IdeaFactory: Autonomous Research Ideation with Adaptive Branching Chain Search

Anonymous submission

Abstract

Autonomous research agents can now perform long-horizon tasks including literature retrieval, code implementation, experiment execution, and paper writing, yet research ideation remains a lightweight front-end step. Recent studies show that, under the same prior-work context, LLM-generated ideas concentrate on a few bridge-and-synthesis patterns and diverge markedly from human research taste. We characterize this as idea exploration collapse and attribute it to one-shot retrieve-generate-rank pipelines, which eliminate heterogeneous branches before they can be adequately examined and repaired. We introduce Autonomous Idea Development (AID), a stateful process that develops literature-grounded opportunities into testable proposals, and IdeaFactory, an end-to-end, traceable multi-agent system for AID. Its core method, Adaptive Branching Chain Search (ABCS), formulates research ideation as quality-diversity search. Under a fixed budget, ABCS organizes research paths into cells defined by research-gap discovery patterns and research-strategy categories, expands underexplored paths, and repairs promising elites using critic feedback. This prevents global ranking from prematurely discarding promising branches and helps uncover high-quality proposals missed by flat generation. We evaluate IdeaFactory on proposal quality and overall idea distribution. Experiments show that it improves proposal quality without compromising feasibility and narrows the distributional gap from human research taste. Further analysis indicates that these gains arise primarily from mitigating idea exploration collapse through quality-diversity search.

1 Introduction

Autonomous research agents are increasingly capable of carrying out long research workflows, including literature retrieval, code implementation, experiment execution, result analysis, and paper writing (Lu et al. 2024; Yamada et al. 2025; Gottweis et al. 2026; Tang et al. 2025; Feng et al. 2026; Jin et al. 2026; Lyu et al. 2026). However, their autonomy remains largely concentrated on execution and optimization after a research direction has been specified. Idea generation is usually treated as a lightweight front-end module that reads a user-provided topic and a small set of related papers, quickly produces several candidate proposals, and passes them to the downstream experimental pipeline. As a result, current systems behave more like efficient research executors than agentic research teams capable of actively identifying promising research opportunities.

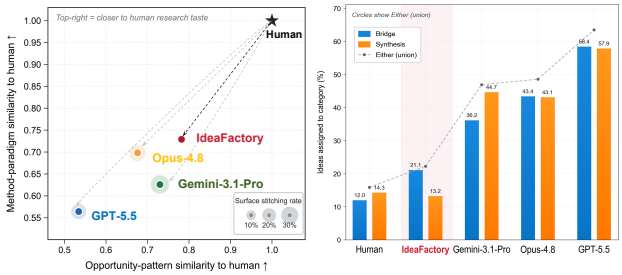

Figure 1: Alignment of IdeaFactory with human research taste. (a) Distributional alignment in opportunity patterns and method paradigms. (b) Bridge, Synthesis, and Either prevalence across systems.

As illustrated in Figure 1, this limitation concerns not only the quality of individual ideas but also a systematic **research-taste distribution gap**. A recent study shows that, even when receiving the same prior-work context, LLM-generated ideas remain concentrated within a narrower set of research patterns (Chen, Zhao, and Cohan 2026). Models are more likely to frame research opportunities as bridges between existing works, whereas human ideas cover a broader range of gap framings and research paradigms. Complementary evidence shows that LLM-generated ideas can be judged more novel while remaining slightly weaker in feasibility (Si, Yang, and Hashimoto 2025). ResearchStudio-Idea further demonstrates the value of evidence grounding, prior-art collision checking, and outcome-informed auditing when developing traceable research proposals (Zhao et al. 2026).

This concentration suggests that the dominant retrieve–generate–rank paradigm is prone to **idea exploration collapse**. When literature understanding, bottleneck identification, claim formulation, and experiment design are compressed into a single generation step, models tend to favor research paths that are safer and easier to justify. Even when repeated sampling produces different candidates, flat

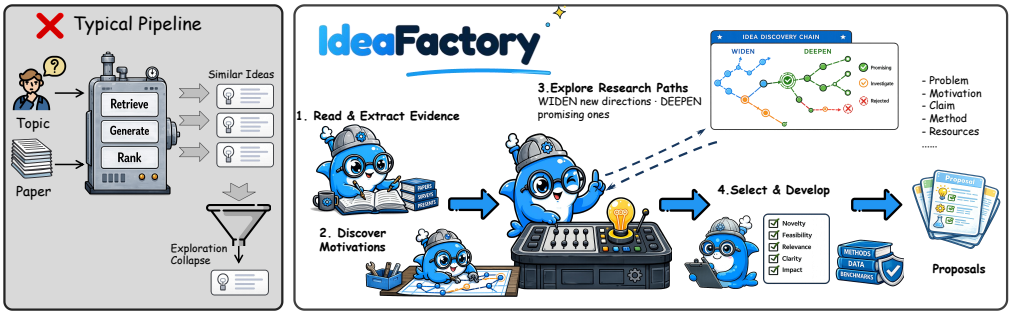

Figure 2: IdeaFactory at a glance. The system develops literature evidence into research motivations, explores and repairs multiple research paths with ABCS, and consolidates promising directions into testable proposals.

Top-K selection may eliminate heterogeneous branches before they have been adequately explored and repaired. Existing ideation systems have explored literature-grounded generation, iterative refinement, and multi-agent collaboration (Wang et al. 2024b; Baek et al. 2025; Su et al. 2025); preserving heterogeneous candidate paths through final selection remains a distinct challenge. We argue that scientific ideation should not be treated as one-shot text generation. Instead, it should be formulated as a stateful process that continuously tracks evidence, compares alternative branches, incorporates criticism, and repairs promising research directions. We define this process as **Autonomous Idea Development (AID)**, which aims to progressively develop an initial research topic and its supporting literature into testable research proposals.

