinely new research paths (Tang and 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. To this end, empirical evaluations of autonomous agents demonstrate that their success on complex downstream engineering tasks is fundamentally bounded by the diversity of their early-stage ideation (Audran-Reiss et al., 2025).

**Quality-Diversity Search in Text Generation** In the context of LLM-based generations, the most closely related work is `QDAIF` (Bradley et al., 2024). `QDAIF` applies an evolutionary algorithm, utilizing language models to generate mutations, evaluating candidates for elitism, and operating on a fixed archive grid to map the search space. However, we maintain a more amorphous setup for evolution without any fixed grid search space or MAP-Elites, and improve the quality of ideas with the help of directed feedback.

## 3 Methodology

### 3.1 Task Formulation as Quality–Diversity Search

IDEAgent formulates scientific ideation as a *Quality–Diversity* (QD) 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 a set of user-provided research papers targeting an area of interest, we define *idea* $I$ as a structured research proposal specifying a problem, causal diagnosis, intervention, underlying mechanism, assumptions, expected effect, and evaluation strategy (see Tables 4 and 5).

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

A raw idea generated in one-shot may contain a promising mechanism, but could be inconsistent, obvious, or under specified. We thus repair and refine the seed idea and track its evolution as a lineage $\ell_I$. Close variants of the same mechanism inherit the same lineage and cannot be counted as separate discoveries. A seed and its lineage are thus a fresh attempt to discover a mechanism and consume exactly one unit of the discovery budget $B$.

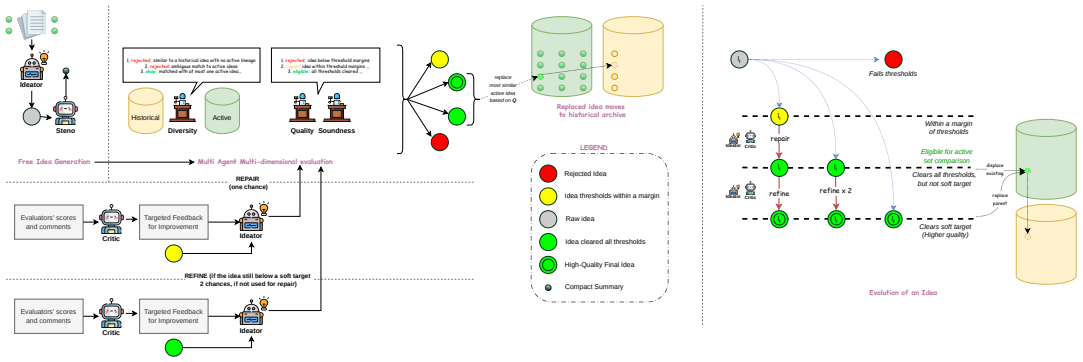



**Figure 1: Left:** An abstract overview of IDEAgent highlighting the crucial components and flow. **Right:** An abstract overview of the evolution of an idea and a lineage, where a raw idea might either be directly rejected, repaired and refined, just refined, or directly accepted as per the assessment by the evaluators. Note that a refinement/repair maintains the lineage of the idea ($\ell_I$). We purposefully omit the exact conditionals at each step for simplicity and refer the readers to the methodology (§3) for it.

## 3.2 Agents and Search Memory

The **Ideator** is the core generator agent that reads a background corpus $\mathcal{X}$ to generate a complete seed idea from scratch or subsequently refine it based on targeted feedback. Since the generated proposal is loosely formed, a **Stenographer** *summarizes* each draft idea $I$ into a concise quintuple:

$$\sigma(I) = (p_I, m_I, v_I, a_I, e_I), \tag{1}$$

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 structured decomposition provides a more manageable and interpretable representation for pairwise comparison of two ideas.

Three evaluators assess each draft idea. (1) The **Quality Evaluator** scores non-obviousness ($\text{NB}_I$), 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.

(2) The **Soundness Panel** produces $M = 5^{\dagger}$ independently sampled judgments $S_I^1, \ldots, S_I^M \in \{0, \ldots, 9\}$ on logical and mathematical consistency of the idea's arguments, and assumptions. These multiple judgments are aimed at mitigating any rollout-specific biases or oversights (Ho et al., 2026). Their mean used a single representative score and rescaled to 0-100 as:

$$S_I = \frac{100}{9M} \sum_{j=1}^{M} S_I^j. \tag{2}$$

A judgment is deemed *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 higher make it *invalid*. The *Invalid* ideas are discarded, whereas the *disputed* ideas are not directly accepted but could be repaired.

(3) The **Diversity Judge** compares core signatures of the active, previously-qualified-but-inactive, and rejected ideas, to the new generation. Based on the similarity with respect all previously qualified ideas[1], a Diversity score $D_I \in [0, 100]$ is issued along with a duplicate list which identifies the closest mechanisms from the active and historical sets for further comparison.

Finally, a **Critic** turns the evaluator scores and failure evidence into one focused feedback for repair or refinement with a goal to improve the current idea rather than propose a replacement direction.

[†]empirically, higher values beyond 5 had diminishing returns, but greatly increased the overall runtime and cost
[1]active and inactive

The controller maintains four *stores* of compact memories. The active archive ($\mathcal{A}$) contains currently accepted ideas. A repair queue ($\mathcal{Q}$) temporarily holds a promising near-miss during its immediate repair. A historical archive ($\mathcal{H}$) retains signatures of previously qualified ideas that left $\mathcal{A}$, including replaced parents and ideas displaced by other higher quality ones. Instead of storing exact discarded ideas in a rejected archive ($\mathcal{R}$), multiple individual failures are aggregated into *failed/rejected patterns*[2] for explicit avoidance during generation. Subsequent seed prompts receive these compact representations rather than all previous full texts for computation efficiency. Full parent text is provided to the Ideator only when that lineage is being revised via repair or refinement.

## 3.3 Search Procedure

For $b = 1, \ldots, B$, the search completes the bounded lineage of one seed before generating the next:

$$I_b^{(0)} \longrightarrow [\textsc{Repair}] \longrightarrow [\textsc{Refine}] \longrightarrow I_{b+1}^{(0)}, \tag{3}$$

where the bracketed operations are conditional.

**Stage I: Generate and assess.** Every fresh-seed prompt asks the Ideator for a problem or mechanism distinct from all ideas represented in $\mathcal{A}$, $\mathcal{H}$, $\mathcal{Q}$, and $\mathcal{R}$. But, if the preceding lineage had ended as a pure historical duplicate or falls in an rejected pattern/logic, the instruction is augmented with an explicit causal pivot to abandon that failing mechanism or problem area in its next generation. Next, once a draft is produced, the Stenographer and evaluators return its structured quintuple summary, quality scores, soundness panel, and archive-relative diversity judgment, respectively.

