IDEAS HAVE GENOMES: BENCHMARKING SCIENTIFIC LINEAGE REASONING AND LINEAGE-GROUNDED IDEA GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific ideas rarely start from a blank page. They inherit mechanisms, repair known limitations, and recombine pieces of earlier work, much like biological genomes. Current benchmarks still say little about whether AI systems can follow this inheritance structure. We present **IdeaGene-Bench (IG-Bench)**, a benchmark for scientific lineage reasoning and lineage-grounded idea generation. IG-Bench is organized around the IdeaGene framework: each paper or proposal is represented as a set of minimal, typed, evidence-grounded *Idea Genome* objects, and a GenomeDiff aligns these objects to record inheritance, mutation, loss, external import, and novel insertion under six operational evolutionary dynamics. The benchmark contains 1,961 golden lineage traces, 1,085 curated *Idea Genome* objects, and 920 pairwise GenomeDiff records across 10 scientific domains. It supports two evaluations. IG-Exam (42 task types, 1,029 instances) tests closed-form lineage reasoning across *Idea Genome* abstraction, inheritance tracing, evolutionary reasoning, and lineage verification. IdeaGene-Gym (IG-Gym) evaluates generation with a Population-Evolution Score (PES): three LLM judges apply domain-expert-annotated rubrics to assess whether a proposal inherits the right *Idea Genome* objects, introduces meaningful variation, and offers selection value within a scientific lineage. Experiments on 14 LLM-based scientists expose a compositional bottleneck. The strongest system reaches only 27.3% exact accuracy on lineage reasoning, and structured lineage context reshuffles system rankings rather than helping every participant uniformly.

1 INTRODUCTION

LLM-based auto-research systems now search literature, synthesize hypotheses, run experiments, and write paper-like reports (Lu et al., 2024; Yamada et al., 2025; Tang et al., 2025). Existing evaluations emphasize retrieval quality, factuality, writing fluency, novelty, workflow competence, or pairwise preference (Ajith et al., 2024; Sun et al., 2024; Xu et al., 2025; Wang et al., 2025; 2024; Si et al., 2024; Zheng et al., 2023). These criteria leave a harder question mostly untested: when a proposal claims to extend a research direction, does it inherit the right mechanism, repair the right limitation, and stay coherent with the lineage it builds on? Figure 1 motivates this shift from paper-centric context gathering to genome-centric lineage comparison.

The distinction matters because scientific progress is not the same as topical proximity. Two papers can share a task without inheriting from one another, and two papers can look textually distant while carrying the same core mechanism forward. YOLOv2 (Redmon & Farhadi, 2017), for example, modifies YOLO's (Redmon et al., 2016) single-shot detection mechanism through anchor boxes, batch normalization, and multi-scale training: the lineage is visible as local mutation. DETR (Carion et al., 2020) also studies object detection, but it replaces the Faster R-CNN (Ren et al., 2015) pipeline with set prediction and a Transformer architecture. It shares the task ecology, not the anchor-based driver mechanism. Titles, abstracts, citation edges, and embeddings tend to blur these cases together. A model can retrieve the right papers and still miss the parent mechanism or repaired limitation; a generated idea can sound novel while inheriting no coherent mechanism.

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

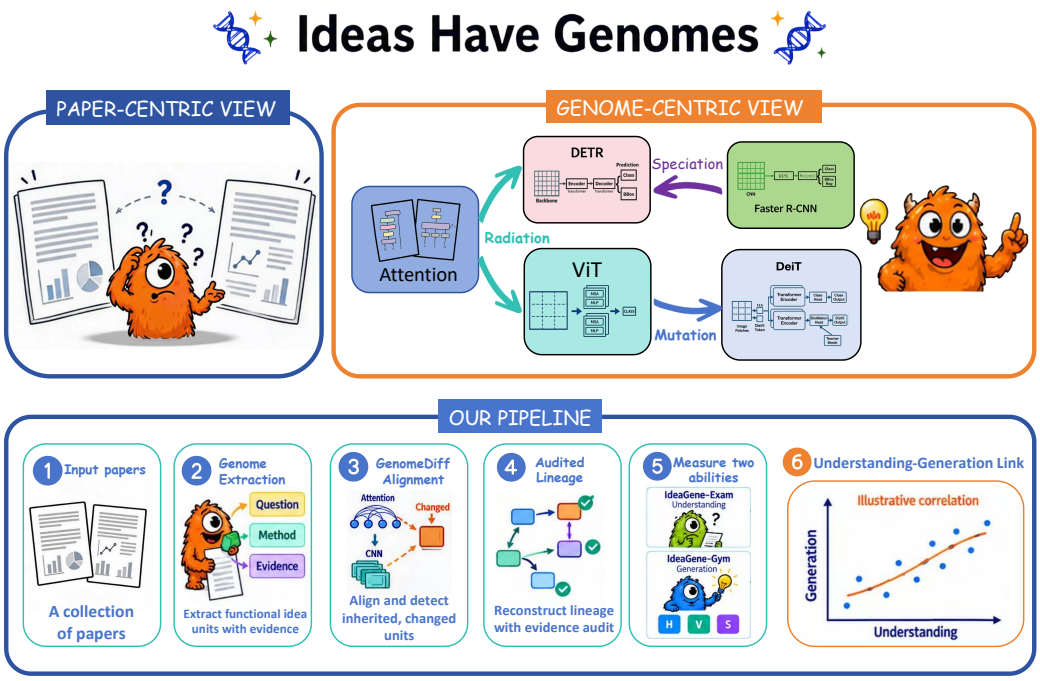

Figure 1: **From paper-centric search to genome-centric lineage.** A paper-centric view leaves a model with many related papers but no explicit unit for deciding which ideas are inherited, mutated, recombined, or merely co-located. The genome-centric view first extracts *Idea Genome* objects, then aligns them through GenomeDiff records into populations and lineages, making evolutionary relations explicit enough to evaluate.

Evaluating this behavior requires a representation below the paper level, one that makes idea inheritance explicit enough to audit. We call the target capability *scientific lineage competence*: the ability to (i) abstract a paper into *Idea Genome* objects, (ii) trace which *Idea Genome* objects persist, mutate, or disappear across papers, (iii) explain the dominant dynamics of a transition, (iv) verify whether a proposed lineage is coherent rather than merely topical, and (v) generate a proposal that can be inserted into an existing lineage as a plausible descendant.

IdeaGene turns this capability into an evaluable object. It represents each paper or proposal as a set of ***Idea Genome*** objects, where each *Idea Genome* is a minimal, typed, evidence-grounded idea structure used for lineage comparison. A **GenomeDiff** aligns *Idea Genome* objects across predecessor and successor work, recording inheritance, mutation, loss, external import, and novel insertion. This is an operational choice rather than a biological theory of science: the framework fixes the granularity and evidence contract needed for comparison.

We instantiate the framework as **IdeaGene-Bench (IG-Bench)**. The benchmark contains 1,961 golden lineage traces across 10 scientific domains, 1,085 curated *Idea Genome* objects, and 920 pairwise GenomeDiff records. It has two parts. IG-Exam is a closed-form benchmark with 42 task types and 1,029 instances, testing *Idea Genome* abstraction, inheritance tracing, evolutionary reasoning, and lineage verification. IG-Gym is an open-ended generation evaluation: systems write proposals under three controlled information settings—Question-only, Library, and Lineage—and receive a lineage-conditioned Population-Evolution Score (PES) from three LLM judges applying domain-expert-annotated rubrics. PES asks whether a proposal can be inserted as a coherent descendant in a specified lineage population: inheriting the right *Idea Genome* objects (Heredity), introducing meaningful variation relative to neighboring work (Variation), and offering selection value for future research (Selection).

Experiments on 14 LLM-based scientists—direct LLMs, research-agent frameworks, and CLI harnesses—show that plausible research text is not the same as lineage competence. On IG-Exam, the best system reaches only 27.3% exact accuracy. The common failure is compositional: models often recover local signals but fail to keep parent choice, driver assignment, object fate, and verifi-

cation flags jointly consistent. In IG-Gym, structured lineage context does not simply raise every score; it reshuffles the ranking, separating systems that can use lineage evidence from those that only benefit from more text.

Our contributions are:

1. **IdeaGene framework for scientific lineage.** We define *Idea Genome*, genome extraction, and GenomeDiff, together with six operational evolutionary dynamics (mutation, adaptive radiation, hybridization, speciation, niche competition, isolation) for aligning, classifying, and auditing idea inheritance across papers.
2. **IG-Bench benchmark.** We build a two-part benchmark: IG-Exam for closed-form lineage reasoning (42 task types, 1,029 instances) and IG-Gym for lineage-grounded idea generation with expert-annotated PES rubrics applied by three LLM judges under controlled information settings.
3. **Empirical findings.** We show that frontier systems face a compositional bottleneck in lineage reasoning, that structured lineage evidence separates systems rather than uniformly helping them, and that generated ideas often sound plausible before they are lineage-coherent.