We propose **IdeaFactory**, a multi-agent system for AID. IdeaFactory coordinates agents with complementary research perspectives to jointly understand the literature, identify research opportunities, formulate and challenge research claims, and review candidate proposals.

IdeaFactory organizes AID into three connected stages. During *Evidence-Grounded Discovery*, the system constructs a literature landscape and derives research motivations from concrete evidence and unresolved gaps. During *Adaptive Research Path Exploration*, **Adaptive Branching Chain Search (ABCS)** organizes these paths according to how research opportunities are identified and addressed. Under a fixed budget, ABCS balances exploration of underexplored directions with repair of promising candidates. Following quality-diversity principles (Mouret and Clune 2015), the system preserves strong candidates across distinct research paths rather than forcing premature global competition. During *Testable Proposal Development*, quality-diversity selection and proposal auditing further develop the selected claims into testable research proposals. Claim-level critiques guide subsequent branches, while proposal-level evaluation and auditing are linked back to their originating paths as delayed feedback for subsequent search within the same run.

Our main contributions are summarized as follows:

- **We introduce IdeaFactory**, an end-to-end and traceable multi-agent system for Autonomous Idea Development that transforms literature evidence into testable research proposals.
- **We propose Adaptive Branching Chain Search (ABCS)**, which uses branching exploration and critic-guided repair to balance proposal quality and path diversity under a finite search budget.
- **We demonstrate that IdeaFactory generates higher-quality research proposals**, explores a broader range of research paths, and more closely matches human research patterns.

2 Related Work

2.1 Autonomous Research Agents

Recent autonomous research agents increasingly cover the full research workflow, from idea generation and literature review to experiment execution and paper writing. End-to-end systems such as AI Scientist, AI Scientist-v2, AI-Researcher, and InternAgent-1.5 connect ideation with implementation and manuscript generation over long horizons (Lu et al. 2024; Yamada et al. 2025; Tang et al. 2025; Feng et al. 2026). Multi-agent systems instead introduce complementary roles for proposing, criticizing, and verifying research directions (Gottweis et al. 2026; Su et al. 2025; Yang, Li, and Li 2026). EvoScientist accumulates idea-generation and experimental experience across tasks through persistent memory, whereas Arbor uses hypothesis trees to support long-horizon experimentation and artifact refinement (Lyu et al. 2026; Jin et al. 2026). In contrast to systems that primarily target complete research workflows or downstream experimental optimization, IdeaFactory focuses on research opportunity discovery and proposal development before experimental commitment.

2.2 Scientific Idea Generation and Search

Prior work generates scientific ideas through direct prompting, literature-grounded retrieval, iterative refinement, and multi-agent collaboration (Si, Yang, and Hashimoto 2025; Wang et al. 2024b; Baek et al. 2025; Su et al. 2025). ResearchStudio-Idea combines literature retrieval, bottleneck identification, pattern-guided generation, and prior-work auditing into a traceable ideation workflow (Zhao et al. 2026). Idea2Story instead pre-constructs reusable methodological structures from scientific literature and retrieves them during online research generation (Xu et al. 2026). However, generating more candidates does not necessarily yield a broader set of final directions, because selection based on a single global score may still favor similar research paths.

IdeaFactory instead formulates ideation as stateful multi-path search. ABCS explores under-covered directions, repairs promising branches, and preserves strong candidates across distinct research paths.

2.3 Research Idea Evaluation

Existing studies evaluate individual ideas in terms of novelty, feasibility, relevance, clarity, and potential impact, or decompose discovery into inspiration retrieval, hypothesis composition, and hypothesis ranking (Si, Yang, and Hashimoto 2025; Liu et al. 2026). Human evaluation further shows that model-generated ideas may appear novel while remaining limited in feasibility and practical research value (Si, Yang, and Hashimoto 2025). TasteGap identifies a complementary issue at the distribution level, showing that model-generated ideas are overly concentrated in a small set of research patterns, such as bridging and synthesis, while other work studies scientific taste as a learnable preference signal (Chen, Zhao, and Cohan 2026; Tong et al. 2026). We therefore evaluate both the quality of individual proposals and the overall distribution of generated ideas, which allows us to examine whether the system improves proposal quality while exploring a broader range of research paths.

3 Method

3.1 Autonomous Idea Development

Existing research ideation systems typically formulate ideation as a short-horizon generation problem: given a research topic q and relevant literature $\mathcal{L}$, a model generates candidate proposals and selects final outputs through a common evaluation procedure. Even with repeated sampling or short refinement loops, candidates are largely generated from a fixed context and subsequently compete under global selection (Baek et al. 2025; Su et al. 2025; Zhao et al. 2026).

This formulation compresses literature understanding, opportunity identification, claim formulation, and proposal validation into a small number of generation decisions. Distinct research directions therefore compete before they are sufficiently developed: familiar or easy-to-justify patterns can gain an early advantage, while promising directions requiring further criticism and repair may be prematurely discarded, leading to the **idea exploration collapse** discussed in Section 1.

We formulate research ideation as **Autonomous Idea Development (AID)**. Rather than treating an idea as complete text produced in a single generation step, AID views ideation as a stateful process in which research directions are progressively developed through evidence accumulation, path exploration, and critical feedback. Given a research topic q, an accessible literature environment $\mathcal{L}$, and a finite budget B, the system maintains an evolving research state:

$$S_{t+1} = \mathcal{T}(S_t, a_t, o_t), \qquad t < B, \tag{1}$$

where a_t denotes a research action and o_t its resulting research object or feedback. The accumulated state is eventually developed into a set of testable research proposals $\mathcal{P} = \{p_1, \ldots, p_K\}$.

