# How Does Knowledge Organization Shape LLM-Based Scientific Ideation

**Anonymous authors**
Paper under double-blind review

## Abstract

AI scientists powered by large language models (LLMs) draw on domain knowledge to propose and develop research ideas. Existing systems organize this knowledge differently, as retrieved papers, literature chains, or concept graphs. Comparing complete systems cannot reveal which organizational choice shapes the resulting ideas, because such comparisons change the delivered evidence and its arrangement at once. We introduce a controlled evaluation framework that separates these two factors: knowledge access (what evidence reaches the model under a fixed budget) and knowledge presentation (how the same evidence is arranged). Across two studies with three generators and two judges, we run more than 68,000 idea generations and 87,000 model evaluations. On ResearchBench, access methods change the evidence the model receives and, where that evidence differs most, the benchmark-matching score. Enlarging the candidate pool dilutes the relevant evidence admitted and does not improve matching. With the evidence held fixed, linear and relation-grouped presentations are practically equivalent in all four generator–judge pairings. In multi-step development built on CHIMERA, a recombination instruction raises the rate of explicit citation of knowledge from other research branches by 52–62 percentage points over refinement, whereas two access methods delivering similar evidence show no consistent difference. These results indicate that knowledge organization shapes the resulting ideas mainly through which evidence reaches the model.

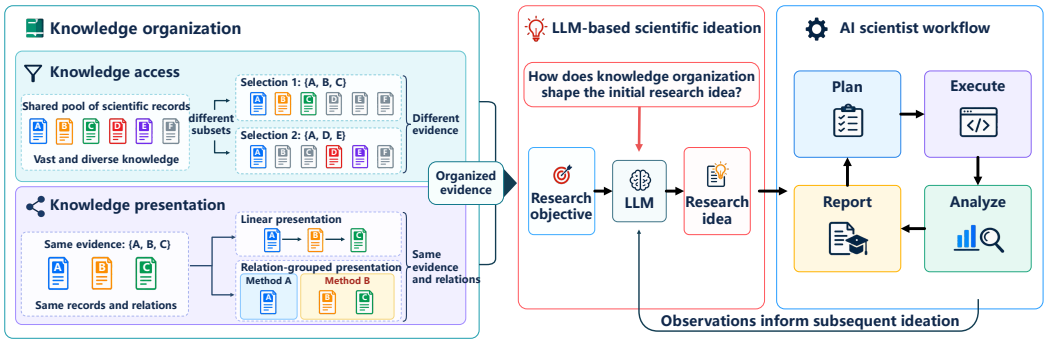


Figure 1: Knowledge organization in an AI scientist. External knowledge informs the initial research idea, which shapes planning, execution, analysis, and reporting. Organization has two dimensions: access selects which knowledge from a shared pool reaches the model, so selected evidence can differ; presentation arranges the same knowledge, for example linearly or grouped by relation.

## 1 Introduction

Large language models (LLMs) increasingly support scientific research, from planning and executing chemical experiments (Boiko et al., 2023) to iteratively developing hypotheses (Gottweis et al., 2026). AI scientist systems connect ideation, experimental execution, and scientific writing within a single workflow (Lu et al., 2026). These workflows can be viewed as dynamic sequential decision



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.

processes: proposed ideas guide which experiments are conducted, and their results inform subsequent revisions. The initial research idea therefore anchors later decisions about what to investigate and how, so designing an AI scientist requires knowing what shapes that idea.

Recent work exposes gaps in LLM-based ideation and in its evaluation. LLM-generated ideas can cover a narrower and systematically shifted range of research opportunities and methodological approaches than human reference papers (Chen et al., 2026). A large scientist-in-the-loop study reports highly similar ideas among non-reasoning models and underrepresented null hypotheses even when reasoning broadens the explored hypothesis space (Bao et al., 2026). Theoretical work distinguishes fitting observed distributions and maintaining internal consistency from discovering mechanisms and achieving externally verified improvement (Zenil et al., 2026). Evaluation can obscure such gaps: LLM judges favor generated research questions in comparisons where domain experts prefer the author-anchored references (Sinhahajari et al., 2026). These studies motivate identifying which controllable components shape an AI scientist's ideas and measuring their effects with outcomes of explicit scope.

We study a design choice that can be varied with the generator held fixed: the organization of external scientific knowledge (Figure 1). Literature-based discovery has long shown that connecting separately published findings can reveal new knowledge (Swanson, 1986). Contemporary ideation systems retrieve related papers, arrange literature into chains, or reason over graphs of scientific concepts and relations (Baek et al., 2025; Li et al., 2025; Ghafarollahi & Buehler, 2025). These choices determine which prior findings and connections are available when an idea is developed. We therefore ask: *given the same LLM, knowledge source, and research task under comparable budgets, can different ways of organizing external scientific knowledge change the ideas that the model proposes and develops?*

Answering this question requires distinguishing what knowledge reaches the model from how it is presented. Arrangement alone can affect model behavior, as studies of evidence position and meaning-preserving prompt formats show (Liu et al., 2024; Sclar et al., 2024). Systems drawing on the same corpus can also supply different papers, relation statements, and amounts of evidence. In multi-step ideation, the current proposal shapes subsequent retrieval, so evidence can diverge as ideas develop. Final outputs alone cannot show which organizational choice caused a difference.

We introduce a controlled evaluation framework that separates *knowledge access* from *knowledge presentation* (Figure 2). Access determines which scientific records and relation statements enter the model's context under a budget, and presentation determines how that evidence is arranged. Access comparisons share the candidate pool and budget but let the selected evidence differ, whereas presentation comparisons hold both evidence content and relational information fixed and vary only their arrangement. We apply this framework to ResearchBench (Liu et al., 2026) with 29,760 idea generations from three generators, scored by two judges in four generator–judge pairings. We then extend the access comparison to a dynamic protocol built on CHIMERA (Sternlicht & Hope, 2026), in which proposals develop over four steps of refinement or recombination. The two studies measure reference-hypothesis matching and explicit cross-branch citation, respectively.

Our contributions are:

- **A controlled framework for knowledge organization.** We separate changes in the evidence that reaches a fixed generator from changes in its arrangement, report the delivered evidence alongside the resulting ideas, and extend the access comparison from single-step generation to iterative refinement and recombination.

- **Static effects follow the delivered evidence.** Across four generator–judge pairings, access methods change the evidence received and, where it differs most, the matching score (up to 0.25 points); larger pools dilute relevant evidence without improving matching; and grouping the same evidence by relation leaves matching practically unchanged.

- **Instructions steer knowledge use in multi-step development.** A recombination instruction raises explicit cross-branch citation by 52–62 percentage points over refinement. Access methods that deliver similar evidence show no consistent difference, in line with the static findings.