Based on the evaluator assessments, we define the core qualification gate as:

$$G(I) = \mathbb{1}\left[(N_I \geq \tau_N \wedge S_I \geq \tau_S \wedge C_I \geq \tau_C) \wedge \neg\,\text{invalid}(I) \wedge \neg\,\text{disputed}(I)\right] \tag{4}$$

The primary search uses $\tau_N = \tau_S = \tau_C = 60$ and a diversity floor $\tau_D = 60$. Here, Feasibility is measured but excluded from the gate as an ambitious, sound mechanism may be valuable before it is immediately executable.

**Stage II: Route and repair.** The Diversity Judge first determines the candidate's relationship to $\mathcal{A}$ and $\mathcal{H}$, after which the controller applies four cases:

1. First, an idea is said to be a *purely historical match* if it only matches historical mechanisms, but not the active refined successors of their lineage. This means the candidate is sub-optimal and rejected. A candidate matching multiple active ideas is also rejected because it cannot be assigned unambiguously to one lineage to comparison.

2. A candidate matching exactly one active idea is treated as another version of that lineage, not as a new discovery. If it passes $G$, it may replace the matched idea when it has a higher combined quality score (Equation (5)); otherwise it is rejected.

3. A candidate no duplicate matches but $D_I < \tau_D$ indicates a combination or reuse of multiple pre-qualified ideas, but not exactly matching one. Such an idea is discarded as it lacks diversity. A repair is not effective in this case to create diversity as it is instructed to preserve the central mechanism.

4. Lastly, a candidate with $D_I \geq \tau_D$ and no duplicate matches is admitted to the active set if it passes $G$. If it instead falls within $\delta = 20$ points of the non-obviousness, soundness, and clarity thresholds and is not *invalid* in its soundness panel evaluation, it has the *opportunity* 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 list, quality evidence, soundness objections, and list of competing/similar mechanisms. The Ideator then returns a complete corrected child with the same lineage. The child is evaluated again and must pass the same qualification and diversity requirements. If the single repair attempt fails, the original seed-level representative is retained and its failure is summarized in $\mathcal{R}$.

[2] As humans, we rarely remember exact failures as they can be numerous and quite diverse, but a higher abstraction/aggregation of failure techniques, causes and patterns for a more broader and generalizable match.

Candidates that qualify compete in $\mathcal{A}$ according to a linear combination of the individual quality scores:

$$Q_b(I) = 0.7N_I + 0.2S_I + 0.1C_I \tag{5}$$

This combination emphasizes innate non-triviality of the idea[3], which is difficult to achieve via simple 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 when its $Q_b$ is higher, and any ties are broken by Feasibility ($F_I$). Whenever an accepted/qualified idea leaves[4] $\mathcal{A}$, its summary is saved in $\mathcal{H}$, not $\mathcal{R}$, as it is a valid and good-quality piece.

**Stage III: Refine accepted ideas.** An accepted idea has an *opportunity* for refinement while it remains below a soft target: $N_I < 90$, $S_I < 80$, or $C_I < 80$. Note that, these targets serve only as a conditional check for refinement and are not additional acceptance gates. The Critic addresses the relevant deficit, and the Ideator produces a same-lineage child. During the child's diversity judgment, its own lineage is naturally excluded while all foreign lineages remain visible to avoid trivial matches.

The child replaces its parent only if it passes $G$, clears $\tau_D$ 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, and position disagreement is treated as a tie. This safeguard prevents improvements in clarity or soundness from trivializing the central move. 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 non-gating consideration throughout this process, and is only used for a tie break during a $Q_b$ match. The Ideator is asked to provide a plausible implementation and evaluation path, but feasibility does not trigger repair or refinement and is not 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 total number of Ideator calls lies between $B$ and $B \cdot (1 + K_{\text{aux}})$, or between 10 and 30 in the primary setting. The search ends after all $B$ fresh seeds and the final seed's pending auxiliary chain are complete. The resulting $\mathcal{A}$ now contains the final set of ideas generated by IDEAgent which are also 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 domains, including `cs.AI`, `cs.CL`, and `cs.RO` to build our background corpus $\mathcal{X}$. In a way, our goal is to generate a set of diverse, non-trivial, sound research ideas that could *potentially* emerge from a shared set of background papers, which serves as a seed knowledge bank. To achieve this, we start from a collection of published papers within each domain and identify the broader set of papers that *might* have influenced them.

For each domain, we first collect papers published within a predefined time frame and filter them based on publication status, and ranking[5]. For every remaining paper, we independently extract its *influential citations* (ICs) (Sinhahajari et al., 2026) and group them into sets based on the overlap of their ICs. Specifically, two papers are assigned to the same set if the Intersection-over-Union (IoU) of their ICs exceeds 0.5. For each resulting set, we define the union of all ICs as its *background papers*.

Next, we retain only those sets of *background papers* that contribute to at least three published papers. This filtering step ensures that each knowledge bank has demonstrated the ability to support multiple novel and high-quality research ideas. Such a collection of background papers as a *topic*, which describes a uniquely identifiable sub-domain or task. Topics that contain between 5 and 8 background papers are retained, and we use an LLM to assign each a descriptive tag. Finally, we then set a budget of $\Gamma = 6$

[3]Note that we did not rigorously optimize any hyperparameters in this study due to budget constraints. The results reported in this paper could be under-reported or unoptimized.
[4]displaced by another higher quality idea, replaced by a new variant of its own lineage
[5]top-$k$%.

and apply greedy maximum diversity selection (Algorithm 1) to obtain the final collection of topics as presented in Table 3.

## 5 Evaluation

In line with our formulation of the ideation process as a *Quality-Diversity* search, we evaluate each method on the returned *set* of $N$ ideas per topic rather than assessing ideas in isolation. A valid set must satisfy independent yet complementary requirements: ideas must individually be high-quality (non-obvious, sound, and clear) while remaining mutually diverse. Near-duplicate high-quality ideas collapse into a single redundant contribution, whereas highly diverse but trivial, invalid, or unclear ideas offer no real scientific value. To capture these, our evaluation protocol leverages an LLM-judge to supply independent per-idea *Quality* scores, and pairwise *Diversity* metrics, and then combines them according to the QD motivation.

**Quality Axis** We evaluate *Quality* of an idea $I$ along two primary† and three secondary axes, respectively:

1. **Non-obviousness**† ($\mathrm{NB}_I$) measures the originality of an idea's core contribution. The judge evaluates the idea according to *Problem-Mechanism pairing*, *Causal Diagnosis*, and *Intervention* (Table 5), which boil down to checking whether the idea would *surprise* a knowledgeable human researcher while ignoring differences in writing quality or presentation. The exact scale value and textual descriptions are presented in Table 6.

2. **Soundness**† ($S_I$) measures whether an idea's mechanism would actually work as claimed, and if the logical flow has internal contradictions or is based on non-rigorous or untested assumptions. A safe, unoriginal idea can still score low here if its own reasoning has a gap, and a highly original idea can score high here if its logic holds up cleanly. The exact scale value and textual descriptions are presented in Table 7.