AID imposes three requirements. First, *evidence-grounded progression* keeps evolving claims traceable to concrete prior work. Second, *branching with preservation* keeps multiple plausible directions active long enough for exploration and repair before global comparison. Third, *controlled research reasoning* separates global exploration decisions, local research construction, and retention judgments. IdeaFactory is designed around these requirements, transforming one-shot candidate generation into a sustained process jointly driven by evidence, search, and critique.

3.2 System Overview

IdeaFactory is an end-to-end multi-agent system for AID. Given a research topic and an accessible literature environment, it develops an evidence-grounded understanding of the surrounding research space, explores directions with distinct scientific motivations and solution strategies, and progressively develops promising directions into testable proposals.

As illustrated in Figure 3, the process follows three stages: *Evidence-Grounded Discovery*, *Adaptive Research Path Exploration*, and *Testable Proposal Development*. The first stage constructs a compact research landscape from relevant literature and extracts evidence-grounded motivations through close reading and cross-paper analysis. The second treats these motivations as a research frontier, where ABCS allocates a finite budget between opening new research paths and repairing promising branches. Finally, mature directions are developed into candidate proposals and examined for quality, novelty, and practical feasibility.

The stages share evidence, decisions, and feedback through the **Idea Discovery Chain**, a run-scoped structured research memory represented at step t as

$$G_t = (V_t, E_t). \tag{2}$$

Nodes represent research objects such as evidence, motivations, search tasks, claims, candidates, and decision records, while edges encode provenance, support, revision, selection, and rejection relations. Unlike a natural-language interaction history, the Chain preserves scientific dependencies among research objects: claims retain their supporting evidence, revisions remain linked to the critiques that triggered them, and failed branches remain available as informative search history.

The Chain represents research progress rather than performing search itself. It provides shared structured state across research roles, while ABCS uses recorded paths and feedback to determine where subsequent search effort should be allocated. Final proposals can consequently be traced back to their supporting evidence and development history.

3.3 Adaptive Branching Chain Search

The central search problem in AID is deciding which research directions should be opened and which existing directions deserve further investment under a finite budget. Simply generating more candidates does not solve this problem: under global quality selection, mature or commonly generated directions gain an inherent advantage, while less-developed branches may disappear before criticism and repair reveal their value.

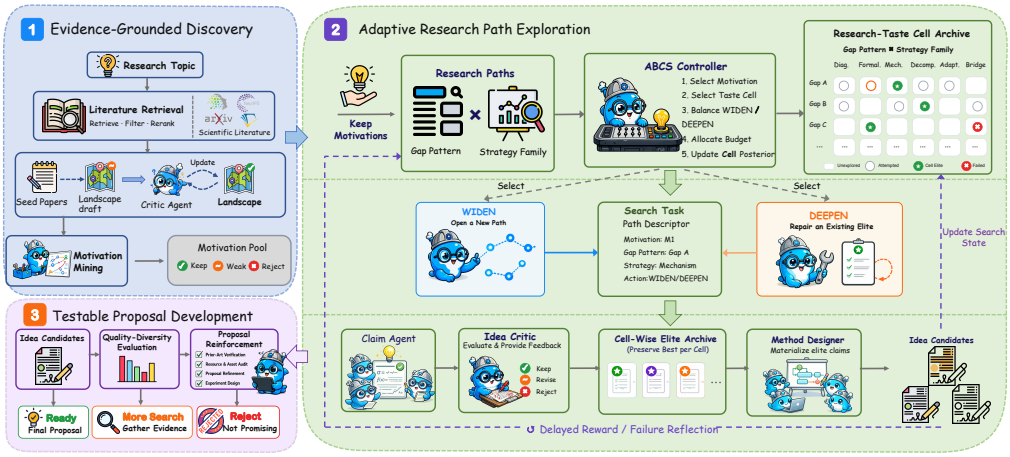

Figure 3: System overview of IdeaFactory. Evidence-grounded discovery extracts literature-supported motivations; ABCS allocates search budget across research paths through WIDEN and DEEPEN; and proposal development turns path elites into testable proposals, with feedback retained in the Idea Discovery Chain.

We introduce **Adaptive Branching Chain Search (ABCS)**, which formulates research ideation as a quality-diversity search problem (Mouret and Clune 2015; Pugh, Soros, and Stanley 2016). Rather than forcing all candidates to compete globally, ABCS organizes research paths according to how opportunities are identified and addressed, preserves promising directions across heterogeneous paths, and adaptively reallocates search budget using accumulated feedback.

Research Paths and Elite Archive. ABCS characterizes each research path by two complementary dimensions. A *gap-discovery pattern* g describes how a research opportunity is identified, while a *strategy family* s describes the high-level research move used to address it. We define a research-path cell as

$$c = (g, s) \in \mathcal{G} \times \mathcal{S}. \tag{3}$$

In our implementation, $\mathcal{G}$ contains ten gap-discovery patterns, such as contradictory evidence, hidden assumptions, and benchmark–reality gaps, while $\mathcal{S}$ contains six strategy families, such as diagnosis, mechanism change, and decomposition. Together they define up to $|\mathcal{G} \times \mathcal{S}| = 10 \times 6 = 60$ research-path cells. This is a theoretical search space: only cells induced by retained motivations are instantiated, and each run explores a budget-limited subset. The complete taxonomy is provided in the Appendix.