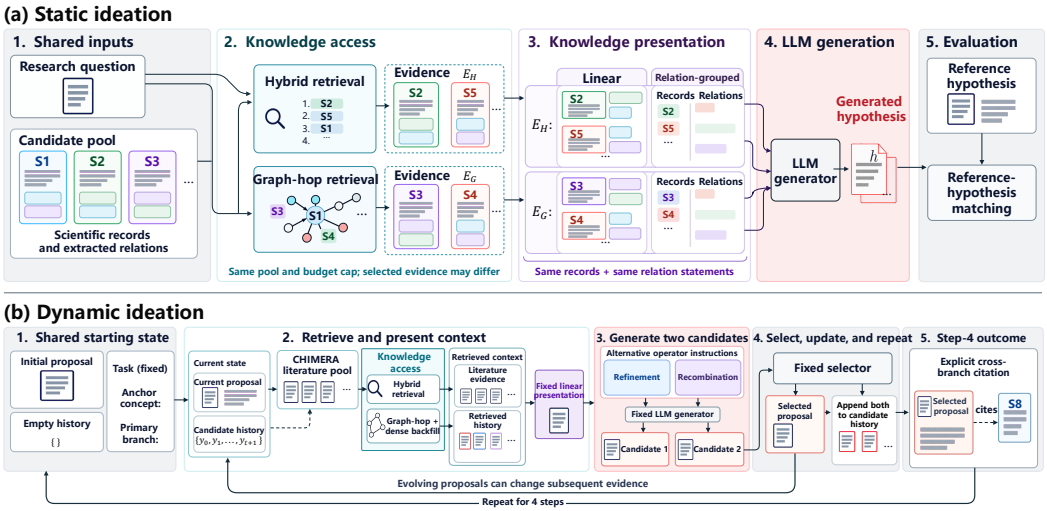


Figure 2: Study design. (a) Static study: access selects records and their relation statements from a shared candidate pool under a token budget, so hybrid retrieval and graph-hop access can deliver different evidence; presentation arranges the same evidence linearly or grouped by relation. A fixed generator writes one hypothesis per condition, scored against the ResearchBench reference hypothesis. (b) Dynamic study: systems share an initial proposal and an empty history. At each of four steps, they retrieve literature and history, generate two candidates under a refinement or recombination instruction, and select one with a fixed selector, so later evidence can diverge. The outcome is explicit citation of a record from another branch at step four.

## 2 Related Work

**LLM-based scientific ideation and evaluation.** LLM-based research systems support experimental execution (Boiko et al., 2023), end-to-end research (Lu et al., 2026), and iterative hypothesis development (Gottweis et al., 2026). Within ideation, SciMON retrieves scientific inspirations and revises ideas for novelty (Wang et al., 2024), and MOOSE-Chem retrieves inspiration papers to compose hypotheses (Yang et al., 2025). Human studies distinguish perceived idea quality from realized outcomes: LLM ideas can receive higher novelty but slightly lower feasibility ratings than human ideas (Si et al., 2025), and these ratings do not fully carry over after execution (Si et al., 2026). We instead fix the generator, vary its evidence, and measure reference matching and explicit citation.

**Knowledge organization for scientific ideation.** Literature-based discovery connects findings across publications (Swanson, 1986), and studies of scientific impact link high-impact work to combinations of conventional and atypical knowledge (Uzzi et al., 2013). Ideation systems organize knowledge through citation graphs and cross-paper entities (Baek et al., 2025), literature chains (Li et al., 2025), and ontological graph paths (Ghafarollahi & Buehler, 2025). CHIMERA records recombinations of scientific ideas (Sternlicht & Hope, 2026), Scideator supports recombining purpose, mechanism, and evaluation facets (Radensky et al., 2026), and Intern-Atlas compares graph-grounded ideation with retrieval from the same paper library (Wu et al., 2026). More generally, retrieval-augmented generation supplies external passages (Lewis et al., 2020), HippoRAG uses graph retrieval with Personalized PageRank (Gutiérrez et al., 2024), and GraphRAG uses community summaries (Edge et al., 2025). Graph-based and standard retrieval show task-dependent differences (Han et al., 2026), and dialogue-memory research separates representation, organization, maintenance, and retrieval (Hu et al., 2026). Beyond access, models can be sensitive to irrelevant context (Shi et al., 2023), evidence position (Liu et al., 2024), and meaning-preserving prompt formats (Sclar et al., 2024). These findings motivate separating the evidence delivered from its arrangement. Because a shared corpus does not fix the selected evidence, we compare access methods under shared candidate pools and budgets, and presentations with the selected records and relation statements held fixed.

# 3 Problem Formulation and Controlled Comparisons

## 3.1 Knowledge organization in scientific ideation

Scientific ideation is a generative step within a sequential research process: an idea specifies a direction to pursue, and later information can motivate its refinement or recombination. We study this step at two resolutions, a single hypothesis and a proposal developed over several steps, to trace how external knowledge enters idea formation.

A task $i$ consists of a research question $q_i$ and a *candidate knowledge pool* $K_i$ of *scientific records* drawn from a source corpus. In the static study, a record is the title and abstract of a candidate paper together with the *relation statements* that an LLM extracted from that abstract. Each statement has a type (a mechanism addresses a problem, the paper uses a mechanism, a mechanism extends another, or two findings contrast), quotes its evidence verbatim, and keeps its source identifier. In the dynamic study, a record is a CHIMERA recombination instance, described by its source abstract, context, and the two entities involved, with relation statements extracted in the same way. An access method $\mathcal{A}$ selects an *evidence set* $E_i$ under a token budget $B$, a presentation $\mathcal{P}$ renders it, and a generator $G$ produces a research idea $y_i$ (a hypothesis in the static study and a proposal in the dynamic study):

$$E_i = \mathcal{A}(K_i, q_i; B), \qquad y_i \sim G\big(q_i, \mathcal{P}(E_i)\big). \tag{1}$$

We vary the context supplied to fixed models and leave their parametric knowledge unchanged. Eq. (1) makes the comparison boundary explicit: conditions that share $K_i$ and $B$ need not share $E_i$.

## 3.2 Separating access from presentation

*Hybrid retrieval* indexes record texts and relation statements, takes the top 25 items by lexical (BM25) and the top 25 by dense similarity to the research question, maps them to records, and reranks the union with a cross-encoder. Each admitted record carries all its relation statements. *Graph-hop access* builds a graph of records and the entities named in their statements, merging identically named entities across records. It starts from the three records most similar to the question, traverses up to two hops, and returns the records reached together with the statements on the traversal. Both methods admit whole records in their own order and stop when the next record would exceed the budget or when they run out of candidates. They can therefore differ in which records are admitted, how much of the budget they use, and which statements accompany a record.

For each access result, the *linear* presentation lists each record followed by its relation statements, whereas the *relation-grouped* presentation lists the record texts first and then the statements grouped by relation type. Admission uses the longer rendering, so both presentations contain the same record texts, identifiers, and multiset of relation statements. The comparison therefore asks whether grouping adds value once relations are present, not whether relations do. With $\bar{S}_{a,p}$ the mean benchmark-matching score of access $a$ and presentation $p$ at a fixed pool size and budget, the presentation and access contrasts are

$$R = \tfrac{1}{2}\textstyle\sum_a \big(\bar{S}_{a,\mathrm{lin}} - \bar{S}_{a,\mathrm{grp}}\big), \qquad A = \tfrac{1}{2}\textstyle\sum_p \big(\bar{S}_{\mathrm{hyb},p} - \bar{S}_{\mathrm{gh},p}\big), \tag{2}$$

computed within task, so positive $R$ favors linear presentation and positive $A$ favors hybrid retrieval. The pool-size contrast compares 75-record pools with 25-record pools that keep all designated inspiration papers and a stratified sample of the other candidates, so enlarging the pool adds only non-inspiration papers.

Auxiliary controls test specific explanations. Supplying only the task's inspiration papers, compared with the question alone, tests whether the score responds to informative evidence. *Sequential admission* fills the same budget with records in the benchmark's list order, which separates ranking from the amount of text. *Dense backfill* adds records by dense similarity when graph-hop access stops before filling the budget. Shuffling records tests their order within the linear presentation, and a relation-removal control deletes the relation statements from fixed hybrid-retrieval records.