2 Related Work

IG-Bench sits between several evaluation traditions. Scientific lineage needs both a representation layer and a benchmark: paper-level relevance alone is too coarse, while generation-only preference is too unconstrained.

Scientific retrieval and QA. Scientific retrieval and representation benchmarks such as LitSearch and SciRepEval (Ajith et al., 2024; Singh et al., 2023) test whether systems can find or embed relevant papers; sentence and document embeddings (Reimers & Gurevych, 2019; Wang et al., 2022) and multi-document scientific summarization (Lu et al., 2020) add complementary representation tasks. Scientific QA, reasoning, general-intelligence, and end-to-end autonomous-research suites (Sun et al., 2024; Xu et al., 2025; Wang et al., 2025; Xu et al., 2026; Welbl et al., 2017; Sawada et al., 2023) test factual, workflow, and problem-solving competence. IG-Bench asks a different question: once the relevant work is available, can a system identify which idea structures are actually transmitted across papers?

Automated research and ideation. Recent systems automate literature search, hypothesis generation, experimentation, and paper drafting (Lu et al., 2024; Yamada et al., 2025; Tang et al., 2025; Schmidgall et al., 2025). Other work studies novelty-aware generation (Wang et al., 2024; Hu et al., 2024; Zhao et al., 2025), iterative ideation from literature (Baek et al., 2025; Si et al., 2024; Li et al., 2024), scientific-literature synthesis (Asai et al., 2026) and retrieval-grounded long-form writing (Shao et al., 2024), and multi-agent or fully autonomous scientific collaboration (Ghafarollahi & Buehler, 2025; Liu et al., 2026). These systems make lineage-grounded evaluation more urgent. A proposal can be fluent and literature-aware, yet still fail to inherit the parent mechanism or repair the stated limitation of the work it claims to extend.

Science of science and citation structure. Science-of-science studies map citations and paradigm shifts (Fortunato et al., 2018; Jurgens et al., 2018; Kuhn, 1962); citation-intent classification identifies functional roles (Cohan et al., 2019), and Intern-Atlas represents method-level evolution (Wu et al., 2026). Drawing on evolutionary concepts (Darwin, 1859), IG-Bench evaluates explicit GenomeDiff alignments to test which mechanisms are inherited or changed across papers.

LLM evaluation methodology. LLM-as-judge and Chatbot Arena support scalable assessment (Zheng et al., 2023; Chiang et al., 2024), with position and length biases motivating controls (Wang et al., 2023; Dubois et al., 2024). Holistic benchmarks assess multiple dimensions (Liang et al., 2023; Lin et al., 2024), while agent benchmarks test tool use (Jimenez et al., 2024; Mialon et al., 2023; Yao et al., 2024). IG-Gym evaluates lineage competence through expert-annotated rubrics applied by three LLM judges to genome-centric lineage and population context.

Table 1: Positioning of IG-Bench relative to adjacent evaluation paradigms.

Paradigm	Unit	Genome?	Diff?	Understand?	Generate?
Paper retrieval / QA	Paper or passage	×	×	Factual	×
Scientific embeddings	Paper vector	×	×	Similarity	×
Citation graph analysis	Citation edge	×	Partial	Structural	×
Literature-grounded ideation	Paper summary	Partial	Implicit	Weak	✓
Automated research agents	Workflow trace	No fixed unit	×	Indirect	✓
IG-Bench	*Idea Genome*	✓	✓	✓	✓

Automated research agents include AI Scientist-v2, AI-Researcher, Agent Laboratory, CoI-Agent, SciAgents, and AutoResearchClaw.

3 IDEAGENE FRAMEWORK

The evolutionary analogy gives lineage evaluation a useful vocabulary (Darwin, 1859), but IG-Bench does not require a broad biological ontology of science. It uses a smaller operational layer: *Idea Genome* objects, genome extraction, and GenomeDiff alignments. These three pieces make inheritance explicit enough to evaluate.

3.1 SCIENTIFIC LINEAGE COMPETENCE

Scientific lineage competence is the ability to identify, verify, and extend the mechanism-level inheritance structure behind scientific work. A lineage-competent system should abstract a paper into heritable *Idea Genome* objects, trace which objects persist or change across papers, explain the dominant dynamics of a transition, reject merely topical links, and generate a proposal that can enter an existing lineage as a plausible descendant.

3.2 IDEA GENOME

Within IdeaGene, an ***Idea Genome*** is the minimal auditable structure used for lineage evaluation: a typed, evidence-grounded, lineage-relevant idea object extracted from a paper or proposal. We represent a paper or proposal as a set of *Idea Genome* objects:

$$G(p) = \{g_i = (t_i, z_i, e_i, c_i)\}_{i=1}^{m_p}. \tag{1}$$

Each *Idea Genome* object g_i has a role type $t_i \in \{\texttt{niche}, \texttt{mechanism}, \texttt{observation}, \texttt{limitation}, \texttt{delta}, \texttt{claim}\}$, a content description z_i, an evidence pointer e_i, and optional constraints c_i. The role type states what the object does in lineage reasoning: `niche` marks the problem environment, `mechanism` marks an inheritable method or design, `observation` marks a motivating empirical pattern, `limitation` marks a defect or bottleneck, `delta` marks the repair or design change relative to prior work, and `claim` marks the asserted outcome.

3.3 GENOME EXTRACTION

Genome extraction $\mathcal{E} : p \mapsto G(p)$ converts a paper or proposal into typed, evidence-grounded objects for GenomeDiff alignment and lineage verification. Each object must have a functional role, a source pointer, and enough context to express one heritable idea. Independently transferable components are split; mutually dependent components remain together. Light abstraction is allowed, but unsupported inference is excluded. We retain details whose inheritance, loss, or modification changes the lineage judgment, and treat the rest as metadata.

3.4 GENOMEDIFF

Given a predecessor p_s and successor p_t, a **GenomeDiff** $\Delta_{s\to t}$ aligns *Idea Genome* objects in $G(p_s)$ to *Idea Genome* objects in $G(p_t)$ by type and semantic role. Source objects are marked INHERITED, MUTATED, or LOST; unaligned target objects are NOVEL or EXTERNAL. Each GenomeDiff record also stores the primary transition driver, the relation to the surrounding task or domain setting, and an evidence-backed rationale. We construct these records through LLM-assisted extraction followed by expert audit.

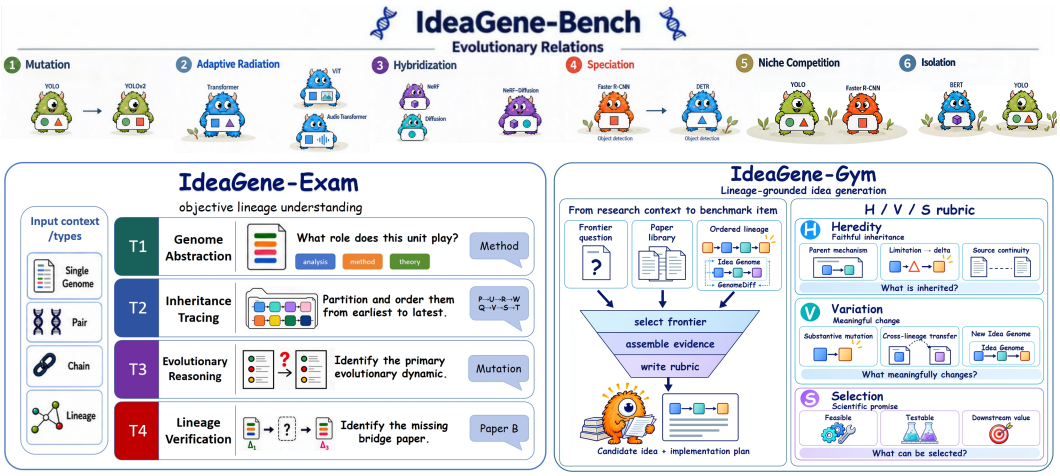

Figure 2: **IG-Bench evaluation design.** IG-Bench converts input papers into a lineage substrate of audited *Idea Genome* objects and GenomeDiff records, then evaluates two complementary capabilities: IG-Exam for closed-form lineage understanding and IG-Gym for lineage-grounded idea generation scored by three LLM judges using expert-annotated H/V/S rubrics.