For each active cell, ABCS maintains an elite archive containing its strongest claim, accumulated feedback, and a Beta pseudo-posterior (α_c, β_c) representing an uncertainty-aware estimate of whether the path merits further search. A new claim therefore competes first with alternatives following similar research logic rather than immediately with the entire candidate pool, giving heterogeneous paths room to mature before downstream comparison.

Adaptive Widening and Deepening. ABCS allocates search budget through two complementary operations: WIDEN and DEEPEN. Using Thompson-style sampling (Russo et al. 2018), each search round samples

$$\theta_c \sim \mathrm{Beta}(\alpha_c, \beta_c) \tag{4}$$

for candidate cells.

For a retained motivation m, we write the WIDEN priority as

$$\begin{aligned} S_{\mathrm{wide}}(c; m) = \theta_c + \lambda_0 \mathbb{I}[n_c = 0] + \lambda_g u_g(g) + \lambda_s u_s(s) \\ + \lambda_m u_m(m) + \lambda_e e(m, c) - \lambda_b o_{\mathrm{bs}}(c), \end{aligned} \tag{5}$$

where n_c denotes previous attempts in cell c; u_g, u_s, and u_m reward underexplored gap patterns, strategy families, and motivations; $e(m, c)$ measures evidence support; and $o_{\mathrm{bs}}(c)$ penalizes bridge/synthesis over-concentration. Exact definitions and coefficients are provided in the Appendix.

WIDEN therefore expands search coverage by combining estimated path value with empty-cell and underexploration bonuses and evidence support. Motivated by the bridge and synthesis concentration observed in LLM ideation (Chen, Zhao, and Cohan 2026), the final term softly discourages excessive concentration without excluding legitimate synthesis-based directions.

DEEPEN develops promising existing branches. When a cell contains a strong elite but the critic identifies a repairable weakness, ABCS allocates additional budget to the same research path and uses prior critique to refine its mechanism, scope, or validation design. Deepening additionally considers elite quality, repair potential, and repeated-attempt penalties; implementation details are provided in the Appendix. Early search emphasizes WIDEN, while accumulated feedback gradually shifts part of the budget toward DEEPEN, balancing exploration and refinement under a fixed budget.

Critic-Guided Search Update. For each selected search task, a research claim is constructed under the assigned path constraints and evaluated by a critic. We aggregate critic rank, novelty, task alignment, direct contribution, scientific value, retention decision, strategy fit, and evaluation completeness

Algorithm 1 Adaptive Branching Chain Search

Require: Motivation pool $\mathcal{M}$, search state Σ, budget B
Ensure: Cross-cell claim elites $\mathcal{E}$ and updated state Σ

Load or initialize cell archive $\mathcal{A}$ from Σ
$b \leftarrow 0$
while $b < B$ **do**
  Sample $\theta_c \sim \mathrm{Beta}(\alpha_c, \beta_c)$ for candidate cells
  Allocate current budget between WIDEN and DEEPEN
  $\mathcal{T}^W \leftarrow \textsc{SelectWiden}(\mathcal{A}, \mathcal{M}, \{\theta_c\})$
  $\mathcal{T}^D \leftarrow \textsc{SelectDeepen}(\mathcal{A}, \mathcal{M}, \{\theta_c\})$
  $\mathcal{T} \leftarrow \mathcal{T}^W \cup \mathcal{T}^D$
  if $\mathcal{T} = \emptyset$ **then**
    break
  end if
  $\mathcal{X} \leftarrow \textsc{ConstructClaims}(\mathcal{T})$
  $\mathcal{F} \leftarrow \textsc{Critique}(\mathcal{X}, \mathcal{M})$
  for each evaluated (τ, x, f) **do**
    $r_x \leftarrow \textsc{Reward}(x, f)$
    Update pseudo-posterior and elite of cell $c(\tau)$ using r_x
    Store critique and repair feedback
  end for
  $b \leftarrow b + |\mathcal{T}|$
end while
$\mathcal{E} \leftarrow \textsc{CrossCellEliteExtraction}(\mathcal{A})$
Update run-scoped search state Σ
return $\mathcal{E}, \Sigma$

into a bounded provisional reward $r_t \in [0, 1]$; exact scoring rules and weights are provided in the Appendix.

The corresponding cell is updated through fractional pseudo-counts:

$$\alpha_c \leftarrow \alpha_c + r_t, \qquad \beta_c \leftarrow \beta_c + 1 - r_t. \tag{6}$$

A stronger claim may replace the current cell elite, while unsuccessful trials still contribute critique and failure information for later repair. Critic feedback therefore affects both local claim development and subsequent search allocation.

Adaptive Search Procedure. Algorithm 1 summarizes the ABCS search loop. ABCS maintains cell-wise elites, allocates search between WIDEN and DEEPEN using uncertainty-aware path estimates and coverage, and updates cells from critic feedback. Implementation-level selection rules, budget configurations, and stopping criteria are provided in the Appendix.

After search, ABCS extracts high-quality claim elites from distinct active cells rather than applying a global Top-K directly to raw generations. These elites are materialized into candidate ideas and passed to proposal-level evaluation and auditing, allowing heterogeneous directions to undergo path-specific exploration and repair before downstream comparison.

3.4 Multi-Agent Research Orchestration

ABCS determines where search budget should be allocated, while complementary research roles carry out the corresponding development. Role specialization and proposer–critic interaction are common in multi-agent scientific systems (Gottweis et al. 2026; Su et al. 2025; Yang, Li, and Li 2026); IdeaFactory differs by separating global search control from bounded research construction and critique. It organizes these responsibilities into three components: **Search Coordinator**, **Constructive Research Specialists**, and **Epistemic Critics**.