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 having to invent missing pieces. This is judged separately from whether the mechanism is correct (soundness) or practical to build (feasibility). The exact scale value and textual descriptions are presented in Table 8.

4. **Feasibility** measures how practical it would be to actually build and test the idea, given realistic constraints such as data, compute, existing tools, and time. It does not factor in whether the idea is correct or important, as an idea can be highly feasible yet unsound, or highly sound yet infeasible. The exact scale value and textual descriptions are presented in Table 9.

5. **Significance** measures how much it would matter if the idea worked exactly as claimed, both for the specific problem and for the broader field. The judge assumes the mechanism works as claimed and ignores its current feasibility or stage of development. The exact scale value and textual descriptions are presented in Table 10.

In all our future discussions for quality, we mainly deal with non-obviousness and soundness together as an idea that reads as highly non-trivial but is actually inconsistent is meaningless and unusable.

**Diversity Axis** In a set of $N$ ideas, *Diversity* is evaluated pairwise. Each idea is first independently decomposed into a semantic signature along eight axes: *failure mode, causal diagnosis, intervention, signal source, intervention locus, objective, evaluation regime,* and *assumption class* (Table 4). For each of the $\binom{N}{2}$ idea pairs in the generated set, a single judgment call compares the pair along all eight axes at once, producing a same/variant/related/distinct relation per axis. A second call then takes these eight axis-level relations as fixed evidence, and generates a single score $D_{ij} \in \{0, \ldots, 9\}$ (Table 11). The classification is instructed to weight disagreement on the *core* axes – *failure mode, causal diagnosis,* and *intervention* – more heavily than the remaining supporting axes. Per topic $T$, we aggregate and report the average pairwise diversity as

$$D(T) = \binom{N}{2}^{-1} \sum_{i} \sum_{i<j} D_{ij} \tag{6}$$

**Yield** Our primary QD performance indicator is Yield(NB $\geq k, S \geq l, C \geq m, D \geq \tau$), which counts the *maximum* number of *sound, clear, non-trivial*, and *mutually diverse* research ideas present in a set of generated ideas. To compute it, we first retain the high-quality ideas that surpass all the pre-defined thresholds of *non-obviousness* ($k$), *soundness* ($l$), and *clarity* ($m$). To correct for diversity, we extract the *largest* subset (maximum clique) of remaining ideas in which every idea pair has distinctness $D_{ij} \geq \tau$ (Algorithm 2). Under this constraint, a set of $N$ highly novel and sound, yet near-identical ideas yields a score of 1 rather than $N$, penalizing methods that exploit quality scores by repeatedly paraphrasing a small number of successful concepts.

**Successful Topics** We define a topic $T$ as successful at yielding $\Phi$ for the method under test if it generates at least $\Phi$ ideas that satisfy our Yield gates. Thus, we report the proportion of successful topics as

$$\mathbb{E}_{T \in \text{Topics}}\, \mathbb{1}[\text{Yield}_T(\text{NB} \geq k, S \geq l, C \geq m, D \geq \tau) \geq \Phi]$$

A general idea generator should ideally have this ratio at 1.

# 6 Experimental Setup

## 6.1 Agents

Our initial experiments with smaller open-source models revealed a stark gap in the capabilities of the agents. We empirically found that current open-source models lacked logical consistency in their generations, often producing ideas with *severely* low soundness scores. Furthermore, they frequently misjudged ideas during internal evaluation, resulting in unreliable scores and subsequent corrections.

To this end, all our experiments use proprietary large-scale models, with the latest `gpt-5.6-sol`[6] (OpenAI, 2026) serving as the core Ideator. The Critic, Soundness Panel, and Quality Judge are built around `gemini-3.1-pro`[7] (Google DeepMind, 2026a), while the Diversity Judge and Steno use `gemini-3.5-flash` (Google DeepMind, 2026b) and `gemini-2.5-flash` (Comanici et al., 2025), respectively. During evaluation, we employ two external large-scale models, `claude-sonnet-5`[7] (Anthropic, 2026b) and `claude-opus-4.7`[7] (Anthropic, 2026a), for all judgments. During analysis, we average the scores from both judges and treat the result as a single *jury verdict* to account for potential biases and sampling stochasticity. Finally, we set the same idea budget of $B = 10$ for all the baselines discussed below.

## 6.2 Baselines

**Stateless** The most naive generation strategy, in which the Ideator produces $B$ ideas in parallel without any correlation or knowledge of one another.

**One-Shot ($\mathcal{OS}$)** In this setup, the Ideator generates all $B$ ideas at once in a single output. Note that, in the case of a thinking-enabled Ideator, the model can use its thinking space as a shared memory for developing and diversifying all the ideas. Furthermore, every generated idea naturally conditions on all preceding ideas through causal generation. Hence, this setup is analogous to a rudimentary form of *stateful* generation.

**Sequential-Memory ($\mathcal{SM}$)** Building on the previous notion of statefulness, we move to a many-shot setup in which each of the $B$ ideas is generated independently, while we *cumulatively* carry forward the core signatures $\sigma(i)$ of all previous ideas so that the model retains a lightweight memory of prior generations for diversification. This setup is sequentially similar to IDEAgent, but does not include the additional quality improvement subroutines (repair and refinement).

**NOVA-inspired $\mathcal{SM}$** Building further on Sequential-Memory, we augment it with NOVA's (Hu et al., 2025) iterative *seed-pool*-based germination and replacement strategy for diversification. Note that NOVA also includes an online retrieval module, which we discard to better suit our generation setting, where the background papers constitute a closed knowledge base. Specifically, we run the Sequential-Memory system for $n = 3$ rounds, with each round generating $B$ ideas. At the end of every round, an external judge reflects on the generated ideas and selects the $m = 3$ most promising ones as seed concepts, replacing those from the previous round. The subsequent round then germinates $B$ new ideas from these seed

[6] `thinking_effort=low`
[7] `thinking_effort=high`

**Table 1:** Results across 32 topics as evaluated by two external judges. Scores are averaged over the two evaluators. Bold marks the best method in each row; blue shading identifies the primary outcomes and quality dimensions.