Lineage versus co-location. The surrounding task, benchmark, dataset convention, and community form an *EcologyContext*: they explain why papers occupy the same research environment, but they do not by themselves establish descent. Genome continuity is what makes a lineage claim. Shared setting without driver inheritance is treated as niche competition; inherited mechanisms moving into a new setting are treated as adaptive radiation.

3.5 Evolutionary Dynamics

Evolutionary dynamics classify GenomeDiff patterns into operational categories. The first check is driver inheritance. Without evidence for it, the relation is co-located rather than lineage.

The six categories are Mutation, Adaptive Radiation, Hybridization, Speciation, Niche Competition, and Isolation; Table 6 in Appendix G gives their criteria and examples.

Ambiguous cases use a fixed priority rule: Hybridization before Speciation, Speciation before Niche Competition when lineage evidence exists, and Adaptive Radiation before Mutation when the setting shift is the driver. These evolutionary dynamics are operational categories for consistent evaluation; they do not claim to exhaust every possible pattern of scientific development.

4 IG-Bench: Dataset and Evaluation Artifact

IG-Bench instantiates the IdeaGene framework as a reusable benchmark. It has two parts: IG-Exam for closed-form lineage understanding and IG-Gym for lineage-grounded idea generation (Figure 2).

4.1 Data Construction and Quality Assurance

IG-Bench contains 1,961 golden lineage traces across 10 scientific domains—including NLP (Devlin et al., 2019; Radford et al., 2019), computer vision (Redmon et al., 2016; Carion et al., 2020; Dosovitskiy et al., 2021), multimodal learning (Radford et al., 2021; Liu et al., 2023), and six additional domains (biology, chemistry, physics, materials, medicine, mathematics). These traces cover 1,085 curated *Idea Genome* objects and 920 pairwise GenomeDiff records. Construction has four stages:

1. **Seed collection.** Experts nominate landmark and frontier papers in each domain.
2. **Trace expansion.** We expand candidate predecessors and successors through citation links, semantic search, and domain curation, producing lineage traces of 3–7 papers.

3. **Genome extraction and diff alignment.** Multi-pass LLM-assisted extraction converts each paper into typed *Idea Genome* objects, which experts then audit. Pairwise GenomeDiff records add object alignment, fate annotation, and driver labels.
4. **Benchmark-level audit.** Programmatic checks and held-out annotators verify schema validity, answer contracts, temporal consistency, anonymization leakage, and trace coherence.

Quality assurance. We recruited 50 graduate annotators (master's and doctoral students across computer science, biology, physics, materials science, and other disciplines). They validated three points: GenomeDiff relation labels and role-type assignments during construction, IG-Exam item difficulty through stratified human solving, and IG-Gym pairwise battles. Construction disagreements were adjudicated by a third annotator. Low human accuracy on IG-Exam was checked to reflect compositional difficulty rather than label noise, and human judges reached 80% agreement with the model-judge panel in IG-Gym. Inter-annotator agreement on dynamics labels is 84.7% before adjudication.

4.2 IG-EXAM: LINEAGE UNDERSTANDING

IG-Exam contains 42 main-challenge task types and 1,029 instances. Each item is

$$\tau = (c, x, \mathcal{A}, y, m), \tag{2}$$

where c is the capability axis, x is anonymized context, $\mathcal{A}$ is the answer space, y is the gold target, and m stores metadata. Scoring uses exact match: every required field must be correct at the same time. A lineage answer is not reliable if it identifies the right parent but assigns the wrong driver or object fate.

Table 7 in Appendix G summarizes the four capability axes, task counts, and representative tasks.

T1 tests whether a model can read a single *Idea Genome*. T2 asks it to trace inheritance across multiple *Idea Genome* objects. T3 requires an explanation of a transition under GenomeDiff criteria. T4 verifies proposed lineage claims and is the closest closed-form analogue to idea generation, because a generated descendant must pass the same parent, coherence, and evidence checks.

4.3 IG-GYM: LINEAGE-GROUNDED GENERATION

IG-Gym evaluates open-ended proposals under controlled information settings. Each instance contains a frontier question, a research domain, an optional paper library, an optional structured lineage, and a scoring rubric. The three settings separate different sources of capability:

- **Question**: domain and frontier question only; tests parametric ideation.
- **Library**: unordered paper summaries; tests paper-centric context use.
- **Lineage**: ordered trace with *Idea Genome* objects and GenomeDiff evidence; tests lineage-structured context use.

Population-Evolution Score (PES). The primary metric is PES, a lineage-conditioned population-insertion score. Rather than judging an idea in isolation, PES evaluates each generated proposal x as a candidate descendant of an ordered lineage $\mathcal{L}$ and a neighboring population $\mathcal{P}$ of related papers or proposals. The three LLM judges receive the task-specific expert rubric and a lineage-structured context packet: the frontier question, the ordered lineage, relevant *Idea Genome* objects, GenomeDiff evidence, neighboring population items, and the candidate proposal. PES then decomposes insertion quality into three dimensions that mirror Darwin's conditions for evolution by natural selection (Darwin, 1859):

- **Heredity** (H): Relative to $\mathcal{L}$, does the proposal inherit and build on the right parent *Idea Genome* objects? This measures source continuity, mechanism preservation, and limitation-delta coherence under the GenomeDiff evidence.
- **Variation** (V): Relative to $\mathcal{P}$, does the proposal introduce meaningful novelty? Cosmetic recombination is penalized; substantive mutation, transfer, or new *Idea Genome* objects are rewarded.
- **Selection** (S): Within the stated lineage population, is the proposal competitively viable? This measures feasibility, fit to the research environment, and whether the proposed work opens productive downstream directions.

For each scientific domain, one domain expert independently annotates task-specific rubric criteria for Heredity, Variation, and Selection. Three LLM judges apply these criteria to each proposal using the lineage and population context $(\mathcal{L}, \mathcal{P})$, returning scores and reasons. Criterion scores are averaged across judges and within each dimension on a 0–100 scale. PES takes the equal-weight mean of H/V/S after dimension-level caps, then applies proposal-level caps for structural or insertion failures (Appendix C.1). Appendix B.2 presents a fusion-control task with its rubric criteria and scoring anchors. We report PES as the primary metric because it measures lineage-conditioned insertion quality rather than standalone appeal. Pairwise ELO (Elo, 1978), computed from judge-panel battles under a Bradley–Terry model (Bradley & Terry, 1952), is a complementary preference diagnostic. ELO measures relative pairwise preference, whereas PES applies the rubric and structural scoring rules to each proposal. PES inter-judge agreement is Krippendorff's $\alpha = 0.74$ (Krippendorff, 2011). Separately, human judges agree with the model-judge panel on 80% of pairwise rankings.

5 EXPERIMENTS

5.1 EVALUATION SETUP

Systems and settings. IG-Exam uses the 42-task, 1,029-instance main-challenge profile after removing calibration diagnostics that reached ceiling performance in pilot runs. IG-Gym covers 30 frontier tasks across 10 domains, yielding 1,260 generated proposals from 14 LLM-based scientists under Question, Library, and Lineage settings. Participants include eight direct LLMs, two GPT-5.5-based research agents, and four CLI harnesses that wrap GPT-5.5 or Claude Opus 4.7 in Codex or Claude Code workflows. This split lets us separate backbone ability, retrieval-heavy agent workflows, and lightweight tool-using scaffolds.

Metrics. IG-Exam reports exact accuracy, so all required fields must be correct simultaneously. IG-Gym reports PES, obtained from three LLM judges applying expert-annotated H/V/S rubrics, and ELO as a pairwise preference diagnostic. The three information settings form a controlled ablation: Question tests parametric ideation, Library tests paper-level context, and Lineage tests ordered *Idea Genome* objects and GenomeDiff records.

5.2 IG-EXAM RESULTS

Table 2: **IG-Bench main leaderboard.** IdeaGene-Exam measures closed-form lineage understanding with exact accuracy (%). IdeaGene-Gym reports expert-rubric-based PES from three LLM judges and pairwise ELO.