Search Coordinator. The Search Coordinator controls global research-path selection. Based on active cells, current elites, coverage, and accumulated feedback, it determines which direction receives the next unit of search budget and whether to apply WIDEN or DEEPEN. In IdeaFactory, this responsibility is instantiated by the ABCS control layer rather than an additional free-form coordinator LLM. Each selected search task specifies a bounded research path for subsequent execution.

Constructive Research Specialists. Constructive Research Specialists transform literature evidence into structured research understanding and motivations, formulate scientific claims along paths assigned by ABCS, and materialize mature claim elites into candidate ideas. They reason within assigned directions but do not independently redirect global search.

Epistemic Critics. Epistemic Critics determine whether research objects should be retained, revised, or abandoned. Critique operates throughout AID: opportunities are screened before exploration, claims are evaluated for scientific value and potential risks, and mature candidates are examined against prior work and experimental feasibility. Critics identify weaknesses and request repair without replacing the direction being evaluated.

These roles interact through structured research objects. ABCS assigns a bounded search task, Constructive Research Specialists produce the corresponding research objects, and Epistemic Critics return structured feedback. Valid outcomes are written back to the Idea Discovery Chain while preserving evidence provenance and critique history, thereby instantiating the state transition in Equation (1).

The three components directly instantiate the AID requirements introduced in Section 3.1. The Idea Discovery Chain supports *evidence-grounded progression* (R1) by preserving links between claims and supporting evidence. ABCS implements *branching with preservation* (R2) through cell-wise elite archives that prevent premature elimination of heterogeneous paths. The separation of search coordination, construction, and critique provides *controlled research reasoning* (R3), keeping global exploration decisions distinct from local research operations.

At later stages, mature elites from different paths are developed into candidate proposals under a quality-first criterion while preserving complementary directions. Proposal-level evaluation and auditing are linked back to their originating paths as delayed feedback, allowing subsequent search rounds within the same run to reuse accumulated evidence about promising, incomplete, and unsuccessful directions.

IdeaFactory thus combines global search control, bounded research construction, and persistent within-run critique into

a traceable process for Autonomous Idea Development.

4 Experimental Setup

4.1 Core Questions

To evaluate IdeaFactory and validate our main claims, we assess the system from two complementary perspectives: idea quality and research-taste alignment. Specifically, our experiments address two questions:

Q1: Does IdeaFactory generate higher-quality research proposals than existing baselines?

Q2: Does IdeaFactory mitigate idea exploration collapse and better align generated ideas with the human research-taste distribution?

4.2 Evaluation Benchmarks and Metrics

IdeaGen. We construct IdeaGen, a proposal-quality evaluation set containing 57 research-ideation tasks derived from papers accepted at ICLR 2026, ICML 2026, and NeurIPS 2025. The tasks span 15 consolidated AI/ML super-domains. For each task, every system generates three research proposals, which are evaluated along four dimensions: novelty, feasibility, impact, and clarity (Si, Yang, and Hashimoto 2025; Lyu et al. 2026).

IdeaSeed-ML. To evaluate the research-taste distribution of generated ideas, we select 1,037 AI/ML instances from the publicly available IdeaSeed benchmark (Chen, Zhao, and Cohan 2026) and refer to this subset as **IdeaSeed-ML**. Each instance pairs a related-work context with a human proposal; evaluated systems receive only the former. Following the benchmark, we measure distributional alignment along two research-taste axes—opportunity pattern and method paradigm—and examine bridge/synthesis concentration and template-like generation.

Evaluation Protocol. All systems receive identical task inputs and follow the same output requirements. For **IdeaGen**, system identities are anonymized, and the order within each proposal pair is randomized to mitigate positional bias (Wang et al. 2024a). Each pair is independently evaluated by GPT-5.4-mini, Claude Sonnet 5, and Gemini 3.5 Flash. Using judges from three model families reduces dependence on any single evaluator. We report win, tie, and loss rates over their judgments and additionally conduct human-expert evaluation on a subset of proposal pairs. For **IdeaSeed-ML**, we follow the IdeaSeed evaluation protocol (Chen, Zhao, and Cohan 2026), including its taxonomy, annotation procedure, and distributional metrics. We report paired bootstrap confidence intervals, with detailed prompts, metric definitions, and implementation configurations provided in the Appendix.

4.3 Baselines

We compare IdeaFactory with representative autonomous research systems, including AI Scientist-v2 (Yamada et al. 2025), EvoScientist (Lyu et al. 2026), InternAgent (Feng et al. 2026), ARIS (Yang, Li, and Li 2026), AI-Researcher (Tang et al. 2025), and Idea2Paper (KAUST-ARK 2026). We also include GPT-5.5 (OpenAI 2026b), Gemini 3.1 Pro (Google DeepMind 2026), Claude Opus 4.8 (Anthropic 2026), GPT-5.4 mini (OpenAI 2026a), DeepSeek-V4-Flash and DeepSeek-V4-Pro (DeepSeek-AI 2026), Qwen3.7-Plus (Alibaba Cloud 2026), Qwen3-32B, and Qwen3-8B (Yang et al. 2025) as direct-generation baselines. Detailed descriptions of all baselines are provided in the Appendix.

5 Experimental Results

5.1 Idea Quality (Q1)

We first evaluate whether IdeaFactory generates higher-quality research proposals than existing autonomous research systems and frontier direct-generation models. Tables 1 and 2 report the automated and human evaluation results, respectively.