| Metric | Sequential Memory | Single-shot | Stateless | NOVA | IDEAgent |
|---|---|---|---|---|---|
| QUALITY–DIVERSITY OUTCOMES gate: $D \geq 7, S \geq 7, C \geq 6$ | | | | | |
| Yield(NB $\geq 7$, gate) | 0.188 | $0.000^{\dagger}_{(-100.0\%)}$ | $0.250_{(+33.3\%)}$ | $0.281_{(+50.0\%)}$ | $\mathbf{1.094}^{\S}_{(+483.3\%)}$ |
| Yield(NB $\geq 6$, gate) | 0.531 | $0.000^{\ddagger}_{(-100.0\%)}$ | $0.812^{\dagger}_{(+52.9\%)}$ | $1.156^{\dagger}_{(+117.6\%)}$ | $\mathbf{2.312}^{\S}_{(+335.3\%)}$ |
| SUCCESSFUL TOPICS gate: $D \geq 7, S \geq 7, C \geq 6$ | | | | | |
| Yield(NB $\geq 7$, gate) $\geq 1$ | 6/32 | 0/32 | 7/32 | 8/32 | **27/32** |
| Yield(NB $\geq 7$, gate) $\geq 2$ | 0/32 | 0/32 | 1/32 | 1/32 | **8/32** |
| Yield(NB $\geq 6$, gate) $\geq 1$ | 14/32 | 0/32 | 24/32 | 25/32 | **31/32** |
| Yield(NB $\geq 6$, gate) $\geq 2$ | 3/32 | 0/32 | 2/32 | 11/32 | **23/32** |
| Yield(NB $\geq 6$, gate) $\geq 3$ | 0/32 | 0/32 | 0/32 | 1/32 | **15/32** |
| PRIMARY QUALITY DIMENSIONS | | | | | |
| Non-obviousness | 6.020 | $5.609^{\S}_{(-6.8\%)}$ | $6.077_{(+0.9\%)}$ | $6.142_{(+2.0\%)}$ | $\mathbf{6.430}^{\S}_{(+6.8\%)}$ |
| Soundness | 6.534 | $5.477^{\S}_{(-16.2\%)}$ | $6.589_{(+0.8\%)}$ | $6.483_{(-0.8\%)}$ | $\mathbf{6.720}^{\dagger}_{(+2.8\%)}$ |
| DIVERSITY AND SECONDARY QUALITY MEASURES | | | | | |
| Mechanism clarity | 5.416 | $4.030^{\S}_{(-25.6\%)}$ | $5.739^{\dagger}_{(+6.0\%)}$ | $5.556_{(+2.6\%)}$ | $\mathbf{6.341}^{\S}_{(+17.1\%)}$ |
| Pairwise diversity | 6.606 | $6.739^{\dagger}_{(+2.0\%)}$ | $5.080^{\ddagger}_{(-23.1\%)}$ | $\mathbf{6.757}^{\dagger}_{(+2.3\%)}$ | $6.641_{(+0.5\%)}$ |
| Feasibility | 4.053 | $4.034_{(-0.5\%)}$ | $4.006_{(-1.2\%)}$ | $\mathbf{4.150}_{(+2.4\%)}$ | $3.864_{(-4.7\%)}$ |
| Significance | 5.683 | $5.316^{\S}_{(-6.5\%)}$ | $\mathbf{5.898}^{\dagger}_{(+3.8\%)}$ | $5.470^{\dagger}_{(-3.7\%)}$ | $5.572_{(-2.0\%)}$ |

*Note.* Subscripts on score entries report change relative to Sequential Memory. $^{\dagger}p < .05$; $^{\ddagger}p < .001$; $^{\S}p < .0001$ (paired topic-level tests; rubric tests Holm-corrected).

directions together with the accumulated global memory[8]. Finally, the critic selects the best $B$ ideas from the complete pool of $n \cdot B$ generations, emphasizing distinctness, to form the final evaluation set. As before, each idea is generated in a single attempt without any subsequent repair or refinement.

## 7 Results and Analysis

In this section, we discuss some interesting analyses, curious questions, and findings based on the experiments conducted with IDEAgent and other baselines. All the scores and numbers we present 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 those produced by 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 2 presents a zoomed-in view of the high-quality region ($NB \geq 6$, $S \geq 6$), while the complete $10 \times 10$ grid is shown in Figure 4.

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 approximately 50%, at ($C \geq 6, D_{ij} \geq 7$). As expected, the surface declines as we move toward the bottom-right with increasingly stringent quality constraints, eventually approaching zero for *both* methods when $NB \geq 8$ and $S \geq 8$. The primary operational 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

[8] Prior idea core signatures.

gap remains substantial at +0.9, demonstrating its ability to generate non-obvious, sound, and mutually diverse research ideas.

**IDEAgent vs. Baselines** IDEAgent also outperforms the secondary baseline, NOVA, as shown in Table 1, with per-judge results reported in Tables 12 and 13. The improvement in Yield is a clear 2× and 3.89× at $NB \geq 6$ and $NB \geq 7$, respectively. Furthermore, the three primary quality metrics—Non-Obviousness, Soundness, and Clarity—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. Selecting the final $B$ ideas from a pool of $3B$ candidates naturally provides greater opportunity for variation. 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.

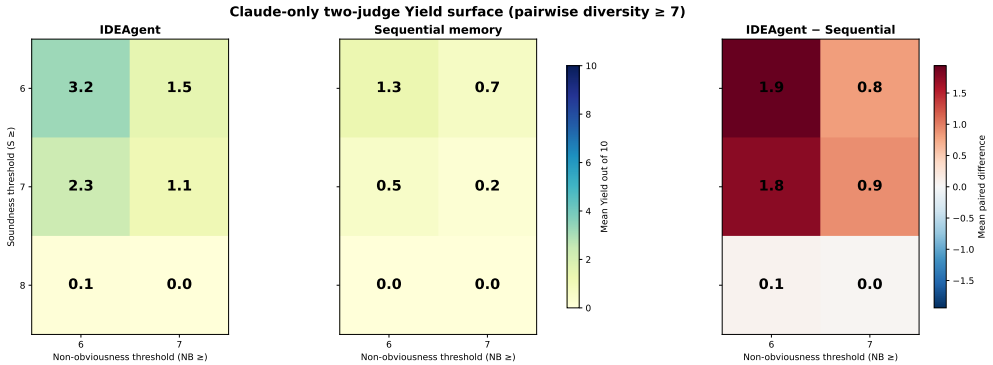



**Figure 2:** Two Judges' average Yield for IDEAgent, Sequential-Memory baseline, and their difference.

For the proportion of successful topics, we see in Table 1 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, under $\mathrm{Yield}(D \geq 7, S \geq 7, C \geq 6, \mathrm{NB} \geq 6) \geq 2$, 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.