## 3.3 Scientific ideation as a sequential process

The dynamic study varies access under one linear presentation. Each task is built around an *anchor*, the source concept of a held-out CHIMERA recombination, whose domain defines the task's

primary *branch*. All systems share the same short initial proposal $y_{i,0}$, and the history $H_{i,0}$ starts empty. At step $t \in \{1, \ldots, 4\}$, the state $s_{i,t-1} = (y_{i,t-1}, H_{i,t-1})$ contains the current proposal and accumulated candidate history. Access uses the task description and the current proposal as its query and returns literature evidence $E_{i,t}$ and retrieved history $U_{i,t} \subseteq H_{i,t-1}$ under separate budgets. For a fixed operator instruction $o$, idea generation and selection are distinct steps:

$$\begin{aligned} C_{i,t} &= \{y_{i,t}^{(1)}, y_{i,t}^{(2)}\}, \quad y_{i,t}^{(j)} \sim G(q_i, y_{i,t-1}, \mathcal{P}(E_{i,t}), U_{i,t}; o)\,,\ j \in \{1,2\}, \\ y_{i,t} &= \mathcal{S}(C_{i,t}; \xi_{i,t}), \qquad H_{i,t} = \mathrm{append}(H_{i,t-1}, C_{i,t}). \end{aligned} \tag{3}$$

The fixed selector $\mathcal{S}$ sees only candidate texts, and seeded randomness $\xi_{i,t}$ resolves ties. Because the selected idea shapes the next query, evidence can diverge over four steps even from a shared initial state. The protocol is fixed rather than learned; Appendix A gives the full recurrence and controls.

*Refinement* improves the proposal along the anchor's branch, borrowing auxiliary mechanisms only from records of that branch. *Recombination* integrates at least one mechanism from another branch, cites its record, and explains its connection to the anchor. Motivated by the static access bottleneck (Section 5.2), we compare hybrid retrieval with *graph-hop plus dense backfill*. Each displayed record keeps its identifier and branch label. Independent-sampling and history-off controls examine refinement without intermediate selection and without retrieved history, respectively (Appendix A). The outcome is *explicit cross-branch citation*: whether the step-four proposal cites at least one displayed literature record from another branch. Invalid identifiers, undisplayed records, and history entries do not count. This outcome needs no judge and records observable knowledge use.

## 4 Experimental Setup

**Tasks.** The static study uses 240 ResearchBench tasks (Liu et al., 2026), the first 20 in each of 12 disciplines, drawn from 192 source papers (DOI clusters; Appendix G). Each task has a research question and 75 candidate papers, of which 2.3 on average are designated inspiration papers. Qwen-Flash extracted 17,111 relation statements from the 4,534 distinct candidate abstracts in a single pass. The generator writes one concrete hypothesis grounded in the evidence and cites record identifiers. Evidence budgets of 2,000 or 6,000 tokens (2k and 6k) are counted with the Qwen-27B tokenizer for every generator. Crossing two pool sizes, two budgets, two access methods, two presentations, and two sampling seeds gives 7,680 generations per generator and 23,040 for the three generators. Single-seed auxiliary controls add 2,880 generations for Qwen-27B and 1,920 for each other generator, for a total of 29,760 generations and 39,360 judge evaluations (39,232 valid).

The dynamic study uses 150 tasks drawn from the CHIMERA recombination-prediction test split, stratified by relation type and cross-domain status (Sternlicht & Hope, 2026). Literature pools are nested samples of 2,000 (small) and 8,000 (large) knowledge-base records, restricted per task to earlier years and excluding the anchor's paper and records mentioning the held-out recombination partner. Branches are the knowledge base's fine-grained domain labels. Twenty tasks were used for protocol development, so primary analyses use the other 130; history-off runs use a 50-task subset containing 39 primary tasks. Literature and history may use at most 4,000 and 1,536 tokens, respectively, within a 6,000-token prompt. Each generation system produces 1,600 endpoints from 12,800 candidate generations, so the dynamic study comprises 38,400 candidate generations, 38,400 selector evaluations, and 9,600 judge evaluations of 4,800 endpoints (Appendices B and D).

**Models and roles.** We abbreviate the three deployed models after their serving identifiers as Qwen-27B (`qwen3.8-27b`), Qwen-Flash (`qwen3.8-flash`), and DeepSeek-Flash (`deepseek-v4.1-flash`), all run with reasoning disabled (Appendix H lists architectures, weights, and licenses). Qwen-27B is the base generator of both studies, and Qwen-27B judged by Qwen-Flash is the base static pairing; the other pairings and generators replicate its comparisons. In the static study, the replication generators received exactly the base generator's input messages, so all generators saw identical evidence. The static pairings change one model at a time: Qwen-27B's outputs are scored by both judges to isolate the judge, and each other generator is compared with Qwen-27B under one fixed judge. In the dynamic study, all three generators run with the same DeepSeek-Flash selector, and both judges score all 4,800 endpoints.

**Outcomes and inference.** The static outcome is the official ResearchBench matching score (0–5), which rates how many key methodological points of the reference hypothesis a generated hypothesis

covers and whether it uses them similarly. The task is the unit of analysis. Seed scores are averaged within task, each contrast uses the tasks with all required scores (missing scores are never imputed), and intervals come from a bootstrap over DOI clusters with 10,000 replicates. The base study's analysis plan, fixed before data collection, specified eight difference contrasts with Holm correction: access, presentation, and their interaction at 75 records and 2k, the changes of these three contrasts from 25 to 75 records, and the changes of access and interaction from 2k to 6k. It also specified two equivalence comparisons with Bonferroni correction, access at 75 records and 2k and presentation at 75 records and 6k. Each uses two one-sided tests (Schuirmann, 1987) against a $\pm 0.25$-point margin (5% of the scale), a practical threshold set before any condition-level result was read. The replication pairings follow their own plan, fixed before those runs, with twelve difference and six equivalence comparisons, each Bonferroni-corrected. All other intervals are unadjusted 95% intervals and are descriptive; Appendix D records when each analysis decision was made.

Explicit cross-branch citation is the dynamic study's primary outcome, fixed with its analysis plan before formal collection. It replaced the originally planned judge-rated proposal quality after the candidate judges missed a pre-set agreement threshold with human annotations on development items (Section 6). Contrasts are paired over the 130 primary tasks with paired score intervals (Tango, 1998). The plan has two primary contrasts, access under recombination and under refinement with the large pool at step four (Bonferroni, 97.5% intervals, exact McNemar tests), and ten secondary contrasts, reported at 99.5% (Appendix E). The replication with two further generators has a family of eight contrasts (99.375%), in which changes relative to the base generator use a task-level bootstrap. Steps, candidates, and repeated judgments are never treated as independent observations.

# 5 Results

## 5.1 Larger candidate pools dilute the evidence that reaches the model

In an exploratory paired analysis, enlarging the pool from 25 to 75 records lowered the mean matching score for all four pairings at both budgets (Figure 3a; Appendix C). Averaged over access methods and presentations, the paired difference ranges from $-0.05$ to $-0.13$ points (2–5% of the 25-record mean). One of the eight 95% intervals, for the base pairing at 2k, excludes zero ($-0.13$ $[-0.26, -0.02]$). In the base study's pre-specified family, the access, presentation, and interaction contrasts showed no corrected change between the two pool sizes.

The admission columns of Table 1 suggest why. All added candidates are non-inspiration papers that compete with the inspiration papers for the same budget, so the fraction of a task's inspiration papers admitted by hybrid retrieval falls from 0.60 to 0.43 at 2k and from 0.97 to 0.73 at 6k.

## 5.2 Access shapes the evidence received, and matching tracks it

The access methods differed most in how much evidence they delivered (Table 1). At 6k, graph-hop access exhausted its traversal on every task after about three records and used 27–28% of the budget, whereas hybrid retrieval admitted about eleven records and used 94–95%. Matching scores track this gap (Figure 3b): at 75 records and 6k, hybrid retrieval exceeds graph-hop access by 0.11–0.25 points (4–9% of the graph-hop mean) in all four pairings, with every unadjusted 95% interval above zero. At 2k, where the two methods admit similar numbers of records, the differences shrink to at most 0.14 points. There, the pre-specified comparison at 75 records meets the equivalence criterion for both pairings judged by DeepSeek-Flash. For the base pairing ($+0.135$, simultaneous interval $[-0.055, 0.326]$) and for DeepSeek-Flash outputs ($+0.124$, $[-0.017, 0.272]$), it is neither significant nor equivalent within the margin.