LLM-based Scientist	**IdeaGene-Exam**					**IdeaGene-Gym**	
	T1 Abstraction	**T2** Tracing	**T3** Reasoning	**T4** Verification	**Total**	**PES**	**ELO**
Direct LLMs							
GPT-5.5	27.5	25.7	23.3	16.0	23.1	86.5	1213.2
Claude Opus 4.7	28.5	21.9	17.1	14.5	19.3	82.6	1007.2
Qwen3.6-Max-Preview	26.9	22.5	18.8	**17.4**	20.6	79.8	882.7
Gemini-3.1-pro-preview	32.4	24.6	17.8	10.7	20.1	82.1	705.7
Kimi-K2-Thinking	27.9	22.1	19.6	14.8	20.4	81.1	793.9
DeepSeek-V4-Pro	23.9	20.6	18.6	8.5	17.9	82.7	918.8
GLM-5.1	28.5	19.7	17.4	8.5	17.7	80.7	668.0
MiniMax-M2.7	22.9	9.1	10.9	11.6	11.9	70.5	540.8
Research agents							
AI Scientist v2 (GPT 5.5)	28.1	26.9	22.3	15.1	23.0	80.6	1253.3
CoI-Agent (GPT 5.5)	26.9	27.4	22.4	13.4	22.7	80.5	1249.9
CLI harnesses							
Codex (GPT-5.5)	31.8	30.3	23.6	13.7	24.6	**86.7**	**1476.1**
Claude Code (GPT-5.5)	31.5	**37.9**	**25.3**	12.7	**27.3**	86.1	1420.6
Codex (Claude Opus 4.7)	**34.4**	29.7	17.0	10.5	21.8	82.7	900.8
Claude Code (Claude Opus 4.7)	27.9	26.3	21.0	13.8	22.0	83.3	1159.7

Lineage reasoning is compositionally hard. The strongest direct LLM, GPT-5.5, reaches 23.1% exact accuracy; the best harness, GPT-5.5 + Claude Code, reaches 27.3% (Table 2; Appendix Figure 5). Errors are rarely simple misses. A model may identify the parent paper but assign the wrong dynamics label, or infer the driver *Idea Genome* but misclassify its fate. Exact scoring exposes the consistency failures that matter for downstream generation.

Capability-axis breakdown. Across the 14 systems, mean accuracy decreases from 28.5% on T1 Genome Abstraction to 24.6% on T2 Inheritance Tracing, 19.7% on T3 Evolutionary Reasoning, and 12.9% on T4 Lineage Verification. T1 requires reading one *Idea Genome*; T4 requires parent identity, *Idea Genome* compatibility, driver consistency, and evidence validity to hold together. The lower joint accuracy motivates examining these field dependencies alongside task difficulty.

Tool scaffolding helps retrieval more than consistency. CLI harnesses improve most on T2 Inheritance Tracing (GPT-5.5: 25.7% $\rightarrow$ 37.9% with Claude Code), where iterative tool use can retrieve and compare extracted *Idea Genome* objects across papers. The gains shrink on T3 Evolutionary Reasoning and nearly vanish on T4 Lineage Verification. Current scaffolds therefore help information gathering more than compositional consistency checking. Research agents (AI Scientist v2 and CoI-Agent in the main ranking) stay close to direct GPT-5.5 on most axes, suggesting that retrieval-heavy workflows alone do not add lineage-reasoning capability.

5.3 IG-Gym Results

Lineage context separates systems. The gain from Question to Lineage varies by system (Figure 3). GPT-5.5 already performs well under Question-only prompts (PES 85.2) and gains only +2.3 from lineage context. Weaker systems such as Kimi-K2-Thinking gain more (+6.9). The median gain is +4.4, so lineage context helps on average, but its main effect is diagnostic: it separates systems that can use *Idea Genome* structure from those that cannot.

Heredity and structural validity distinguish the groups. Appendix Figure 7 groups proposals by setting and inferred dynamics. Variation and Selection vary less across groups (82.7–84.7 and 79.7–82.2) than Heredity. Question-only Mutation has Heredity 61.9 and PES 69.2, compared with 84.2 and 83.6 for Lineage-setting Hybridization. The former group contains 64 proposals, 22 of which receive proposal-level caps; its reported PES therefore differs from the mean of its displayed H/V/S scores. Within Hybridization, Heredity rises by 5.4 points from Question to Lineage.

Dynamics labels describe the move, not its quality. Question-only prompts produce a broader mixture of dynamics, whereas Library and Lineage concentrate on Hybridization (Appendix Figure 4). A Mutation label does not itself imply invalid inheritance, nor does Hybridization guarantee a coherent proposal. PES assesses the evidence and insertion conditions for each individual proposal.

PES and pairwise preference. PES and ELO produce different system rankings (Spearman $\rho \approx 0.56$, computed from the 14 displayed system scores in Table 2). The two measures summarize rubric-based scores and pairwise preferences, respectively; this comparison alone does not identify the cause of their disagreement.

5.4 Key Findings

Finding 1: Paper-level evidence is not enough. Library context provides more information than Question-only, but unordered paper summaries do not say which *Idea Genome* objects are inherited, mutated, lost, or externally imported. The Lineage setting isolates that added structure. Its gain in lineage-conditioned PES, especially Heredity, shows that GenomeDiff structure carries signal beyond paper summaries.

Finding 2: Verification defines a shared evaluation target. Both IG-Exam verification and IG-Gym heredity require valid parents, compatible mechanisms, and supporting evidence. They assess this target through different tasks. A reliable quantitative understanding–generation association remains to be established using matched system outputs and sensitivity analysis.

Finding 3: Tool scaffolds yield uneven lineage gains. Among five GPT-5.5-backbone systems (Appendix Figure 8), research agents nearly match the direct LLM on understanding but score lower on PES. CLI harnesses improve T2 Inheritance Tracing and maintain T1 Genome Abstraction, with lit-

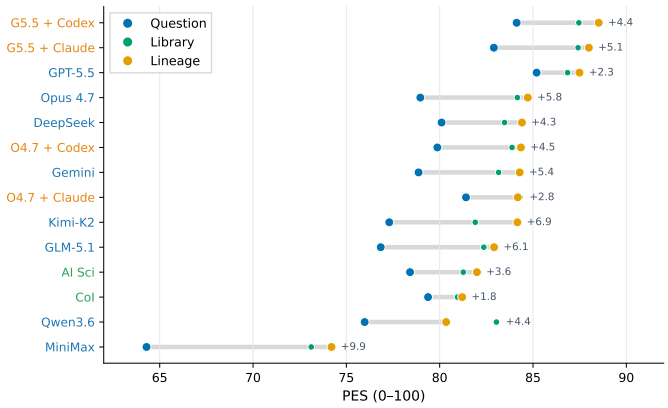

Figure 3: **PES across information settings.** Question, Library, and Lineage scores for all 14 systems. Gains from Question to Lineage vary across systems.

tle T3 gain and lower T4 accuracy. These profiles suggest that tool scaffolds aid retrieval-dependent tasks without resolving compositional lineage reasoning.

Finding 4: Systems overproduce plausibility relative to lineage coherence. Generated ideas often sound useful before they preserve the exact parent mechanism or limitation-delta relation. The PES dynamics decomposition (Appendix Figure 7) shows why: Variation stays high across the displayed groups, while Heredity and structural caps distinguish their PES scores. Ungrounded combination can be high-variation but low-heredity; genome-grounded recombination can be novel and coherent at the same time.

6 Conclusion

IG-Bench reframes automated-research evaluation around scientific lineage. The central question is not only whether a system can write plausible research text, but whether it can identify, verify, and extend the inherited mechanisms that make a proposal a coherent descendant of prior work. The IdeaGene framework—typed *Idea Genome* objects, GenomeDiff records, and evolutionary dynamics—provides the representation layer for making that judgment precise and auditable.

The results make that distinction visible. Even the best system reaches only 27.3% exact accuracy on IG-Exam; tool scaffolds help T2 Inheritance Tracing but barely move T4 Lineage Verification; and structured lineage context separates systems rather than lifting them uniformly. In IG-Gym, expert-rubric-based PES decomposition shows the same plausibility–coherence gap: proposals can sound novel while failing to preserve the parent mechanism or limitation-delta relation needed for coherent descent. These findings point to a concrete design direction: auto-research systems need compositional verification modules, not only better retrieval.

AI Use Statement

We used AI tools to polish the manuscript's language and presentation. After human consistency checks of evolutionary-dynamics annotations, we used LLMs to scale dynamics annotation to larger collections of papers. Three LLM judges applied domain-expert-annotated rubrics for PES scoring; LLM judges also performed pairwise comparisons under the evaluation protocol described in Section 4.3. Human validation and agreement checks are described in Section 4.1 and Appendix D. The authors take responsibility for the manuscript and reported results.

Ethics Statement

This benchmark evaluates scientific lineage reasoning and generated research proposals. Lineage annotations represent evidence-based interpretations of scientific relationships, rather than definitive

judgments of priority or credit. Model-judge scores may reflect evaluator biases and do not establish experimental validity or the safety of a proposed study. Generated proposals, particularly in biology and medicine, require domain-specific scientific and ethical review before implementation.

Reproducibility Statement

Section 4 specifies dataset construction, annotation checks, input settings, and scoring rules. Appendix A lists the task inventory, Appendix B provides evaluation examples, including the PES rubric and scoring anchors in Appendix B.2, and Appendix C describes decoding budgets, response diagnostics, and failure logging. Together, these details define the evaluation procedure and the information needed to interpret reruns. Results from hosted models can vary with provider updates and stochastic generation.