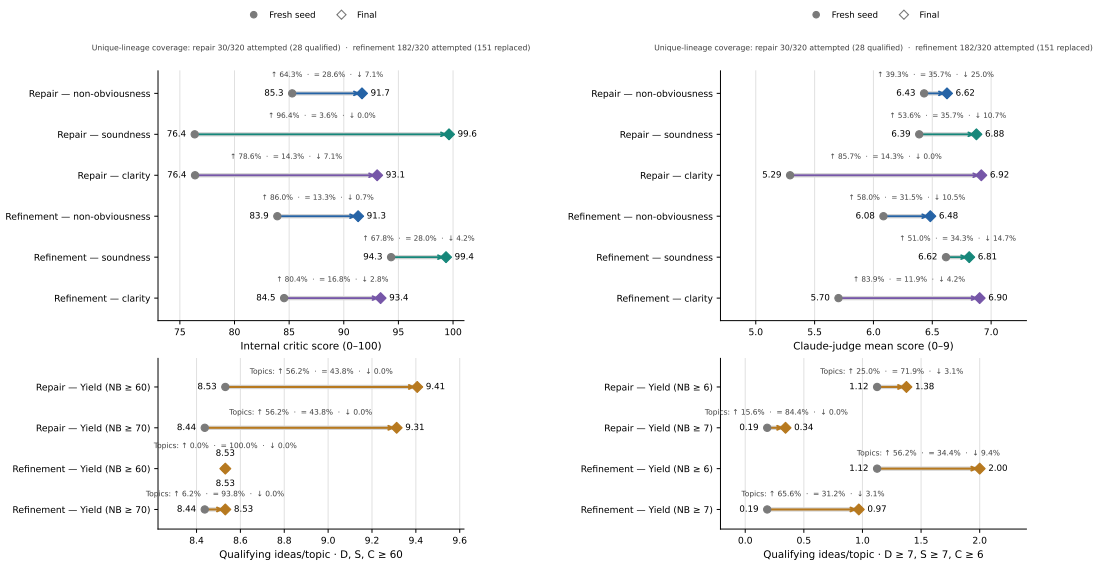
**Quality improvements due to repair and refinement** The primary distinction between all the baselines and IDEAgent is the inclusion of quality improvement subroutines in the form of repair and refinement, which we analyze in this section across all 320[9] lineages. Figure 3a illustrates the quality improvements as assessed by the internal judge agents during generation. Specifically, we compare the quality of the initial seed idea with its final version after any repair or refinement has been applied. Out of the 320 lineages, only 30 required repair, as they fell just below the qualification thresholds, and among these, 28 were successfully improved and ultimately *qualified*. Similarly, 182 of the 320 lineages underwent refinement

[9] 32 topics × 10 ideas per topic

after qualifying, and in a substantial 82% of these cases, the refined version replaced the original idea, demonstrating the overall effectiveness of the process.

Taking a closer look, we further analyze improvements across the individual quality rubrics together with the resulting Yield. Since refinement is applied only to ideas that have already qualified the thresholds, whereas repair is reserved for ideas that narrowly miss them, the 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 a maximum gain of approximately 0.97 at both non-obviousness thresholds, whereas the gains from refinement are comparatively small.

It is also important to verify that the improvements observed by the internal agents during generation are reflected in downstream evaluation. Figure 3b 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, primarily 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 relative effectiveness of repair and refinement is reversed according to the external judges. Refinement contributes the largest gains in Yield, improving it by +0.88 and +0.78 at the NB thresholds of 6 an 7, respectively. Finally, we refrain from directly correlating the score deltas across the internal and external evaluations because the two systems employ different rubrics and scoring scales. Instead, we emphasize their qualitative agreement regarding the consistent persistence of the observed improvements.




**(a)** *wrt* Internal judges **(b)** *wrt* External judges

**Figure 3:** Quality improvements due to repair and refinement techniques

**Inter Annotator Agreement** To establish the reliability of our evaluation framework, we analyze the inter-annotator agreement between the two judges. To compute the agreement per rubric as shown in

**Table 2:** Agreement between Claude Opus 4.7 and Claude Sonnet 5 across all methods and topics. Linear-weighted Cohen's $\kappa$ measures agreement in absolute ordinal scores; Spearman's $\rho$ measures agreement in within-topic rankings.

| Rubric | Linear-weighted $\kappa$ | Median Spearman $\rho$ |
|---|---|---|
| Non-obviousness | 0.455 | 0.610 |
| Soundness | 0.268 | 0.409 |
| Mechanism clarity | **0.604** | 0.600 |
| Feasibility | 0.210 | **0.761** |
| Significance | 0.352 | 0.504 |
| Pairwise diversity | 0.424 | 0.679 |

Table 2, we use the *linear-weighted* Cohen's Kappa on the pooled evaluations of all topics across all runs[10]. Across the primary rubrics, soundness shows the least agreement of 0.26, demonstrating that models find judging soundness in terms of logical and mathematical consistency of the mechanism and assumptions subjective (Ho et al., 2026). Clarity has the highest agreement of 0.60, indicating that a clear idea is easily identified by the model. The rest of the core rubrics, Non-obviousness and diversity, hold a moderate agreement of $\sim 0.43$, providing a decent indicator about the internal alignment of the jury. Taken together, agreement is consistently higher for rubrics with a concrete, checkable referent in the idea's text (clarity, non-obviousness, diversity) than for rubrics requiring a more holistic or forward-looking judgment (feasibility, significance, and to a lesser extent, soundness), suggesting that judge disagreement tracks each rubric's inherent ambiguity rather than a general unreliability of the evaluation framework.

The Spearman column in Table 2 reports whether the two evaluators rank the same ideas relatively higher or lower (Spearman, 1904). Rank correlations are positive across all rubrics, ranging from 0.409 for soundness to 0.761 for feasibility. Following the descriptive convention of 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.

## 8 Conclusion

In this work, we presented IDEAgent, a multi-agent framework that reframes scientific ideation as a *Quality-Diversity* (QD) search to build *portfolios of ideas* with maximum quality density. Through sequential generation and multi-agent evaluation centered on the core requirements of QD search, we introduced a simple quality improvement subroutine that repairs or optionally refines ideas based on their identified *opportunities*. We further demonstrated lineage-inspired management of ideas throughout this evolutionary process using dedicated archives for completed, rejected, and historical ideas, efficiently enabled by core signature comparison. To jointly evaluate both quality and diversity, we proposed Yield, which filters ideas using fixed quality thresholds across its primary indicators and then selects the largest subset whose ideas satisfy a pairwise diversity threshold. Finally, we evaluated IDEAgent across 32 Computer Science research topics and presented an extensive analysis demonstrating the contributions of both the quality improvement loop and stateful memory. We hope that our problem formulation and the open-sourced experimental framework provide a new perspective on AI-driven scientific ideation.

## 9 Limitations

Our work has several limitations that warrant an open discussion.

1. All internal and external evaluations rely primarily on LLM-based judges to assess non-obviousness, soundness, clarity, feasibility, and diversity. To mitigate evaluator-specific bias, we employ two independent large-scale judge models and aggregate their scores with equal weightage. Similarly, to calibrate the crucial soundness evaluation, we sample 5 independent scores and aggregate them for downstream usage to mitigate any accidental biases. Moreover, our primary metric, Yield, is model-agnostic and can be paired with alternative evaluation protocols, including human judgments.

[10] baselines and IDEAgent

2. We restrict the scope of this work to generating portfolios of diverse and 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 a thorough implementation, debugging, and end-to-end experimental and devoted expert validation. Likewise, we **do not** claim that the generated ideas are necessarily publishable or absent from the existing literature. Verifying such claims would require a comprehensive retrieval over both published and emerging research to identify potentially related prior work. However, we do show that our ideas demonstrate significant soundness and logical consistency across multiple rigorous evaluations.