The control rows link matching to evidence exposure. For Qwen-27B and Qwen-Flash, which usually answer even without evidence, supplying the inspiration papers raises the score over the question alone by 0.48 to 1.34 points. DeepSeek-Flash declines to answer without evidence and is scored 0, so its gain of 3.27 includes this abstention. Sequential admission fills the budget as fully as hybrid retrieval but admits few inspiration papers, and both retrieval methods exceed it in all 32 single-seed comparisons, by up to 1.05 points (26 intervals exclude zero). Filling the budget that graph-hop access leaves unused at 75 records and 6k with dense backfill raises admitted records from about 3 to 12 and inspiration recall from 0.39 to 0.71. Matching improves by 0.20 to 0.41 points (0.37 for the base pairing), with every interval above zero, and backfill ends 0.10 to 0.15 points above hybrid

Table 1: Knowledge access on ResearchBench: evidence admitted per task (records, fraction of inspiration papers, % of budget; identical for all generators) and mean matching score (0–5) per generator/judge pairing (Q27 = Qwen-27B, QF = Qwen-Flash, DS = DeepSeek-Flash). Hybrid and graph-hop scores average presentations and seeds over the 223–240 tasks of the paired contrast, so their difference is the access contrast $A$ ($^*$: 95% interval excludes zero). Controls: one seed, linear; DS declines every question-only task (scored 0).

| | | Evidence admitted | | | Matching score, generator/judge | | | |
|---|---|---|---|---|---|---|---|---|
| Setting | Condition | Records | Recall | Budget % | Q27/QF | Q27/DS | QF/DS | DS/QF |
| 25/2k | Sequential | 3.6 | 0.14 | 85 | 2.14 | 2.36 | 2.08 | 2.20 |
| | Hybrid | 3.2 | 0.60 | 85 | 2.94 | 2.86 | 2.88 | 3.07 |
| | Graph-hop | 2.8 | 0.54 | 74 | 2.81 | 2.86 | 2.87 | 3.07 |
| | *Hybrid − graph-hop* | | | | +0.13$^*$ | 0.00 | +0.01 | 0.00 |
| 75/2k | Question only | 0 | – | – | 1.69 | 2.00 | 2.61 | 0.00 |
| | Inspiration papers | 2.2 | 0.96 | 62 | 3.03 | 3.05 | 3.09 | 3.27 |
| | Sequential | 3.8 | 0.04 | 85 | 1.99 | 2.17 | 1.86 | 1.97 |
| | Hybrid | 3.2 | 0.43 | 85 | 2.83 | 2.80 | 2.86 | 3.08 |
| | Graph-hop | 3.0 | 0.37 | 74 | 2.70 | 2.76 | 2.79 | 2.96 |
| | *Hybrid − graph-hop* | | | | +0.14 | +0.04 | +0.07 | +0.12$^*$ |
| 25/6k | Sequential | 11.9 | 0.49 | 95 | 2.75 | 2.79 | 2.77 | 3.05 |
| | Hybrid | 10.8 | 0.97 | 94 | 3.03 | 2.95 | 3.06 | 3.25 |
| | Graph-hop | 3.1 | 0.57 | 27 | 2.85 | 2.89 | 2.94 | 3.13 |
| | Graph-hop + backfill | 11.6 | 0.97 | 95 | 3.04 | 3.06 | 3.05 | 3.30 |
| | *Hybrid − graph-hop* | | | | +0.18$^*$ | +0.06 | +0.12$^*$ | +0.12$^*$ |
| 75/6k | Sequential | 11.9 | 0.16 | 95 | 2.38 | 2.51 | 2.63 | 2.81 |
| | Hybrid | 11.4 | 0.73 | 95 | 2.97 | 2.90 | 3.00 | 3.23 |
| | Graph-hop | 3.3 | 0.39 | 28 | 2.72 | 2.79 | 2.82 | 2.98 |
| | Graph-hop + backfill | 12.1 | 0.71 | 95 | 3.02 | 3.04 | 3.10 | 3.37 |
| | *Hybrid − graph-hop* | | | | +0.25$^*$ | +0.11$^*$ | +0.18$^*$ | +0.25$^*$ |

retrieval (one interval excludes zero). At 25 records, the gain is 0.13 to 0.19. These controls point to an evidence-access bottleneck behind the graph-hop deficit; backfill removes it by changing both how many and which records reach the model.

### 5.3 Arranging fixed evidence by relation leaves matching unchanged

With the admitted records and relation statements held fixed, presentation leaves matching practically unchanged (Figure 3c; Table 4 in Appendix C). All 16 presentation contrasts, across four pool–budget settings and four pairings, lie within $\pm 0.07$ points, at most 2.3% of the grouped mean. At the pre-specified setting of 75 records and 6k, every pairing meets the equivalence criterion ($\pm 0.25$ points) under its planned correction.

The result holds as each model is replaced in turn. Rescoring the same Qwen-27B hypotheses with the DeepSeek-Flash judge, or replacing the generator under either judge, leaves grouping without a detectable effect, and none of the twelve replication difference contrasts excludes zero after correction. The two judges agree on identical hypotheses with ICC 0.68–0.71 and reach ICCs of 0.74 and 0.84 against a human reference on 40 Qwen-27B outputs (Appendix F).

Two further controls produced no detectable change: shuffling record order within the linear presentation (four contrasts) and removing the relation statements from fixed hybrid-retrieval records, which shortened the evidence by about a third (Table 7).

### 5.4 Instructions steer knowledge use in multi-step development

The operator instruction separates citation behavior in every condition (Figure 4, Table 2). At step four, recombination proposals cite a record from another branch in 67.7–80.8% of primary tasks, and refinement proposals in 6.2–22.3%. The paired difference ranges from 52 to 62 percentage points across the twelve combinations of generator, access, and pool, with every 95% interval above 43 points. It widens in every condition from 30–49 points at step one. Because each operator in-

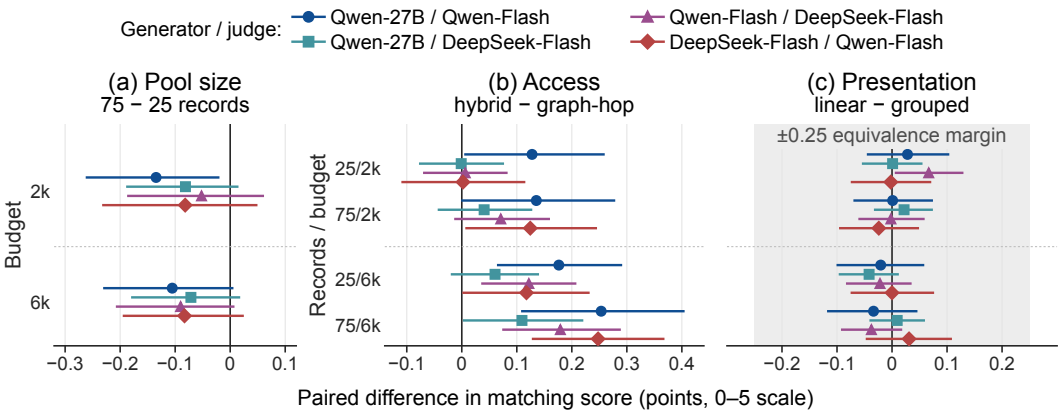


Figure 3: Static comparisons paired within task, in points (0–5 scale) with unadjusted 95% DOI-cluster intervals, for the pairings of Table 1. (a) Averaged over access and presentation (exploratory). (b) Averaged over presentations. (c) Averaged over access; shaded: $\pm 0.25$ equivalence margin.