As shown in Table 1, IdeaFactory achieves a positive Avg. gap against every autonomous research system and direct-generation LLM. Against GPT-5.5, Gemini 3.1 Pro, and Claude Opus 4.8, the margins reach $+72.66$, $+71.20$, and $+66.33$, respectively. The gains are particularly consistent in feasibility and clarity, but are not limited to better proposal formulation. Against GPT-5.5, IdeaFactory also achieves win rates of 80.12% in novelty and 72.71% in impact, while the corresponding rates against EvoScientist reach 84.02% and 81.48%. These results suggest that the benefit of AID extends beyond presenting ideas more completely to improving the research direction and underlying mechanism themselves.

Human evaluation in Table 2 shows the same preference pattern. IdeaFactory obtains Avg. gaps of $+63.02$, $+82.90$, and $+81.87$ against AI Scientist-v2, EvoScientist, and GPT-5.5, respectively, while retaining clear advantages in novelty and impact. The agreement between automated and human evaluation therefore supports Q1: **IdeaFactory consistently develops higher-quality research proposals across complementary quality dimensions.**

5.2 Research-Taste Alignment (Q2)

High-quality individual proposals do not necessarily imply broad exploration of the research space. Following the research-taste evaluation perspective of Chen, Zhao, and Cohan (2026), we compare generated and human ideas at the distribution level along opportunity patterns and method paradigms. This directly tests the phenomenon introduced in Section 1: whether IdeaFactory mitigates *idea exploration collapse* rather than merely producing stronger proposals within a narrow set of familiar research moves.

Table 3 shows that **IdeaFactory substantially closes the distributional gap to human research taste**. It achieves the lowest TVD and JSD among all non-human systems on both axes, with Opportunity TVD/JSD of $0.218/0.043$ and Method TVD/JSD of $0.271/0.082$. Most notably, its Opportunity entropy reaches 0.868, nearly matching the human value of 0.866, indicating that the diversity of opportunity discovery is recovered to a near-human level. Method entropy also increases to 0.861, and although it remains below the human value of 0.980, IdeaFactory still yields the closest overall method distribution among the evaluated systems. The improvement therefore reflects closer alignment with the human distribution rather than diversity for its own sake.

Method	Feasibility			Novelty			Impact			Clarity			Avg. gap
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	
Autonomous Research Systems													
IdeaFactory vs. AI Scientist-v2	**92.01**	0.19	7.80	**50.88**	17.93	31.19	**60.23**	13.26	26.51	**90.84**	1.75	7.41	**+55.26**
IdeaFactory vs. EvoScientist	**91.81**	0.58	7.60	**84.02**	8.38	7.60	**81.48**	8.77	9.75	**95.13**	0.00	4.87	**+80.65**
IdeaFactory vs. InternAgent	**82.65**	2.53	14.81	**50.29**	17.93	31.77	**62.38**	16.37	21.25	**61.60**	18.52	19.88	**+42.30**
IdeaFactory vs. ARIS	**71.93**	1.56	26.51	**86.55**	5.65	7.80	**79.14**	8.58	12.28	**79.14**	2.53	18.32	**+62.96**
IdeaFactory vs. AI-Researcher	**88.11**	1.36	10.53	**46.20**	20.27	33.53	**48.93**	15.98	35.09	**83.04**	4.48	12.48	**+43.66**
IdeaFactory vs. Idea2Paper	**98.64**	0.19	1.17	**75.24**	8.77	15.98	**75.44**	7.60	16.96	**98.25**	0.97	0.78	**+78.17**
Direct Generation LLMs													
IdeaFactory vs. GPT-5.5	**86.94**	0.39	12.67	**80.12**	8.58	11.31	**72.71**	16.76	10.53	**92.01**	1.36	6.63	**+72.66**
IdeaFactory vs. Gemini 3.1 Pro	**97.27**	0.39	2.34	**67.45**	15.20	17.35	**71.93**	11.11	16.96	**90.25**	4.29	5.46	**+71.20**
IdeaFactory vs. Claude Opus 4.8	**84.60**	1.36	14.04	**78.75**	7.21	14.04	**73.29**	12.09	14.62	**83.82**	3.70	12.48	**+66.33**

Table 1: Pairwise proposal-quality evaluation on IdeaGen. Win/tie/loss rates compare IdeaFactory with each baseline; Avg. gap is the mean win–loss margin across the four dimensions.

Method	Feasibility			Novelty			Impact			Clarity			Avg. gap
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	
Autonomous Research Systems													
IdeaFactory vs. AI Scientist-v2	**90.64**	8.77	0.58	**47.37**	33.92	18.71	**56.14**	25.73	18.13	**95.91**	3.51	0.58	**+63.02**
IdeaFactory vs. EvoScientist	**98.25**	1.17	0.58	**83.04**	14.04	2.92	**69.01**	24.56	6.43	**93.57**	4.09	2.34	**+82.90**
Direct Generation LLMs													
IdeaFactory vs. GPT-5.5	**92.40**	5.26	2.34	**74.85**	19.88	5.26	**76.02**	18.13	5.85	**97.66**	2.34	0.00	**+81.87**

Table 2: Human evaluation on a stratified IdeaGen subset for three representative baselines. Win/tie/loss rates and Avg. gap follow Table 1.