3. Due to budget constraints, each idea is only given one repair opportunity and at most two refinements, if one was not consumed for repair. Both repair and refinement are computationally heavy subroutines as they require the involvement of all Agents (Quality, Diversity, Soundness, Ideator, and Critic) in our system. Due to using proprietary LLMs for our setup, additional rounds would significantly increase the monetary load of the setup. Nevertheless, our aim in this work was to demonstrate the concept of follow-up repair and refinement techniques to improve ideas along the Quality axis.

4. All our experiments use proprietary LLMs used via their APIs. This was because we noticed that **current** smaller open-source models severely lacked knowledge, idea formulation/refinement, and evaluation capacity. However, our framework is agnostic to any model type or family, and hence we expect the underlying principles of sequential generation, quality improvements via 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 Preview, Claude Sonnet 5, and Claude Opus 4.8. These models generally produced non-obvious and diverse ideas with clearly articulated mechanisms; however, the evaluators *frequently* identified serious soundness flaws in their underlying mechanisms, substantially reducing overall idea 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 nine-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.

## 10 Acknowledgements

We extend our sincere gratitude to Google DeepMind for their *Gemini Academic Program Award* and GCP credits which enabled us to run the Gemini And Claude models at scale for our project.

## References

Anthropic. Introducing Claude Opus 4.7. https://www.anthropic.com/news/claude-opus-4-7, April 2026a. Accessed 2026-07-16.

Anthropic. Introducing Claude Sonnet 5. https://www.anthropic.com/news/claude-sonnet-5, June 2026b. Accessed 2026-07-16.

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, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6709–6738, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. . 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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

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

Google DeepMind. Gemini 3.1 Pro - model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, February 2026a. Accessed 2026-07-16.

Google DeepMind. Gemini 3.5 Flash - model card. https://deepmind.google/models/model-cards/gemini-3-5-flash/, May 2026b. Accessed 2026-07-16.

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. . 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, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21330–21359, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. . 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, page 211–218, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450305570. . 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*, pages 331–366. Springer Nature Singapore, Singapore, 2024. ISBN 978-981-99-3814-8. . 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, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8971–9004, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. . 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.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, July 2026. Accessed 2026-07-16.

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. . 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, page 348–374, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400724152. . 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, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5977–6043, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. . 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. .

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.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. .

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, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28201–28240, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. . 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, editors, *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 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 List of Domains & Topics

## B Evaluation: Rubrics and Algorithms

## C Additional Results

## D Cost and Futher Scaling

**Table 3:** 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 |

## D.1 Cost Metrics

**Cost / Successful Topic** One natural question in effective idea generation is whether one can simply sample more ideas using cheap methods over fewer ideas with more advanced but expensive approaches. To address this question, we report Cost per Successful Topic (CST), defined as the total generation cost across all attempted topics divided by the number of successful topics:

For method $M$, we define

$$\mathrm{CST}_M(k,l,m,\tau,\Phi) = \frac{\sum_{T\in\mathrm{Topics}} C_{M,T}}{\sum_{T\in\mathrm{Topics}} \mathbb{1}\left[\mathrm{Yield}_{M,T}(\mathrm{NB}\geq k,\ S\geq l,\ C\geq m,\ D\geq\tau)\geq\Phi\right]}, \tag{7}$$

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

The primary objective of quality–diversity search is to produce diverse high-quality ideas for each topic. The pairwise diversity requirement is non-vacuous only when at least two qualifying ideas are produced; consequently, $\Phi = 2$ is our principal diverse-portfolio criterion.

**Algorithm 1** Greedy Max-Min Topic Diversity Selection

**Require:** Topic names $\{t_1, \ldots, t_N\}$, sentence encoder $\mathcal{F}_\theta$, budget $\Gamma$
**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 4:** The eight diversity axes; three are key.

| Axis | What it asks | Key |
|---|---|---|
| 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? | |

## D.2 Analysis

**CST** For Cost per Successful Topic (CST), 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× cheaper per successful topic and succeeds on 2.1× as many topics at NB ≥ 6 and 8× as many topics at NB ≥ 7. At the greater depth of three qualifying ideas, IDEAgent is approximately 11× 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 subset of four topics to examine its scaling trend. From each generation pool, up to 10 ideas were retained by Gemini-3.1-pro-preview using the same selection criteria described above.

As shown in Figure 5, scaling Stateless from 10 to 50 generations improves mean Yield by approximately 27% at the NB ≥ 6 gate and 62% at the NB ≥ 7 gate. Nevertheless, IDEAgent achieves approximately 2.06× and 2.46× the Yield of the scaled Stateless baseline at these respective gates. The per-topic Yield distributions show the same pattern: additional fresh generations improve Stateless, but do not match IDEAgent's ability to produce multiple qualifying ideas for the same topic.

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

**Table 5:** 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 6:** Non-obviousness rubric for a single idea ($NB_i$). The accepted region used in our study in green.

| $\sigma$ | 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. |

**Table 7:** Soundness rubric ($S_i$). The *severe* cases are marked in red, and the accepted region used in our study in green.

| $\rho$ | 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 8:** Mechanism clarity rubric ($C_i$). The accepted region used in our study in green.

| $\kappa$ | 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 9:** Feasibility rubric.

| $\phi$ | 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 10:** Significance rubric.

| $\zeta$ | 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 11:** Holistic pairwise distinctness rubric ($D_{ij}$). The accepted region used in our study in green.

| $\delta$ | 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. |

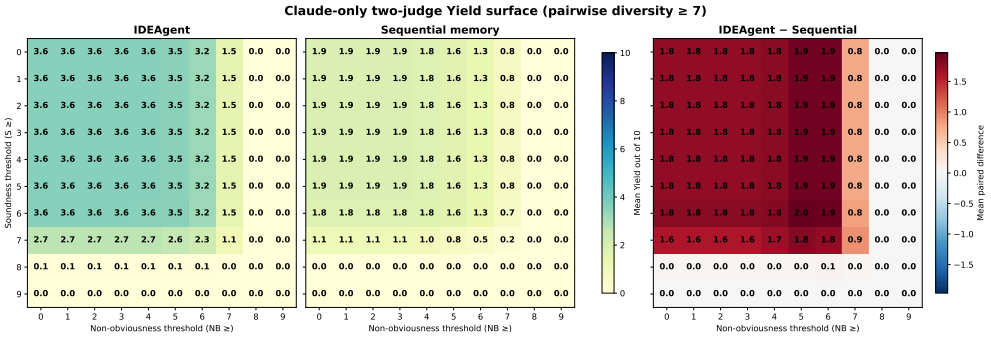
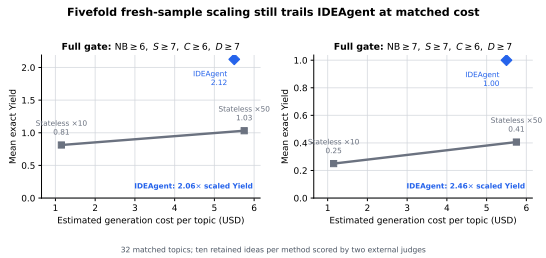