References

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. LitSearch: A retrieval benchmark for scientific literature search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15068–15083, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.840.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'Arcy, et al. Synthesizing scientific literature with retrieval-augmented language models. *Nature*, 650:857–863, 2026. doi: 10.1038/s41586-025-10072-4.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 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)*, pp. 6709–6738, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.342.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of ECCV*, 2020. doi: 10.1007/978-3-030-58452-8_13.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3586–3596, 2019. doi: 10.18653/v1/N19-1361.

Charles Darwin. *On the Origin of Species by Means of Natural Selection*. John Murray, 1859.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019. doi: 10.18653/v1/N19-1423.

Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco, 1978.

Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. Science of science. *Science*, 359 (6379):eaao0185, 2018. doi: 10.1126/science.aao0185.

Alireza Ghafarollahi and Markus J. Buehler. SciAgents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523, 2025. doi: 10.1002/adma.202413523.

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

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *Proceedings of ICLR*, 2024.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406, 2018.

Klaus Krippendorff. Computing krippendorff's alpha-reliability. Technical report, Annenberg School for Communication, University of Pennsylvania, 2011. URL `https://www.asc.upenn.edu/sites/default/files/2021-03/Computing%20Krippendorff%27s%20Alpha-Reliability.pdf`.

Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Deli Zhao, Yu Rong, Tian Feng, and Lidong Bing. Chain of ideas: Revolutionizing research via novel idea development with LLM agents. *arXiv preprint arXiv:2410.13185*, 2024.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. WildBench: Benchmarking LLMs with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023.

Jiaqi Liu, Shi Qiu, Mairui Li, Bingzhou Li, Haonian Ji, Siwei Han, Xinyu Ye, Peng Xia, Zihan Dong, Meng Chen, Congyu Zhang, Letian Zhang, Guiming Chen, Haoqin Tu, Xinyu Yang, Lu Feng, Xujiang Zhao, Haifeng Chen, Jiawei Zhou, Xiao Wang, Weitong Zhang, Hongtu Zhu, Yun Li, Jieru Mei, Hongliang Fei, Jiaheng Zhang, Linjie Li, Linjun Zhang, Yuyin Zhou, Sheng Wang, Caiming Xiong, James Zou, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. AutoResearchClaw: Self-reinforcing autonomous research with human-AI collaboration. *arXiv preprint arXiv:2605.20025*, 2026.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Yao Lu, Yue Dong, and Laurent Charlin. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 8068–8074, 2020. doi: 10.18653/v1/2020.emnlp-main.648.

Grégoire Mialon et al. GAIA: A benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL `https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Joseph Redmon and Ali Farhadi. YOLO9000: Better, faster, stronger. In *Proceedings of CVPR*, 2017.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of CVPR*, 2016. doi: 10.1109/CVPR.2016.91.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP-IJCNLP*, 2019. doi: 10.18653/v1/D19-1410.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015.

Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J. Nay, Kshitij Gupta, and Aran Komatsuzaki. ARB: Advanced reasoning benchmark for large language models. *arXiv preprint arXiv:2307.13692*, 2023.

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. *arXiv preprint arXiv:2501.04227*, 2025.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. Assisting in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6252–6278, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.347.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*, 2024.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5548–5566, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.338.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. SciEval: A multi-level large language model evaluation benchmark for scientific research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19053–19061, 2024. doi: 10.1609/aaai.v38i17.29872.

Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. AI-Researcher: Autonomous scientific innovation. *arXiv preprint arXiv:2505.18705*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 279–299, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.18.

Yiheng Wang et al. SciEvalKit: An open-source evaluation toolkit for scientific general intelligence. *arXiv preprint arXiv:2512.22334*, 2025. doi: 10.48550/arXiv.2512.22334.

Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the Third Workshop on Noisy User-generated Text*, 2017. doi: 10.18653/v1/W17-4413.

Yujun Wu, Dongxu Zhang, Xinchen Li, Jinhang Xu, Yiling Duan, Yumou Liu, Jiabao Pan, Qiyuan Zhu, Xuanhe Zhou, Jingxuan Wei, Siyuan Li, Jintao Chen, Conghui He, and Cheng Tan. InternAtlas: A methodological evolution graph as research infrastructure for AI scientists. *arXiv preprint arXiv:2604.28158*, 2026. doi: 10.48550/arXiv.2604.28158.

Wanghan Xu et al. Probing scientific general intelligence of LLMs with scientist-aligned workflows. *arXiv preprint arXiv:2512.16969*, 2025. doi: 10.48550/arXiv.2512.16969.

Wanghan Xu et al. ResearchClawBench: A benchmark for end-to-end autonomous scientific research. *arXiv preprint arXiv:2606.07591*, 2026. doi: 10.48550/arXiv.2606.07591.

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. *arXiv preprint arXiv:2504.08066*, 2025.

Shunyu Yao et al. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

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. *arXiv preprint arXiv:2511.02238*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2023.

A Main-Challenge Task Inventory

Table 3: Main IG-Exam challenge: 42 task types across 4 capability dimensions, totaling 1,029 instances. Calibration/local-matching diagnostics are retained in the code release but excluded from the paper leaderboard.

Axis	ID	Task Type	N
T1	T1-01	Contribution Role	30
	T1-02	Genome-Field Type	25
	T1-03	Driver-Passenger Roles	20
	T1-04	Lineage Position	25
	T1-05	Cross-Lineage Bridge	25
T2	T2-01	Five-Genome Lineage Reconstruction	25
	T2-02	Six-Genome Lineage Reconstruction	25
	T2-03	Seven-Genome Lineage Reconstruction	25
	T2-04	Lineage Grouping (8)	25
	T2-05	Medium Lineage Grouping (8)	25
	T2-06	Three-Lineage Grouping (9)	25
	T2-07	Limitation-Delta Match	25
	T2-08	Mixed Limitation-Delta Match	25
	T2-09	Chained Limitation-Delta Match	25
	T2-10	Two-Paper Genome-Field Assignment	29
	T2-11	Three-Paper Genome-Field Assignment	29
	T2-12	Genome Alignment	30
T3	T3-01	Single-Step Dynamics	25
	T3-02	Genome-Field Fate	25
	T3-03	Driver Dynamics	25
	T3-04	Shown Genome-Field Fate	22
	T3-05	Driver-Fate Summary	23
	T3-06	Mechanism-Only Dynamics	25
	T3-07	Blind Genome-Field Fate	25
	T3-08	Unlabeled Driver Inference	24
	T3-09	Relation Classification	21
	T3-10	Directional Genome Choice	25
	T3-11	Evolutionary Tempo	30
	T3-12	Evolutionary Pattern	25
	T3-13	Hidden Genome Fate	24
	T3-14	Hybrid Provenance	25
	T3-15	Multi-Hop Genome Tracking	20
	T3-16	Dynamics Boundary	25
	T3-17	Multi-Citation Relation	20
T4	T4-01	Genome Consistency Check	25
	T4-02	Domain Intruder Detection	25
	T4-03	Local Lineage Repair	25
	T4-04	Next-Hop Prediction	24
	T4-05	Parent-Genome Identification	25
	T4-06	Missing-Link Recovery	18
	T4-07	Genome-Bridge Validation	20
	T4-08	Citation Consistency	20

Table 4: Active IG-Gym task inventory for the main run: 30 tasks balanced across 10 scientific domains.

Domain	Task Type	Count
Computer Science	AgentFramework, LLMReasoning, NativeMultimodal	3
Biology	ProteinDesign, SingleCellFoundation, SpatialTranscriptomics	3
Chemistry	DrugDiscovery, MolecularGeneration, Retrosynthesis	3
Climate	AtmosphericChemistry, ClimateSimulation, WeatherForecasting	3
Energy	BatteryOptimization, PerovskiteSolarCell, SmartGridOptimization	3
Materials	BatteryElectrolyte, Catalysis, Discovery	3
Mathematics	AIforCombinatorics, AutomatedTheoremProving, MathReasoning	3
Medicine	DrugRepurposing, MedicalFoundationModels, MedicalImageAnalysis	3
Neuroscience	BrainComputer, Connectomics, VisualCortex	3
Physics	FusionPlasmaControl, NeuralNetworkPotentials, QuantumErrorCorrection	3
Total		**30 active**

B IG-BENCH EXAMPLES

B.1 IG-EXAM

We present one representative example per capability dimension (T1–T4), showing the full question, choices, and gold answer.

Example of T1 · Genome Abstraction (T1-01: Contribution Role)

Question. Given an anonymized genome extracted from a paper, classify its contribution type. Options: `method`, `dataset`, `analysis`, `system`, `theory`.