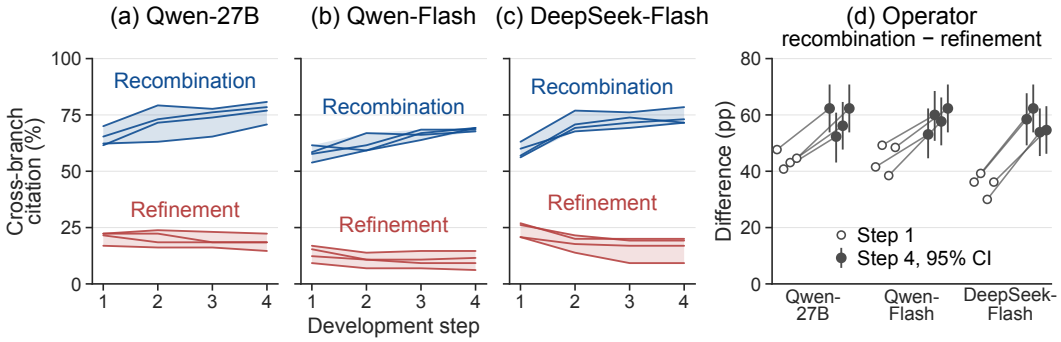


Figure 4: Explicit cross-branch citation of the selected proposal (% of 130 primary tasks) for the three generators with the DeepSeek-Flash selector. (a–c) Rate at development steps 1–4; each line is one access × pool condition and the band spans the four. (d) Recombination minus refinement on the same tasks (percentage points) at step 1 (open) and step 4 (filled, descriptive 95% task-bootstrap interval), one pair per condition.

structs the behavior being measured, this separation is the dynamic counterpart of the static positive controls: it shows that the citation measure registers changes in how knowledge is used.

On this measure, the two access methods show no consistent difference (Table 8), in line with the static findings. Both admitted about seven literature records per step, 40–50% of them from other branches, so the exposure gap of the static study does not arise. No pre-specified large-pool interval (half-widths up to about 12 points) excludes zero, for the base generator or in the eight-contrast replication family, and none of the six small-pool 95% intervals does. Under refinement with the large pool, the estimate is $-3.8$ points for Qwen-27B but $+7.7$ for DeepSeek-Flash, whose unadjusted 95% interval $[1.9, 14.4]$ excludes zero. This sign reversal points to generator-dependent variation rather than a consistent access effect (Appendix E). The refinement controls likewise detect no effect of iteration on this behavior. For the base generator, step-four refinement differs from independent best-of-eight sampling by $-3.8$ and $-3.1$ points, withholding the history changes the rate by $0.0$ and $+7.7$ points on 39 tasks, and none of the ten secondary contrasts is rejected (Appendix E).

## 6 Analysis and Discussion

**Where organization acts.** Across both studies, the organizational differences that changed outcomes were differences in what reached the model. Static pool-size, access, and backfill results moved with the amount and relevance of admitted evidence, whereas arranging fixed content by re-

Table 2: Explicit cross-branch citation of the selected proposal (% of 130 primary tasks) at steps 1 and 4; Difference is recombination minus refinement on the same tasks, access, and pool (percentage points, computed before rounding). The difference widens from step 1 to step 4 in all twelve conditions; every step-4 difference has a 95% interval above 43 points (Table 8).

| | | Refinement | | Recombination | | Difference | |
|---|---|---|---|---|---|---|---|
| Pool | Access | Step 1 | Step 4 | Step 1 | Step 4 | Step 1 | Step 4 |
| *Generator: Qwen-27B* | | | | | | | |
| Small | Hybrid | 22.3 | 18.5 | 70.0 | 80.8 | +47.7 | **+62.3** |
| | Graph-hop + backfill | 22.3 | 22.3 | 65.4 | 78.5 | +43.1 | **+56.2** |
| Large | Hybrid | 21.5 | 18.5 | 62.3 | 70.8 | +40.8 | **+52.3** |
| | Graph-hop + backfill | 16.9 | 14.6 | 61.5 | 76.9 | +44.6 | **+62.3** |
| *Generator: Qwen-Flash* | | | | | | | |
| Small | Hybrid | 16.9 | 14.6 | 58.5 | 67.7 | +41.5 | **+53.1** |
| | Graph-hop + backfill | 15.4 | 11.5 | 53.8 | 69.2 | +38.5 | **+57.7** |
| Large | Hybrid | 12.3 | 9.2 | 61.5 | 69.2 | +49.2 | **+60.0** |
| | Graph-hop + backfill | 9.2 | 6.2 | 57.7 | 68.5 | +48.5 | **+62.3** |
| *Generator: DeepSeek-Flash* | | | | | | | |
| Small | Hybrid | 26.9 | 20.0 | 63.1 | 78.5 | +36.2 | **+58.5** |
| | Graph-hop + backfill | 26.2 | 19.2 | 56.2 | 73.1 | +30.0 | **+53.8** |
| Large | Hybrid | 20.8 | 9.2 | 60.0 | 71.5 | +39.2 | **+62.3** |
| | Graph-hop + backfill | 20.8 | 16.9 | 56.9 | 71.5 | +36.2 | **+54.6** |

lation left scores unchanged. In the dynamic study, access methods admitting similar evidence produced similar citation behavior. This pattern holds across two corpora, access implementations, and outcome measures, and points to organization acting mainly through which knowledge is received. In the graph-hop condition, lower scores at the larger budget coincide with an access bottleneck: the traversal over a sparse relation graph stopped after about three records, and the gap closed once dense backfill filled the unused budget. Access methods should therefore be assessed by what they deliver under the actual budget, and arrangements tested with their content held fixed. Comparisons of graph-based and text-based systems should report the realized evidence (records admitted, budget used, and, where possible, recall of known relevant material) alongside pool and graph size.

**Outcome measures and scope.** The matching score measures agreement with a reference hypothesis and responds strongly to evidence (Section 5.2); novelty, feasibility, and scientific value lie outside its scope. For developed proposals, we measure judge-free, directly observable citation because automatic quality judges missed a pre-set agreement threshold with human annotations (Spearman 0.18 and 0.30 against 0.4) and agreed weakly with each other (sensitivity analysis in Appendix F). Our evidence covers three generators, two judges, abstract-level records, one relation-grouping scheme, and four development steps with a fixed selector (analysis timeline in Appendix D). Richer cross-record relations, other knowledge structures, and long-horizon outcomes remain to be tested.

## 7 Conclusion

Comparisons that separate knowledge access from presentation indicate that knowledge organization shapes LLM-based scientific ideation mainly through the evidence it delivers. Access methods changed what was admitted and, where admission differed most, the matching score. Larger pools diluted the relevant evidence without raising scores, and grouping the same records and relation statements by relation type yielded practically equivalent scores in all four generator–judge pairings. In multi-step development, a recombination instruction raised cross-branch citation by 52–62 percentage points over refinement; access methods delivering similar evidence showed no consistent difference. For AI scientists, organizational choices are best judged by what they deliver to the model, using outcomes that match the claim being made.

### AI USE STATEMENT

Language models are part of the experimental system: they extract relation statements, generate hypotheses and proposals, select candidates, and serve as automatic judges. The research idea was conceived by the authors. We used generative AI tools to assist with code development, to conduct broad information gathering and literature discovery, including searching for relevant and related research papers and other domain-specific information, and to polish the language and improve the readability of the manuscript. The AI-assisted code underwent human review, and information collected with the assistance of generative AI was independently examined and verified by the authors before being used in the research. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

All datasets and model weights used in this work are publicly available. We provide an anonymized code repository for data preparation, experiment execution, evaluation, and analysis, with the repository link included in the supplementary material. The repository contains the configurations and implementation details needed to reproduce the main experimental pipeline. Exact model outputs can vary across reruns because generation is sampled and the serving backends honor seeds only on a best-effort basis; Appendices D, G, and H specify the protocols, datasets, and models.

## REFERENCES

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. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.342. URL https://aclanthology.org/2025.naacl-long.342/.

Honglin Bao, Siyang Wu, Xiao Liu, Sida Li, Shiyun Cao, and James A. Evans. Contemporary ai lacks the imagination to diverge or negate in science, 2026. URL https://arxiv.org/abs/2606.08251.