**Figure 4:** Full Yield surfaces for IDEAgent, Sequential-Memory baseline, and their difference.

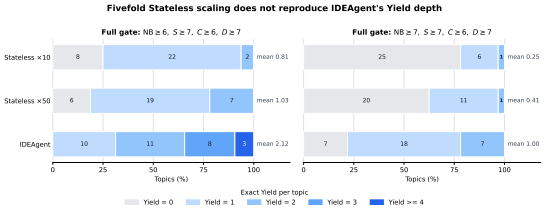
**Algorithm 2** Exact Novel-Sound-Clear Direction Count Yield($NB \geq k,\ S \geq l,\ C \geq m, D \geq \tau$)

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

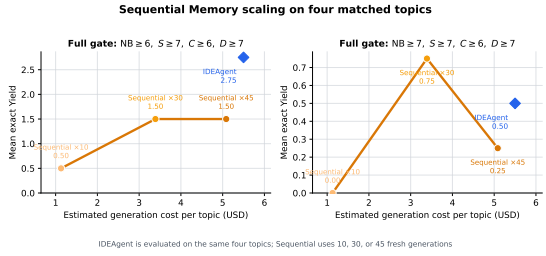
$\mathcal{N} \leftarrow \{\, i \in \mathcal{I} : NB_i \geq k \ \wedge\ S_i \geq l \ \wedge C_i \geq m \}$ ▷ joint eligibility gate
$E \leftarrow \{\, (i,j) \in \mathcal{N} \times \mathcal{N} : D_{ij} \geq \tau \,\}$ ▷ adjacency list on $\mathcal{N}$
$\mathcal{K}^\star \leftarrow \emptyset$
**function**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



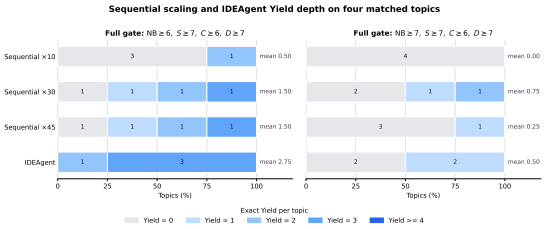

**(a)** Cost-Yield scaling: Stateless on 32 topics.




**(b)** Yield-topic distribution: Stateless on 32 topics.




**(c)** Cost-Yield scaling: Sequential Memory on four topics.




**(d)** Yield-topic distribution: Sequential Memory on four topics.

**Figure 5:** Scaling fresh generation does not reproduce IDEAgent's quality–diversity performance.

**Table 12:** Results across 32 topics as evaluated by Claude Opus 4.7. Bold marks the best method in each row; blue shading identifies the primary outcomes and quality dimensions.

| Metric | Sequential Memory | Single-shot | Stateless | NOVA | IDEAgent |
|---|---|---|---|---|---|
| QUALITY–DIVERSITY OUTCOMES gate: $D \geq 7, S \geq 7, C \geq 6$ | | | | | |
| Yield(NB $\geq 7$, gate) | 1.312 | $0.031^{\S}_{(-97.6\%)}$ | $1.312_{(+0.0\%)}$ | $1.562_{(+19.0\%)}$ | $\mathbf{3.531}^{\S}_{(+169.0\%)}$ |
| Yield(NB $\geq 6$, gate) | 2.375 | $0.031^{\S}_{(-98.7\%)}$ | $1.719^{\dagger}_{(-27.6\%)}$ | $2.844_{(+19.7\%)}$ | $\mathbf{4.656}^{\S}_{(+96.1\%)}$ |
| SUCCESSFUL TOPICS gate: $D \geq 7, S \geq 7, C \geq 6$ | | | | | |
| Yield(NB $\geq 7$, gate) $\geq 1$ | 22/32 | 1/32 | 28/32 | 23/32 | **32/32** |
| Yield(NB $\geq 7$, gate) $\geq 2$ | 14/32 | 0/32 | 12/32 | 17/32 | **31/32** |
| Yield(NB $\geq 7$, gate) $\geq 3$ | 6/32 | 0/32 | 2/32 | 7/32 | **26/32** |
| Yield(NB $\geq 6$, gate) $\geq 1$ | 27/32 | 1/32 | 31/32 | 31/32 | **32/32** |
| Yield(NB $\geq 6$, gate) $\geq 2$ | 21/32 | 0/32 | 19/32 | 27/32 | **31/32** |
| Yield(NB $\geq 6$, gate) $\geq 3$ | 17/32 | 0/32 | 5/32 | 18/32 | **30/32** |
| COST / SUCCESSFUL TOPIC gate: $D \geq 7, S \geq 7, C \geq 6$ | | | | | |
| Yield(NB $\geq 7$, gate) $\geq 1$ | \$1.64 | \$6.72 | \$**1.31** | \$5.45 | \$5.50 |
| Yield(NB $\geq 7$, gate) $\geq 2$ | \$**2.58** | *NA* | \$3.07 | \$7.38 | \$5.68 |
| Yield(NB $\geq 7$, gate) $\geq 3$ | \$**6.03** | *NA* | \$18.40 | \$17.92 | \$6.77 |
| Yield(NB $\geq 6$, gate) $\geq 1$ | \$1.34 | \$6.72 | \$**1.19** | \$4.05 | \$5.50 |
| Yield(NB $\geq 6$, gate) $\geq 2$ | \$**1.72** | *NA* | \$1.94 | \$4.65 | \$5.68 |
| Yield(NB $\geq 6$, gate) $\geq 3$ | \$**2.13** | *NA* | \$7.36 | \$6.97 | \$5.87 |
| PRIMARY QUALITY DIMENSIONS | | | | | |
| Non-obviousness | 6.175 | $5.609^{\S}_{(-9.2\%)}$ | $6.206_{(+0.5\%)}$ | $6.331_{(+2.5\%)}$ | $\mathbf{6.612}^{\S}_{(+7.1\%)}$ |
| Soundness | 6.862 | $5.794^{\S}_{(-15.6\%)}$ | $6.891_{(+0.4\%)}$ | $6.834_{(-0.4\%)}$ | $\mathbf{6.978}_{(+1.7\%)}$ |
| DIVERSITY AND SECONDARY QUALITY MEASURES | | | | | |
| Mechanism clarity | 5.447 | $3.878^{\S}_{(-28.8\%)}$ | $5.728_{(+5.2\%)}$ | $5.597_{(+2.8\%)}$ | $\mathbf{6.353}^{\S}_{(+16.6\%)}$ |
| Pairwise diversity | 7.062 | $7.190^{\dagger}_{(+1.8\%)}$ | $5.309^{\S}_{(-24.8\%)}$ | $\mathbf{7.268}^{\dagger}_{(+2.9\%)}$ | $7.047_{(-0.2\%)}$ |
| Feasibility | 4.862 | $4.772_{(-1.9\%)}$ | $4.847_{(-0.3\%)}$ | $\mathbf{5.006}_{(+3.0\%)}$ | $4.637_{(-4.6\%)}$ |
| Significance | 5.975 | $5.534^{\S}_{(-7.4\%)}$ | $\mathbf{6.137}_{(+2.7\%)}$ | $5.775^{\dagger}_{(-3.3\%)}$ | $5.822^{\dagger}_{(-2.6\%)}$ |