Claim: "A probabilistic model that accounts for repertoire-specific maturation age can reliably recover the true landscape and length statistics of antibody somatic indels from sequencing data."
Delta: "Provides a dedicated probabilistic method for unbiased learning of antibody indel statistics, compared with prior work relying on biased annotation-based indel calls."
Observation: "Applied to large human heavy-chain datasets, the model finds distinct insertion/deletion hotspots and shows indel lengths are approximately geometrically distributed."

Answer: `method` *(The paper introduces a new probabilistic model — a clear methodological contribution.)*

Example of T2 · Inheritance Tracing (T2-04: Lineage Grouping)

Question. Eight *Idea Genome* objects are presented in shuffled order. Partition them into exactly two lineage groups of four objects each, ordered from earliest to latest.

P: *"Directly optimize a neural architecture policy using RL to assign layer-wise operations."*
Q: *"NIN replaces fully-connected layers with global average pooling and* 1×1 *conv to deepen non-linearity."*
R: *"DARTS relaxes the discrete architecture search space into a continuous one via differentiable mixed-operation weights."*
S: *"Inception-v3 factorizes large convolutions into asymmetric sequences to reduce parameters."*
T: *"EfficientNet jointly scales width, depth, and resolution via compound coefficients."*
U: *"ENAS shares weights across child networks to reduce NAS search cost by orders of magnitude."*
V: *"GoogLeNet stacks Inception modules to achieve high accuracy at low parameter cost."*
W: *"MnasNet uses RL-based NAS with mobile hardware latency as a proxy reward."*

Answer:

Group 1 (NAS-RL lineage): P $\rightarrow$ U $\rightarrow$ R $\rightarrow$ W
Group 2 (Inception lineage): Q $\rightarrow$ V $\rightarrow$ S $\rightarrow$ T

Why: Group 1 shares a lineage of learning architectural decisions (RL search $\rightarrow$ weight sharing $\rightarrow$ differentiable relaxation $\rightarrow$ mobile NAS); Group 2 shares the Inception niche of efficient multi-scale feature extraction.

Example of T3 · Evolutionary Dynamics (T3-01: Single-Step Dynamics)

Question. Given the Niche and Mechanism of predecessor A and successor B, identify the primary evolutionary dynamic.

Predecessor A:
Niche: "Understanding and improving robustness of neural network classifiers to small, intentionally constructed input perturbations that cause misclassification."
Mechanism: "FGSM linearizes the loss at the input point, computes the gradient w.r.t. the

input, takes its element-wise sign, and adds a small ϵ-perturbation under ℓ_∞ constraint to generate an approximate worst-case sample."

Successor B:
Niche: "Understanding *why* deep neural network classifiers remain vulnerable to adversarial perturbations despite strong generalization on natural inputs."
Mechanism: "Uses first-order linear analysis to explain adversarial examples, modeling classifiers as locally linear in high-dimensional input space: a perturbation chosen along the input gradient accumulates across many dimensions to significantly alter class scores."

Choices:

A. Mutation B. Adaptive Radiation C. Hybridization D. Speciation E. Niche Competition

Answer: A. Mutation
Why: B inherits and refines A's first-order gradient-sign mechanism, while narrowing to a theoretical niche (explaining vulnerability) — a local modification, not a structural departure.

Example of T4 · Structural Integrity (T4-06: Missing-Link Recovery)

Question. Paper A and Paper C are given; identify the missing bridge paper B from four candidates.

Paper A: "Balance control for legged robots using ZMP-based model predictive control; focuses on stability under static and slow dynamic gaits."
Paper C: "Sim-to-real RL for agile locomotion; trains a neural network policy end-to-end in simulation and deploys on hardware using domain randomization."

Candidates:
X: "A dataset of human gait trajectories for imitation learning."
Y: "A periodic gait stabilization framework that decouples cycle generation from balance correction, enabling classical controllers to handle dynamic terrain."
Z: "A transformer-based language model for robot task planning."
W: "A convolutional terrain classifier for foothold selection."

Answer: Y
Why: Y bridges A's model-predictive stability focus and C's RL-based locomotion by introducing a periodic gait framework — a structural predecessor that made dynamic, policy-compatible gait generation feasible before the RL era.

B.2 IG-Gym

The following fusion-control example illustrates the expert-annotated PES rubric applied by the three LLM judges. Criterion descriptions are condensed for presentation; the paired-idea examples below illustrate the separate preference evaluation.

IG-Gym: Fusion Plasma Control

Task. Predict-first TCV and reinforcement-learning plasma-control studies show that data-driven control can learn device-specific dynamics, but ITER-scale deployment needs cross-device transfer, uncertainty-aware safety, and simulator-to-shot adaptation. What fusion-control foundation model would combine differentiable plasma simulators, multi-tokamak logs, and constrained RL to produce robust ramp-up, sustainment, and ramp-down policies across machines?

Lineage context. The task packet includes disruption prediction, TCV magnetic control, offline model-based control, tearing-instability avoidance, TORAX differentiable simulation,

HL-3 data-driven dynamics, and TCV rampdown control. Judges inspect which mechanisms the proposal inherits, changes, and makes available to later work.

PES Rubric for Fusion Plasma Control

Heredity: grounding the inherited genome.

1. **Inferred lineage fit.** Does the proposed controller form a coherent next step from the supplied fusion-control works, with a plausible source for each inherited mechanism?
2. **Reused-part support.** Are the claimed simulator, learned dynamics, and policy components supported by the source work? Distinguish transport simulation, disruption prediction, and active control rather than treating them as interchangeable.
3. **Lineage reasoning quality.** Does the proposal explain how an inherited component addresses a specific predecessor limitation, and which assumptions remain valid when transferring between devices or discharge phases?

Variation: specifying the methodological change.

1. **Limitation validity.** Does the proposal target a supported bottleneck, such as device-specific dynamics, simulator-to-shot mismatch, or uncertainty during ramp transitions?
2. **Method specificity.** Does it explain how observations, model states, control actions, and safety constraints interact, beyond naming a simulator and an RL algorithm?
3. **Nontrivial change.** Does it introduce a meaningful change relative to the inherited method, rather than simply combining existing modules or changing the model name?
4. **Evaluation rigor.** Can the proposed comparisons isolate that change? Judges examine baselines, ablations, cross-device tests, and criteria for failure as well as success.

Selection: assessing scientific value and viability.

1. **Frontier importance.** Would addressing the stated bottleneck advance transferable, reliable fusion control?
2. **Non-redundancy.** Does the proposal occupy a distinct methodological niche relative to the supplied population of simulator-based, offline, and instability-aware controllers?
3. **Feasibility.** Are the required diagnostics, training data, actuator interfaces, inference latency, and safety checks consistent with the proposed deployment setting?
4. **Future-lineage potential.** Does the work produce a reusable mechanism, model, interface, or validated finding that could support subsequent control methods or new devices?

Scoring anchors. Each criterion uses the same 0–100 anchors: 0–20, missing, incorrect, fabricated, or unusable; 21–40, weak or largely unsupported; 41–60, plausible but ordinary or partially supported; 61–80, concrete and mostly supported; 81–100, exceptional, precise, and well supported. Judges return criterion scores, a parent-plausibility judgment, an insertion verdict (Accept, Conflict, or Redundant), and a reason. Paper-title matching alone is not required: the inherited mechanism and its support determine lineage fit.

Pairwise comparison examples.

Example of IG-Gym · Pairwise Idea Battle (Physics / Fusion Plasma Control)

Question. Given a frontier research question and two anonymously generated ideas, judge which idea is stronger based on lineage grounding, meaningful variation, and selection value for future research. Options: **A**, **B**, or **Tie**.

Frontier question: Predict-first TCV and RL plasma-control studies show that data-driven control can learn device-specific dynamics, but ITER-scale deployment needs cross-device transfer, uncertainty-aware safety, and simulator-to-shot adaptation. What fusion-control foundation model would combine differentiable plasma simulators, multi-tokamak logs, and constrained RL to produce robust ramp-up, sustainment, and ramp-down policies across machines?

Idea A — Uncertainty-Aware Tokamak Control Foundation Model. Pretrain a transformer/state-space latent dynamics model on multi-tokamak time-series data (DIII-D, TCV, JET, EAST, C-Mod, HL-3 and TORAX rollouts), coupled to differentiable TORAX physics with learned residuals. Train constrained model-based RL policies with Bayesian uncertainty estimates and runtime safety shields (tearing, shape, density, disruption budgets). Before deployment, calibrate via a small number of predict-first target-machine shots updating Bayesian residuals — without retraining the full model.