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

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Ziyu Chen, Yilun Zhao, and Arman Cohan. Measuring the gap between human and llm research ideas, 2026. URL https://arxiv.org/abs/2607.01233.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2025. URL https://arxiv.org/abs/2404.16130.

Alireza Ghafarollahi and Markus J. Buehler. Sciagents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22): 2413523, 2025. doi: https://doi.org/10.1002/adma.202413523. URL https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/adma.202413523.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, and Dan Popovici. Accelerating scientific discovery with co-scientist. *Nature*, pp. 1–3, 2026.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 59532–59569. Curran Associates, Inc., 2024. doi: 10.52202/079017-1902.

Haoyu Han, Li Ma, Yu Wang, Harry Shomer, Kai Guo, Yongjia Lei, Zhisheng Qi, Zhigang Hua, Bo Long, Hui Liu, Charu Aggarwal, and Jiliang Tang. Rag vs. graphrag: A systematic evaluation and key insights. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 8966–8977, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817575. URL https://doi.org/10.1145/3770855.3817575.

Sen Hu, Yuxiang Wei, Jiaxin Ran, Xueran Han, Zhiyuan Yao, Huacan Wang, Ronghao Chen, and Lei Zou. Does memory need graphs? a unified framework and empirical analysis for long-term dialog memory. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26758–26782, San Diego, California, United States, 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1232. URL https://aclanthology.org/2026.acl-long.1232/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Yu Rong, Deli Zhao, Tian Feng, and Lidong Bing. Chain of ideas: Revolutionizing research via novel idea development with LLM agents. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 8971–9004, Suzhou, China, 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.477. URL https://aclanthology.org/2025.findings-emnlp.477/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. ResearchBench: Benchmarking LLMs in scientific discovery via inspiration-based task decomposition. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13187–13207, San Diego, California, United States, 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.644. URL https://aclanthology.org/2026.findings-acl.644/.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of ai research. *Nature*, 651(8107):914–919, 2026.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S. Weld. Scideator: Human-llm compound system for scientific ideation through facet recombination and novelty evaluation. In *Proceedings of the ACM Conference on AI and Agentic Systems*, CAIS '26, pp. 348–374, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400724152. doi: 10.1145/3786335.3813161. URL https://doi.org/10.1145/3786335.3813161.

Donald J Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, 15:657–680, 1987.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models'sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pp. 25055–25083, 2024.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 23–29 Jul 2023.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. In *International Conference on Learning Representations*, volume 2025, pp. 94003–94092, 2025.

Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. The ideation-execution gap: Execution outcomes of llm-generated versus human research ideas. In *International Conference on Learning Representations*, volume 2026, pp. 74195–74240, 2026.

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

Noy Sternlicht and Tom Hope. CHIMERA: A knowledge base of scientific idea recombinations for research analysis and ideation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1871–1905, San Diego, California, United States, 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.85.

Don R. Swanson. Undiscovered public knowledge. *The Library Quarterly: Information, Community, Policy*, 56(2):103–118, 1986. ISSN 00242519, 1549652X.

T Tango. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in medicine*, 17(8):891–908, 1998.

Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013. doi: 10.1126/science.1240474.

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. URL https://aclanthology.org/2024.acl-long.18/.

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, 2026. URL https://arxiv.org/abs/2604.28158.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. Moose-chem: Large language models for rediscovering unseen chemistry scientific hypotheses. In *International Conference on Learning Representations*, volume 2025, pp. 33251–33277, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/51fd9a7d1706023cb9f8210cc6ac357c-Paper-Conference.pdf.

Hector Zenil, Abicumaran Uthamacumaran, and Luan Ozelim. Large language models as shannon lossy compressors not solomonoff induction estimators: The singularity is not near without symbolic model synthesis, 2026. URL https://arxiv.org/abs/2601.05280.

# A SEQUENTIAL IDEATION PROCESS

This appendix states the dynamic protocol of Section 3.3 as a recurrence. The protocol is fixed: no component is trained, no reward is observed, and no proposed experiment is executed.

**Objects.** Task $i$ provides a task description $q_i$ (the anchor concept, its primary branch $b_i$, and the recombination context), a task-specific literature pool $L_i$ (Appendix D), and a shared initial proposal $y_{i,0}$, written once by Qwen-Flash from the task description without evidence. A system is a tuple $(\mathcal{A}, \mathcal{P}, G, \mathcal{S}, o)$ of access method, presentation, generator, selector, and operator instruction $o \in \{\text{refine}, \text{recombine}\}$. The presentation is linear in every system and shows each literature record with its identifier and branch label; history entries carry their own identifiers.

**Recurrence.** With $H_{i,0} = \emptyset$, each step $t = 1, \ldots, T$ with $T = 4$ performs

$$(E_{i,t}, U_{i,t}) = \mathcal{A}\big(L_i, H_{i,t-1};\ q_i, y_{i,t-1};\ B_{\text{lit}}, B_{\text{hist}}\big), \tag{4}$$

$$y_{i,t}^{(j)} \sim G\big(q_i,\ y_{i,t-1},\ \mathcal{P}(E_{i,t}),\ \mathcal{P}(U_{i,t});\ o\big), \qquad j = 1, 2, \tag{5}$$

$$y_{i,t} = \mathcal{S}\big(\{y_{i,t}^{(1)}, y_{i,t}^{(2)}\};\ \xi_{i,t}\big), \qquad H_{i,t} = H_{i,t-1} \cup \{y_{i,t}^{(1)}, y_{i,t}^{(2)}\}. \tag{6}$$

Access retrieves literature first under $B_{\text{lit}} = 4{,}000$ tokens and then history under $B_{\text{hist}} = 1{,}536$ tokens, within a 6,000-token prompt; at $t = 1$ the history is empty. Hybrid retrieval queries separate literature and history indexes, and graph-hop access starts from three literature and three history anchors and backfills literature only. Eq. (3) in the main text is this recurrence with $\mathcal{P}(U_{i,t})$ written as $U_{i,t}$.

**Idea conception.** Eq. (5) is the step at which an idea is formed. The generator receives the task, the current proposal, the admitted literature, and the retrieved history, and returns a proposal of at most 250 words with motivation, method, and expected-result sections that keeps the anchor central and cites the records it uses by identifier. The operator constrains which knowledge may enter: refinement admits auxiliary mechanisms only from records of the primary branch, whereas recombination requires at least one mechanism from a record of another branch, its citation, and an explanation of how it connects to the anchor. Access thus determines which knowledge is available at conception, the operator how it is to be used, and the selector which conceived idea is carried forward.

**Selection and divergence.** The selector rates each candidate from 0 to 10 for scientific soundness, novelty, and specificity. It sees only the candidate text, neither the evidence nor the operator instruction, so it cannot favor a condition. The two candidates of a step come from the same prompt and often receive the same score; such ties (61–81% of selections) are broken by a seeded hash $\xi_{i,t}$ of task and step, which does not depend on the condition, so the tie rule is condition-blind as well. Because $y_{i,t-1}$ enters the query in Eq. (4), two systems that start from the same $y_{i,0}$ receive different $E_{i,1}$ when their access methods differ, and their proposals, queries, and histories can diverge at every later step. This divergence is what the dynamic comparison is designed to admit.

**Outcome.** Let $D_{i,t} = \bigcup_{t' \le t} E_{i,t'}$ be the literature records displayed up to step $t$, and let $\text{other}_i(r) = 1$ when neither of record $r$'s two domains is the primary branch $b_i$. The step-$t$ outcome is

$$Z_{i,t} = \mathbf{1}\big[\exists\, r \in \text{cite}(y_{i,t}) \cap D_{i,t} :\ \text{other}_i(r) = 1\big], \tag{7}$$

where $\text{cite}(\cdot)$ extracts the cited identifiers; history entries, pool records never displayed, and unresolved identifiers do not count. The primary outcome is $Z_{i,4}$. Table 2 reports $Z_{i,1}$ and $Z_{i,4}$, and Figure 7 all four steps.