*Note.* Subscripts on score entries report change relative to Sequential Memory. $^{\dagger}p < .05$; $^{\ddagger}p < .001$; $^{\S}p < .0001$ (paired topic-level tests; rubric tests Holm-corrected).

**Table 13:** Results across 32 topics as evaluated by Claude Sonnet 5. Bold marks the best method in each row; blue shading identifies the primary outcomes and quality dimensions.

| Metric | Sequential Memory | Single-shot | Stateless | NOVA | IDEAgent |
|---|---|---|---|---|---|
| QUALITY–DIVERSITY OUTCOMES gate: $D \geq 7, S \geq 7, C \geq 6$ | | | | | |
| Yield(NB $\geq 7$, gate) | 0.406 | $0.000^{\ddagger}_{(-100.0\%)}$ | $0.375_{(-7.7\%)}$ | $0.500_{(+23.1\%)}$ | $\mathbf{1.219}^{\S}_{(+200.0\%)}$ |
| Yield(NB $\geq 6$, gate) | 0.844 | $0.031^{\S}_{(-96.3\%)}$ | $1.000_{(+18.5\%)}$ | $1.250^{\dagger}_{(+48.1\%)}$ | $\mathbf{2.406}^{\S}_{(+185.2\%)}$ |
| SUCCESSFUL TOPICS gate: $D \geq 7, S \geq 7, C \geq 6$ | | | | | |
| Yield(NB $\geq 7$, gate) $\geq 1$ | 12/32 | 0/32 | 11/32 | 14/32 | **28/32** |
| Yield(NB $\geq 7$, gate) $\geq 2$ | 1/32 | 0/32 | 1/32 | 2/32 | **10/32** |
| Yield(NB $\geq 6$, gate) $\geq 1$ | 20/32 | 1/32 | 28/32 | 28/32 | **31/32** |
| Yield(NB $\geq 6$, gate) $\geq 2$ | 6/32 | 0/32 | 4/32 | 11/32 | **26/32** |
| Yield(NB $\geq 6$, gate) $\geq 3$ | 1/32 | 0/32 | 0/32 | 1/32 | **15/32** |
| COST / SUCCESSFUL TOPIC gate: $D \geq 7, S \geq 7, C \geq 6$ | | | | | |
| Yield(NB $\geq 7$, gate) $\geq 1$ | $\mathbf{3.01} | *NA* | $3.35 | $8.96 | $6.29 |
| Yield(NB $\geq 7$, gate) $\geq 2$ | $36.16 | *NA* | $36.80 | $62.72 | $\mathbf{17.60} |
| Yield(NB $\geq 6$, gate) $\geq 1$ | $1.81 | $6.72 | $\mathbf{1.31} | $4.48 | $5.68 |
| Yield(NB $\geq 6$, gate) $\geq 2$ | $\mathbf{6.03} | *NA* | $9.20 | $11.40 | $6.77 |
| Yield(NB $\geq 6$, gate) $\geq 3$ | $36.16 | *NA* | *NA* | $125.44 | $\mathbf{11.73} |
| PRIMARY QUALITY DIMENSIONS | | | | | |
| Non-obviousness | 5.866 | $5.609^{\dagger}_{(-4.4\%)}$ | $5.947_{(+1.4\%)}$ | $5.953_{(+1.5\%)}$ | $\mathbf{6.247}^{\S}_{(+6.5\%)}$ |
| Soundness | 6.206 | $5.159^{\S}_{(-16.9\%)}$ | $6.287_{(+1.3\%)}$ | $6.131_{(-1.2\%)}$ | $\mathbf{6.463}^{\ddagger}_{(+4.1\%)}$ |
| DIVERSITY AND SECONDARY QUALITY MEASURES | | | | | |
| Mechanism clarity | 5.384 | $4.181^{\S}_{(-22.3\%)}$ | $5.750^{\dagger}_{(+6.8\%)}$ | $5.516_{(+2.4\%)}$ | $\mathbf{6.328}^{\S}_{(+17.5\%)}$ |
| Pairwise diversity | 6.150 | $\mathbf{6.288}_{(+2.2\%)}$ | $4.850^{\S}_{(-21.1\%)}$ | $6.246_{(+1.6\%)}$ | $6.234_{(+1.4\%)}$ |
| Feasibility | 3.244 | $\mathbf{3.297}_{(+1.6\%)}$ | $3.166_{(-2.4\%)}$ | $3.294_{(+1.5\%)}$ | $3.091_{(-4.7\%)}$ |
| Significance | 5.391 | $5.097^{\ddagger}_{(-5.4\%)}$ | $\mathbf{5.659}^{\dagger}_{(+5.0\%)}$ | $5.166^{\dagger}_{(-4.2\%)}$ | $5.322_{(-1.3\%)}$ |

*Note.* Subscripts on score entries report change relative to Sequential Memory. $^{\dagger}p < .05$; $^{\ddagger}p < .001$; $^{\S}p < .0001$ (paired topic-level tests; rubric tests Holm-corrected).

**Table 14:** 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 | Sequential Memory | Single-shot | Stateless | NOVA | IDEAgent |
|---|---|---|---|---|---|
| COST / SUCCESSFUL TOPIC gate: $D \geq 7, S \geq 7, C \geq 6$ | | | | | |
| Yield(NB $\geq 7$, gate) $\geq 1$ | $6.03 | *NA* | $\mathbf{5.07} | $16.32 | $6.52 |
| Yield(NB $\geq 7$, gate) $\geq 2$ | *NA* | *NA* | $35.52 | $130.56 | $\mathbf{22.00} |
| Yield(NB $\geq 6$, gate) $\geq 1$ | $2.58 | *NA* | $\mathbf{1.48} | $5.22 | $5.68 |
| Yield(NB $\geq 6$, gate) $\geq 2$ | $12.05 | *NA* | $17.76 | $11.87 | $\mathbf{7.65} |
| Yield(NB $\geq 6$, gate) $\geq 3$ | *NA* | *NA* | *NA* | $130.56 | $\mathbf{11.73} |