Source	Opportunity Pattern			Method Paradigm		
	TVD ↓	JSD ↓	Ent. ↑	TVD ↓	JSD ↓	Ent. ↑
Human	–	–	0.866	–	–	0.980
GPT-5.5	0.466	0.197	0.639	0.436	0.204	0.674
GPT-5.4 mini	0.485	0.205	0.617	0.439	0.189	0.691
Gemini 3.1 Pro	0.270	0.098	0.750	0.374	0.135	0.792
Claude Opus 4.8	0.324	0.105	0.784	0.302	0.098	0.823
DeepSeek-V4-Pro	0.359	0.148	0.692	0.395	0.176	0.725
DeepSeek-V4-Flash	0.376	0.158	0.681	0.460	0.234	0.637
Qwen3.7-Plus	0.365	0.128	0.727	0.384	0.157	0.741
Qwen3-32B	0.539	0.265	0.549	0.531	0.283	0.561
Qwen3-8B	0.273	0.085	0.775	0.484	0.244	0.615
IdeaFactory	**0.218**	**0.043**	**0.868**	**0.271**	**0.082**	**0.861**

Table 3: Research-taste distributions on 1,037 paired IdeaSeed-ML instances. TVD/JSD measure distance to the human distribution, and Ent. denotes normalized entropy; bold marks the best non-human result.

Source	Bridge/Synth. (%)			Diagnostics ↓		
	Br.	Syn.	Either	Surf.	Flag (%)	Boil.
Human	12.0	14.3	15.9	0.60	7.3	0.90
GPT-5.5	58.4	57.9	63.6	0.99	21.2	0.80
GPT-5.4 mini	60.5	58.2	65.0	0.96	18.3	0.87
Gemini 3.1 Pro	36.2	44.7	47.0	1.14	31.4	0.98
Claude Opus 4.8	43.4	43.1	48.6	1.02	25.5	0.92
DeepSeek-V4-Pro	47.7	53.5	57.4	1.29	39.6	0.97
DeepSeek-V4-Flash	49.6	60.3	62.6	1.71	63.3	1.14
Qwen3.7-Plus	48.4	52.7	56.5	1.16	32.0	0.95
Qwen3-32B	65.9	67.4	73.5	1.57	56.8	1.06
Qwen3-8B	38.2	62.7	63.6	1.84	74.1	1.35
IdeaFactory	**21.1**	**13.2**	**22.3**	**0.10**	**0.2**	**0.30**

Table 4: Bridge/Synthesis concentration and proposal diagnostics on IdeaSeed-ML. Either is the union of Bridge and Synthesis; Flag denotes surface-stitching score ≥ 2. Diagnostics use a 0–3 scale.

Table 4 further reveals where this alignment comes from. Direct-generation models exhibit strong concentration on Bridge opportunities and Synthesis/Unification paradigms, whereas IdeaFactory reduces these rates to 21.1% and 13.2%,

with an Either rate of 22.3%, much closer to the human reference of 15.9%. The goal is not to eliminate Bridge or Synthesis as legitimate research strategies, but to prevent them from becoming default research moves across unrelated contexts. IdeaFactory also produces substantially lower surface-stitching and boilerplate diagnostics, suggesting that broader exploration does not arise from superficial recombination or template variation. Together with Q1, these results show that **IdeaFactory narrows the human–LLM research-taste gap without sacrificing proposal quality**, directly supporting our claim that it mitigates idea exploration collapse.

5.3 Ablation Study

We finally examine how the two central mechanisms of ABCS contribute to proposal quality and research-taste alignment.

Variant	Quality↑	Opp. TVD↓	Method TVD↓	Bridge↓	Synth.↓
Full ABCS	**8.336**	**0.218**	**0.271**	**21.1**	**13.2**
w/o Adaptive Branching	8.108	0.302	0.326	33.7	21.4
w/o QD Preservation	8.247	0.281	0.357	30.2	29.1
w/o ABCS	7.846	0.368	0.416	45.8	38.6

Table 5: Ablation of ABCS on proposal quality and research-taste alignment.

Table 5 shows that Adaptive Branching and quality-diversity preservation play complementary roles. Removing Adaptive Branching only moderately reduces average quality from 8.336 to 8.108, but increases Opportunity TVD from 0.218 to 0.302 and Bridge concentration from 21.1% to 33.7%. Its main contribution is therefore to broaden where the system explores rather than simply increasing the score of individual proposals.

Removing quality-diversity preservation leaves proposal quality relatively high at 8.247, yet Method TVD rises to 0.357 and Synthesis concentration more than doubles from 13.2% to 29.1%. Global quality selection can therefore retain strong proposals while still collapsing onto similar methodological patterns. Removing ABCS entirely further degrades both quality and alignment, with average quality falling to 7.846 and Opportunity/Method TVD increasing to $0.368/0.416$. Since this variant retains the underlying AID pipeline, the comparison isolates the incremental benefit of adaptive search over a strong multi-stage ideation scaffold. Overall, **Adaptive Branching determines where the search expands, while quality-diversity preservation prevents heterogeneous directions from collapsing during selection**; together they enable IdeaFactory to maintain strong proposal quality while mitigating idea exploration collapse.

6 Conclusion

We presented IdeaFactory, which formulates research ideation as Autonomous Idea Development. Rather than one-shot generation and selection, IdeaFactory progressively develops research directions through evidence accumulation, quality-diversity search, and critic-guided repair. Adaptive Branching Chain Search organizes heterogeneous paths into research cells, balancing exploration of underexplored directions with refinement of promising candidates. Experiments demonstrate that IdeaFactory generates higher-quality proposals while exploring a broader range of research paths, substantially narrowing the distributional gap between LLM-generated and human research ideas.

References

Alibaba Cloud. 2026. Qwen3.7-Plus Model Documentation. https://help.aliyun.com/en/model-studio/newly-released-models. Model identifier qwen3.7-plus; released June 1, 2026; accessed July 28, 2026.

Anthropic. 2026. Introducing Claude Opus 4.8. https://www.anthropic.com/news/claude-opus-4-8. Released May 28, 2026; accessed July 28, 2026.

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

Chen, Z.; Zhao, Y.; and Cohan, A. 2026. Measuring the Gap Between Human and LLM Research Ideas. *arXiv preprint arXiv:2607.01233*.