**Controls as modified recurrences.** Independent sampling replaces the four steps by one: $E_{i,1}$ is retrieved once from $y_{i,0}$, eight candidates are drawn from the same prompt, and $\mathcal{S}$ selects one. This matches the eight candidate generations of a trajectory while removing intermediate selection and retrieval. History-off sets $U_{i,t} = \emptyset$ and removes history from access, while $y_{i,t-1}$ is still passed to the next step.

**Relation to the static study.** The static study is the one-step case of the same scheme: $T = 1$, no current proposal or history, the research question as query, one candidate, and no selection, which reduces Eqs. (4)–(6) to Eq. (1). The static study also varies the presentation, which the dynamic study holds fixed.

## B COUNTS

Table 3 separates generations, scores, and endpoints. A second judge applied to existing outputs adds scores without adding generations; the 23,040 static generations are the three generators' main matrices. Missing scores are judge responses without a parseable score after one retry, nearly all because the response reached the judge's output limit. They are kept missing and never imputed, and each contrast uses the tasks for which all required scores exist. For the base generator, the 75-record, 2k cells of 120 tasks (960 generations) come from a pilot stage run under the same configuration and are reused unchanged. The static analysis plan, including its contrasts and equivalence margin, had been fixed before any condition-level result of this pilot was read, so the plan did not depend on how these cells scored; the pilot's results were examined before the main protocol was finalized (Appendix D gives the timeline, including the dense-backfill diagnostic added at that point). In the dynamic study, 560 candidate generations and selections of the base generator on development tasks were reused from the development pilot; they fall outside the 130 primary tasks.

Table 3: Generations, selections, and scores. Static counts are per generator–judge pairing; dynamic counts are per generation system.

*Static study (ResearchBench, 240 tasks)*

| Set | Generator | Judge | Generations | Valid scores | Missing |
|---|---|---|---|---|---|
| Main matrix | Qwen-27B | Qwen-Flash | 7,680 | 7,629 | 51 |
| | Qwen-27B (same outputs) | DeepSeek-Flash | – | 7,680 | 0 |
| | Qwen-Flash | DeepSeek-Flash | 7,680 | 7,680 | 0 |
| | DeepSeek-Flash | Qwen-Flash | 7,680 | 7,630 | 50 |
| Auxiliary controls | Qwen-27B | Qwen-Flash | 2,400 | 2,388 | 12 |
| | Qwen-27B (same outputs) | DeepSeek-Flash | – | 1,920 | 0 |
| | Qwen-Flash | DeepSeek-Flash | 1,920 | 1,920 | 0 |
| | DeepSeek-Flash | Qwen-Flash | 1,920 | 1,910 | 10 |
| Relation removal | Qwen-27B | Qwen-Flash | 480 | 475 | 5 |
| Total | | | 29,760 | 39,232 | 128 |

*Dynamic study (CHIMERA-based, per generation system)*

| Condition | Tasks | Steps × candidates | Candidate generations | Endpoints |
|---|---|---|---|---|
| Four-step trajectories (2 operators × 2 access × 2 pools) | $8 \times 150$ | $4 \times 2$ | 9,600 | 1,200 |
| Independent best-of-eight (2 access, large pool) | $2 \times 150$ | $1 \times 8$ | 2,400 | 300 |
| History-off refinement (2 access, large pool) | $2 \times 50$ | $4 \times 2$ | 800 | 100 |
| Per system | | | 12,800 | 1,600 |
| Three systems | | | 38,400 | 4,800 |

Each candidate generation is followed by one selector call, so each system also made 12,800 selections, which yield 5,500 selected step outputs and 1,600 endpoints. Both dynamic judges scored all 4,800 endpoints (9,600 scores, none missing). Relation extraction was run once, before all experiments: Qwen-Flash extracted 17,111 relation statements from the 4,534 ResearchBench records and 33,561 from the 8,000 CHIMERA records. The 150 initial proposals shared by all dynamic systems were generated once by Qwen-Flash from the anchor, its branch, and its context, without evidence.

Static generation requests carried the sampling seed, which the serving backends honor on a best-effort basis; dynamic generation requests carried no seed, and the seed only drives the selector's tie rule. The Qwen-Flash deployment reported the same serving fingerprint in the base and replication rounds, whereas the DeepSeek-Flash deployment reported none. Token budgets are counted with the Qwen-27B tokenizer for all generators. Qwen-Flash tokenizes the static inputs identically, whereas DeepSeek-Flash's own tokenizer counts 6.7% fewer tokens for the same messages in the static study and 5.5% fewer in the dynamic study; admission is unaffected because it always uses the reference count.

## C Additional Static Results

Figure 5 shows all access ($A$), presentation ($R$), and interaction ($I$) contrasts from the four-pairing analysis with unadjusted 95% intervals; $I$ is the presentation effect under hybrid retrieval minus that under graph-hop access. One of the 16 presentation intervals excludes zero (Qwen-Flash/DeepSeek-Flash at 25/2k: $+0.067$, $[0.005, 0.130]$). These intervals are descriptive. For the base pairing, they are computed from the same scores as the base study's analysis but with a different bootstrap seed, so they can differ from its report in the third decimal.

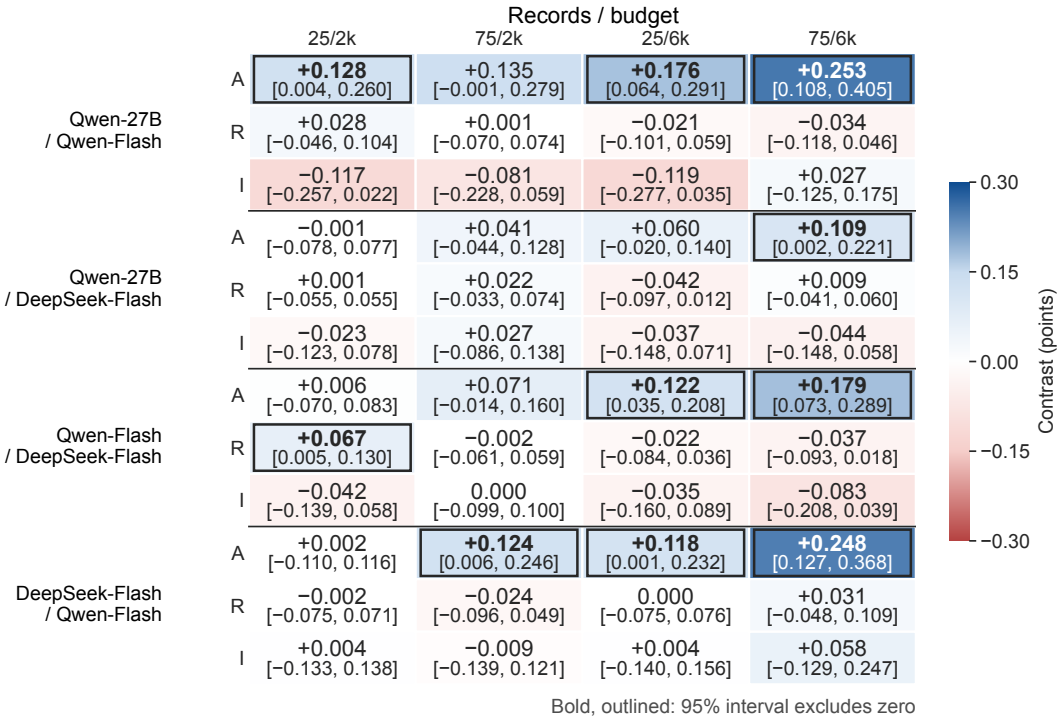


Figure 5: All static contrasts in points, with unadjusted 95% DOI-cluster bootstrap intervals. Shading follows the estimate; bold, outlined cells mark intervals that exclude zero. Tasks per contrast: Qwen-27B / Qwen-Flash 223–231; Qwen-27B / DeepSeek-Flash 240; Qwen-Flash / DeepSeek-Flash 240; DeepSeek-Flash / Qwen-Flash 226–234.