Idea B — FusionFoundationRL: Cross-Tokamak Differentiable-Constrained Policy. Combine TORAX's differentiable physics engine with a neural state-space model (NSSM) into a hybrid dynamics foundation model trained on multi-tokamak logs. Device-specific parameters (geometry, coil limits) are encoded as context vectors. A constrained RL agent uses cross-machine tearing/disruption predictors as safety bounds, and TORAX's differentiability initializes the policy via gradient-based optimization. Unified ramp-up, sustainment, and ramp-down in a single policy, with simulator-to-shot fine-tuning for adaptation.

Which idea better addresses the frontier? Consider: how well does each inherit and build on prior work, how meaningfully does it advance beyond it, and how much does it open new research directions?

Example of IG-Gym · Battle with Judge Result (Biology / Protein Design)

Frontier question. Design de novo protein binders against therapeutic targets where high in silico affinity does not reliably translate into selective, stable, functional activity in living cells. What generative framework would close the loop between computational design and in vivo biological function?

Idea A — Cell-in-the-Loop Multi-Objective Protein Binder Design. Couple structure-based binder generation (diffusion/flow-matching conditioned on target epitope, off-target panels, and stability constraints) with a learned multi-fidelity fitness model. Candidate designs are scored by a differentiable surrogate ensemble covering binding affinity, target selectivity, folding stability, expression, and cell-based efficacy. A causal latent variable model separates true target engagement from confounders (expression level, trafficking, toxicity). The surrogate is updated iteratively via active learning from barcoded pooled yeast/mammalian display and cellular reporter assays, converting protein binder design into a self-improving closed-loop system optimized for experimentally meaningful activity.

Idea B — Multi-Constraint Latent Diffusion for In-Vivo-Aware Protein Binder Design. Learn a unified latent representation from multi-modal data (structures, binding assays, biophysical stability, in vivo functional readouts) using a multi-modal VAE, then run a conditional diffusion model in this latent space guided by target structure embeddings and a learned biological constraint manifold. A differentiable in vivo efficacy surrogate provides gradient feedback during generation; selectivity is enforced by contrastive learning against off-target proteomes; stability is modeled via physics-informed neural networks. A Pareto-frontier multi-objective reward balances competing constraints, aiming to raise experimental validation rates from $\sim$5–10% to $>$30%.

Judge result: Idea A wins.
Judge reason: *"Idea A more cleanly extends the frontier lineage by preserving structure-first generative design while adding assay-grounded active learning and causal components that directly target the known in silico-to-function gap, with a more feasible validation plan. Idea B is ambitious but relies on weakly transferable in vivo surrogates and expensive mouse-model validation."*

C Evaluation Details

All IG-Exam runs write the decoding budget into the top-level JSON metadata as `max_output_tokens`. Per-instance logs also store input/output token counts when returned by

the API, response length, extraction diagnostics, and an error class. Empty responses are therefore separated into API failures, missing-format outputs, and token-budget exhaustion rather than being silently counted as ordinary content errors.

This detail matters for reasoning-heavy deployments. In pilot GPT-5.5 runs, a 4,096-token completion cap sometimes produced empty visible answers while the API reported the full output budget as consumed. We therefore use a 16,384-token cap for GPT-5.5 main-challenge reruns and keep the cap in the released logs so that scores remain auditable across providers and reruns. The paper leaderboard reports the same exact-match metric for all models; token-budget metadata is used only for diagnosis and reproducibility, not for changing an instance score.

C.1 PES Aggregation and Structural Caps

The scoring implementation first averages the available valid judge scores for each rubric criterion, then equally averages the criteria within H, V, and S (three, four, and four criteria, respectively). Categorical parent-plausibility and insertion judgments use majority voting. Dimension caps are applied before the equal-weight H/V/S mean; proposal caps are applied afterward. Every applicable cap is an upper bound, so the lowest applicable bound prevails. System and group scores average these final proposal scores. Consequently, a displayed group PES need not equal the mean of its displayed H/V/S values.

Table 5: Original PES caps. A dash denotes no additional cap at that stage.

Condition	Dimension cap	PES cap
Parent/source judged implausible	H: 40	65
Claimed inherited parts have zero source matches[a]	H: 35	50
Insertion judged redundant	S: 45	60
Nearest mechanism similarity > 0.85	S: 45	55
Inferred dynamics is Out-of-lineage	H: 30	55
Invalid JSON or genome completeness < 0.5	—	40
Insertion judged conflicting	—	45
Hybridization with fewer than two inferred sources	—	60
Speciation with nearest mechanism similarity > 0.85	—	65
All model judges unusable	—	40

[a] Applies when at least one inherited part is claimed and the inferred dynamics is not Niche Competition. These are implementation checks, not additional expert annotations. If a criterion has no usable judge values, the original implementation assigns 50 before applying caps; missing categorical votes do not activate a categorical cap. These fallback rules are part of the archived scoring protocol.

For example, the Question–Mutation group contains 64 proposals. Its mean dimension scores are H=61.94, V=82.73, and S=79.70 after dimension caps; their mean is 74.79 before proposal caps (74.77 if computed from the one-decimal labels in the figure). Proposal caps apply to 22 items, yielding the reported mean PES of 69.20. The raw, intermediate, and capped scores are retained separately in the evaluation records.

D Human Agreement

We recruited 50 graduate annotators (master's and doctoral students across computer science, biology, physics, materials science, and other disciplines) to validate the benchmark at three stages: GenomeDiff relation labels and role-type assignments during construction (disagreements adjudicated; ambiguous items excluded), IG-Exam item difficulty through stratified human solving to confirm low accuracy reflects compositional challenge rather than label noise, and IG-Gym ELO battles where human judges reached 80% agreement with the strongest model-judge panel on pairwise preference comparisons. PES scoring uses the same lineage packet format for model and human audit: judges see the frontier question, relevant *Idea Genome* objects, GenomeDiff evidence, and population neighborhood before assigning H/V/S. The 80% concordance concerns pairwise preferences. It is distinct from rubric annotation, for which one expert in each domain independently specifies the task-level PES criteria, and from the three LLM judges who assign the reported rubric-based scores.

E Additional IG-Gym Analysis

The main system comparison uses the aggregate PES summaries in Table 2. Per-system dimension and domain plots are omitted because their available source records could not be reconciled with those summaries. The setting–dynamics analysis below is retained as an aggregate diagnostic.

E.1 Generated Dynamics Distribution

This figure reports the frequency of generated dynamics labels; Figure 7 relates those labels to PES. It shows that Question-only prompts produce a broader mix of evolutionary moves, while Library and Lineage prompts concentrate on Hybridization; by itself, however, this distribution cannot tell whether the recombination is genetically coherent.

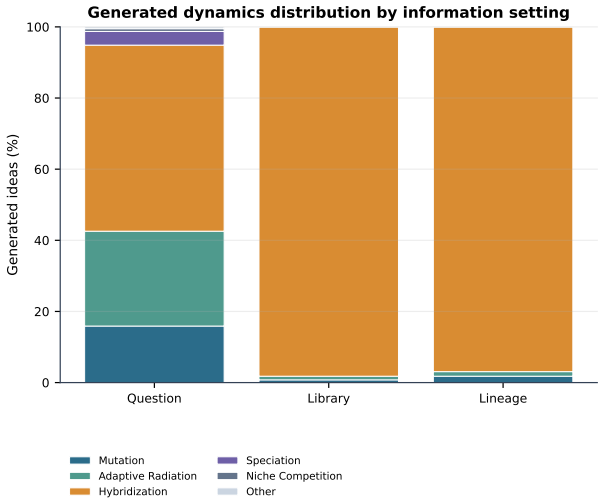

Figure 4: **Generated dynamics distribution by information setting.** Stacked proportions of post-hoc dynamics labels for generated IG-Gym proposals. Figure 7 provides the corresponding quality breakdown: the same dynamics label can have different Heredity and PES depending on whether lineage evidence is available.

F Additional IG-Exam Analysis

F.1 IG-Exam Main-Leaderboard Heatmap

This heatmap reports the main leaderboard's exact accuracy across T1–T4. It complements the backbone-controlled radar profile in Figure 8.

F.2 Error Analysis: Failure Flow

This Sankey view aggregates wrong answers by capability tier, failed field family, and error class. It is meant to localize the bottleneck: most failures originate in evolutionary reasoning and verification, then propagate through driver and dynamics fields before becoming exact-match errors.