DeepSeek-AI. 2026. DeepSeek V4 Preview Release. https://api-docs.deepseek.com/news/news260424/. Released April 24, 2026; accessed July 28, 2026.

Feng, S.; Ma, R.; Yan, X.; et al. 2026. InternAgent-1.5: A Unified Agentic Framework for Long-Horizon Autonomous Scientific Discovery. *arXiv preprint arXiv:2602.08990*.

Google DeepMind. 2026. Gemini 3.1 Pro Model Card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Pro-Model-Card.pdf. Released February 19, 2026; accessed July 28, 2026.

Gottweis, J.; et al. 2026. Accelerating scientific discovery with Co-Scientist. *Nature*, 655(8122): 487–496.

Jin, J.; Hu, Y.; Qiu, K.; Dai, Q.; Luo, C.; Dong, G.; et al. 2026. Toward Generalist Autonomous Research via Hypothesis-Tree Refinement. *arXiv preprint arXiv:2606.11926*.

KAUST-ARK. 2026. Idea2Paper. https://idea2paper.org/. Accessed 2026-07-28.

Liu, Y.; Yang, Z.; Xie, T.; Ni, J.; Gao, B.; Li, Y.; Tang, S.; Ouyang, W.; Cambria, E.; and Zhou, D. 2026. ResearchBench: Benchmarking LLMs in Scientific Discovery via Inspiration-Based Task Decomposition. In *Findings of the Association for Computational Linguistics: ACL 2026*, 13187–13207. San Diego, California, United States: Association for Computational Linguistics.

Lu, C.; Lu, C.; Lange, R. T.; Foerster, J.; Clune, J.; and Ha, D. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *arXiv preprint arXiv:2408.06292*.

Lyu, Y.; Zhang, X.; Yi, X.; Zhao, Y.; Guo, S.; Hu, W.; et al. 2026. EvoScientist: Towards Multi-Agent Evolving AI Scientists for End-to-End Scientific Discovery. *arXiv preprint arXiv:2603.08127*.

Mouret, J.-B.; and Clune, J. 2015. Illuminating Search Spaces by Mapping Elites. *arXiv preprint arXiv:1504.04909*.

OpenAI. 2026a. Introducing GPT-5.4 Mini and Nano. https://openai.com/index/introducing-gpt-5-4-mini-and-nano/. Released March 17, 2026; accessed July 28, 2026.

OpenAI. 2026b. Introducing GPT-5.5. https://openai.com/index/introducing-gpt-5-5/. Released April 23, 2026; accessed July 28, 2026.

Pugh, J. K.; Soros, L. B.; and Stanley, K. O. 2016. Quality Diversity: A New Frontier for Evolutionary Computation. *Frontiers in Robotics and AI*, 3: 40.

Russo, D.; Van Roy, B.; Kazerouni, A.; Osband, I.; and Wen, Z. 2018. A Tutorial on Thompson Sampling. *Foundations and Trends in Machine Learning*, 11(1): 1–96.

Si, C.; Yang, D.; and Hashimoto, T. 2025. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. In *The Thirteenth International Conference on Learning Representations*.

Su, H.; Chen, R.; Tang, S.; Yin, Z.; Zheng, X.; Li, J.; Qi, B.; Wu, Q.; Li, H.; Ouyang, W.; Torr, P.; Zhou, B.; and Dong, N. 2025. Many Heads Are Better Than One: Improved Scientific Idea Generation by A LLM-Based Multi-Agent System. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 28201–28240. Vienna, Austria: Association for Computational Linguistics.

Tang, J.; Xia, L.; Li, Z.; and Huang, C. 2025. AI-Researcher: Autonomous Scientific Innovation. In Belgrave, D.; Zhang, C.; Lin, H.-T.; Pascanu, R.; Koniusz, P.; Ghassemi, M.; and Chen, N., eds., *Advances in Neural Information Processing Systems*, volume 38, 9481–9520. Curran Associates, Inc.

Tong, J.; Li, M.; Li, H.; Yang, Y.; Mou, Y.; Ma, W.; Xi, Z.; Chen, H.; Liu, X.; Cheng, Q.; et al. 2026. AI Can Learn Scientific Taste. *arXiv preprint arXiv:2603.14473*.

Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Kong, L.; Liu, Q.; Liu, T.; and Sui, Z. 2024a. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9440–9450. Bangkok, Thailand: Association for Computational Linguistics.

Wang, Q.; Downey, D.; Ji, H.; and Hope, T. 2024b. SciMON: Scientific Inspiration Machines Optimized for Novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 279–299. Bangkok, Thailand: Association for Computational Linguistics.

Xu, T.; Qian, Z.; Liu, G.; Ling, L.; Zhang, Z.; et al. 2026. Idea2Story: An Automated Pipeline for Transforming Research Concepts into Complete Scientific Narratives. *arXiv preprint arXiv:2601.20833*.

Yamada, Y.; Lange, R. T.; Lu, C.; Hu, S.; Lu, C.; Foerster, J.; Clune, J.; and Ha, D. 2025. The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search. *arXiv preprint arXiv:2504.08066*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; et al. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Yang, R.; Li, Y.; and Li, S. 2026. ARIS: Autonomous Research via Adversarial Multi-Agent Collaboration. *arXiv preprint arXiv:2605.03042*.

Zhao, Q.; Huang, Y.; Dai, Y.; Xiao, L.; Gao, J.; Zhang, X.; et al. 2026. ResearchStudio-Idea: An Evidence-Grounded Research-Ideation Skill Suite from ML Conference Outcomes. *arXiv preprint arXiv:2607.04439*.