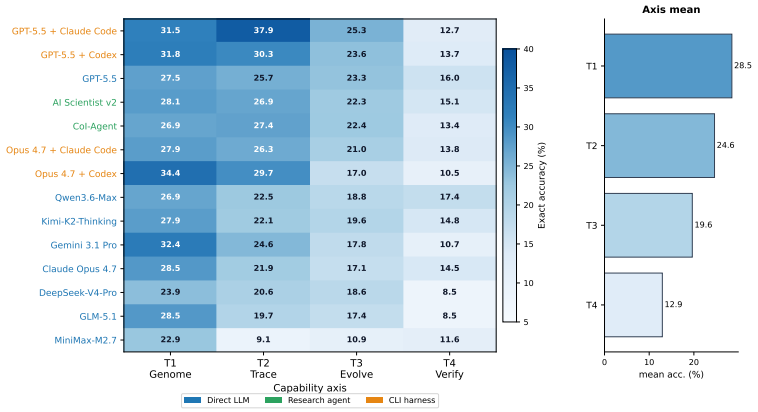

Figure 5: **IG-Exam exact-accuracy heatmap.** Exact accuracy across T1–T4 for all evaluated systems, with verification remaining the hardest axis.

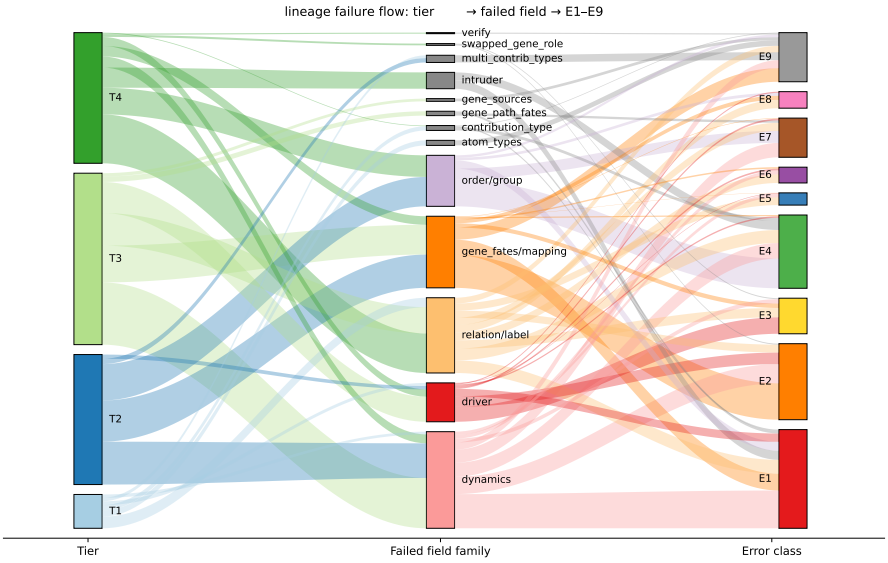

Figure 6: **IG-Exam failure flow.** Sankey diagram tracing errors from capability tier (T1–T4) through failed field family to error class (E1–E9). T3 and T4 errors dominate in volume. The thickest flows connect T3/T4 to dynamics and driver fields, which then fan out to E1 (dynamics misclassification) and E2 (driver misidentification). Genome-fate and relation errors (E3–E4) are the next largest classes, confirming that compositional field interactions—not isolated label failures—drive the accuracy bottleneck.

G Framework and Task Taxonomy

Tables 6 and 7 detail the evolutionary criteria and the four IG-Exam capability axes introduced in the main text.

Table 6: Evolutionary dynamics as operational GenomeDiff criteria. Each dynamics type corresponds to a specific alignment pattern between two *Idea Genome* objects.

Dynamics	Lineage?	GenomeDiff criterion	Canonical example
Mutation	Yes	Driver mechanism is inherited or locally mutated; niche remains same or nearby.	YOLO (Redmon et al., 2016) → YOLOv2 (Redmon & Farhadi, 2017): one-stage detection persists; anchors, batch normalization, and multi-scale training repair local limits.
Adaptive Radiation	Yes	Driver mechanism persists, but moves into a new task, domain, or evaluation ecology.	Transformer (Vaswani et al., 2017) → ViT (Dosovitskiy et al., 2021): self-attention is inherited, but moved from token sequences to image patches.
Hybridization	Yes	Successor imports driver objects from two or more distinct lineages.	CLIP-style visual encoder (Radford et al., 2021) + instruction-tuned LLM → LLaVA (Liu et al., 2023): visual alignment and chat behavior are imported from two lineages.
Speciation	Yes	Same or nearby niche, but the predecessor's driver mechanism is replaced by a new lineage-forming mechanism.	Faster R-CNN (Ren et al., 2015) → DETR (Carion et al., 2020): CNN region proposals are replaced by Transformer set prediction for detection.
Niche Competition	No	Same ecology or problem niche, but no driver inheritance.	Faster R-CNN (Ren et al., 2015) vs. YOLO (Redmon et al., 2016): both solve object detection, but region proposals and one-stage regression are competing mechanisms.
Isolation	No	Neither shared ecology nor driver inheritance.	BERT (Devlin et al., 2019) vs. YOLO (Redmon et al., 2016): language understanding and object detection share neither task ecology nor driver mechanism.

Table 7: IG-Exam main challenge: 42 task types, 1,029 instances, four capability axes.

Capability	Tasks	Inst.	Representative tasks
T1 Genome Abstraction	5	125	Identify field type, driver *Idea Genome*, contribution role, or lineage position
T2 Inheritance Tracing	12	313	Reconstruct ordered lineages, align inherited *Idea Genome* objects, match limitations to deltas
T3 Evolutionary Reasoning	17	409	Infer dynamics, driver, object fates, hybrid provenance, or multi-hop changes
T4 Lineage Verification	8	182	Detect intruders, wrong steps, missing links, citation conflicts, or parent mismatch

H PES Dynamics and Capability Profiles

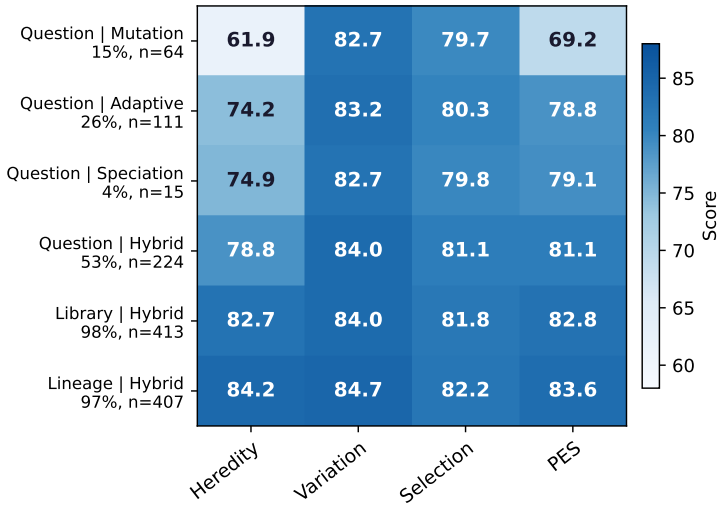

Figure 7: **PES decomposition by dynamics mode.** Groups are sorted by PES; only groups with $n \geq 15$ are shown. Question produces four dynamics modes, while Library and Lineage concentrate on Hybridization ($\geq$97%). H/V/S include dimension caps; PES additionally includes proposal caps (Appendix C.1). Heredity ranges from 61.9 (Question|Mutation) to 84.2 (Lineage|Hybrid).

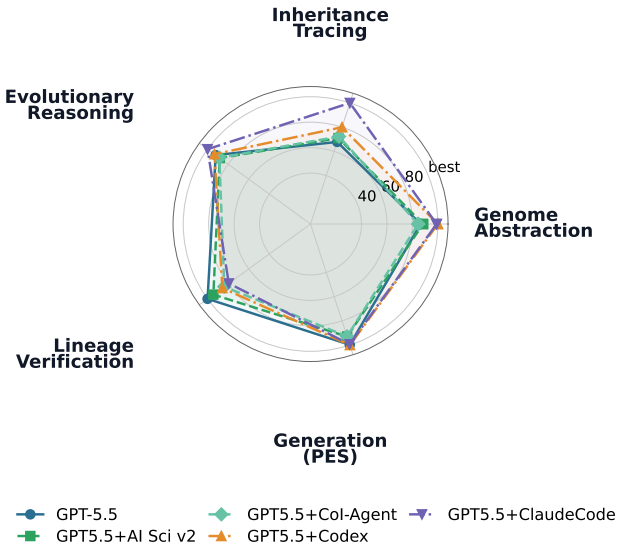

Figure 8: **Backbone-controlled capability profile.** Five GPT-5.5-backbone systems are compared across five capability dimensions: the direct LLM, AI Scientist v2, CoI-Agent, Codex, and Claude Code. Research agents nearly overlap with the direct LLM, while CLI harnesses redistribute strengths rather than uniformly improving every capability.