**Pre-specified families.** Figure 6 shows the corrected families. In the base study's family, every Holm-adjusted $p$ is at least 0.452 and every simultaneous interval contains zero. The simultaneous intervals for the access contrast and the two budget changes extend to 0.3–0.4 points. In the replication's difference family, no 99.583% interval excludes zero. Of its six equivalence comparisons, reported in Sections 5.2 and 5.3, five lie within $\pm 0.25$; the exception is access for DeepSeek-Flash outputs ($[-0.017, 0.272]$).

**Presentation.** Table 4 gives the presentation contrasts of Figure 3c with the underlying means, their size relative to the grouped mean, and the pre-specified equivalence intervals.

**Pool size.** Table 5 gives the paired pool-size contrast by access method. It is exploratory and follows the conventions of the pre-specified analyses (tasks weighted equally, seeds averaged within task, complete cases per contrast, 10,000 DOI-cluster bootstrap replicates). All 24 estimates are negative, and three of their 95% intervals exclude zero, all for the base pairing.

**Auxiliary controls.** Table 6 lists the auxiliary contrasts for the four pairings, and Table 7 the shuffled-order and relation-removal controls, which were run for the base pairing only. Without evidence, DeepSeek-Flash declined to propose a hypothesis on all 240 question-only tasks, and the

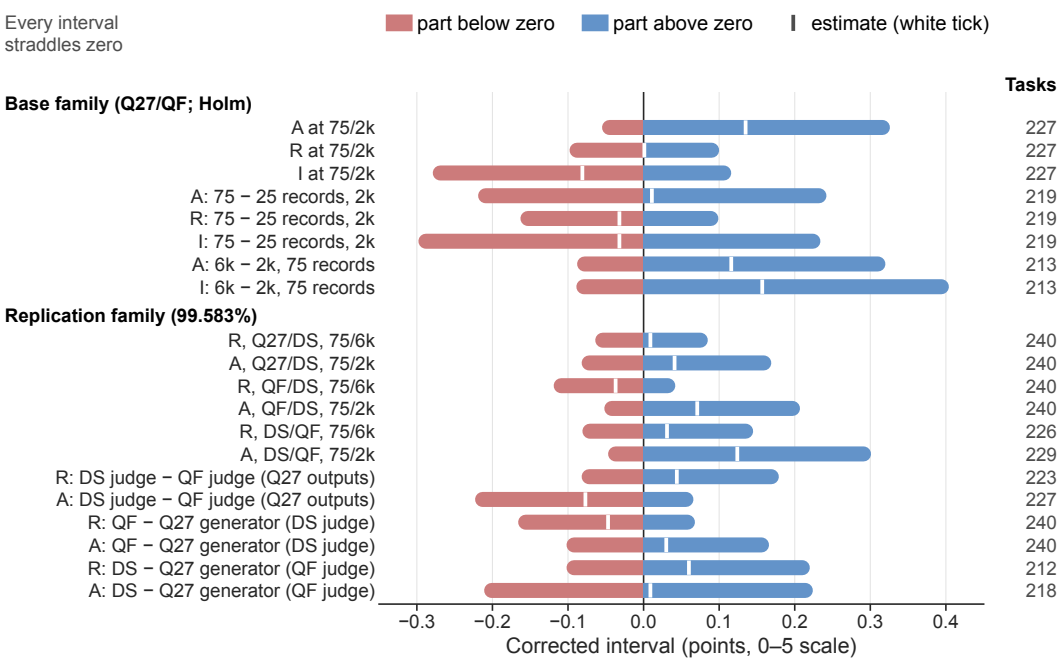


Figure 6: Pre-specified static difference families (points). Base: Qwen-27B with the Qwen-Flash judge, Holm correction with simultaneous max-$|t|$ intervals. Replication: Bonferroni correction over twelve contrasts; its last six rows compare a contrast between pairings on common tasks. Each bar spans the corrected interval (red below zero, blue above), the white tick marks the estimate, and the right column gives the tasks. Q27, QF, and DS abbreviate Qwen-27B, Qwen-Flash, and DeepSeek-Flash.

Table 4: Presentation with admitted records and relation statements held fixed. Means average access methods and seeds over the $n$ tasks of the paired contrast; $R$ is linear minus grouped, with an unadjusted 95% DOI-cluster interval; Relative is $R$ as a percentage of the grouped mean. Equivalence: pre-specified interval at 75/6k (95% for Q27/QF; 98.3% for the replication pairings), all within the $\pm 0.25$ margin.

| Generator/judge | Setting | $n$ | Linear | Grouped | $R$ | Relative | 95% interval | Equivalence interval |
|---|---|---|---|---|---|---|---|---|
| Q27/QF | 25/2k | 231 | 2.887 | 2.859 | +0.028 | +1.0% | [−0.046, 0.104] | |
| | 75/2k | 227 | 2.765 | 2.764 | +0.001 | 0.0% | [−0.070, 0.074] | |
| | 25/6k | 231 | 2.929 | 2.949 | −0.021 | −0.7% | [−0.101, 0.059] | |
| | 75/6k | 223 | 2.826 | 2.860 | −0.034 | −1.2% | [−0.118, 0.046] | [−0.116, 0.049] |
| Q27/DS | 25/2k | 240 | 2.860 | 2.859 | +0.001 | 0.0% | [−0.055, 0.055] | |
| | 75/2k | 240 | 2.790 | 2.768 | +0.022 | +0.8% | [−0.033, 0.074] | |
| | 25/6k | 240 | 2.896 | 2.938 | −0.042 | −1.4% | [−0.097, 0.012] | |
| | 75/6k | 240 | 2.850 | 2.841 | +0.009 | +0.3% | [−0.041, 0.060] | [−0.051, 0.072] |
| QF/DS | 25/2k | 240 | 2.907 | 2.841 | +0.067 | +2.3% | [0.005, 0.130] | |
| | 75/2k | 240 | 2.821 | 2.823 | −0.002 | −0.1% | [−0.061, 0.059] | |
| | 25/6k | 240 | 2.985 | 3.007 | −0.022 | −0.7% | [−0.084, 0.036] | |
| | 75/6k | 240 | 2.888 | 2.925 | −0.037 | −1.3% | [−0.093, 0.018] | [−0.105, 0.031] |
| DS/QF | 25/2k | 228 | 3.068 | 3.070 | −0.002 | −0.1% | [−0.075, 0.071] | |
| | 75/2k | 229 | 3.008 | 3.032 | −0.024 | −0.8% | [−0.096, 0.049] | |
| | 25/6k | 234 | 3.188 | 3.188 | 0.000 | 0.0% | [−0.075, 0.076] | |
| | 75/6k | 226 | 3.121 | 3.090 | +0.031 | +1.0% | [−0.048, 0.109] | [−0.065, 0.125] |

judge scored these responses 0, so its inspiration-paper contrast includes this abstention. By comparison, an abstention classifier flags 76 of the 240 Qwen-27B and 34 of the Qwen-Flash question-only outputs. Each auxiliary condition was run only at the settings where it addresses its question. The question-only input is the same at every pool size and budget. The inspiration-paper input differs across settings only in record indices and, at 2k, in the few inspiration papers that do not fit (recall 0.96). Both were therefore run at 75 records and 2k only. Dense backfill targets the budget that

Table 5: Paired pool-size contrast, 75 minus 25 candidate records, in points with unadjusted 95% intervals and relative to the 25-record mean (exploratory). Tasks are given for 2k/6k.

| | | 2k | | 6k | | |
|---|---|---|---|---|---|---|
| Generator / judge | Averaged over | Difference [95%] | Relative | Difference [95%] | Relative | Tasks |
| Qwen-27B / Qwen-Flash | all four | $-0.13$ $[-0.26, -0.02]$ | $-4.6\%$ | $-0.11$ $[-0.23, 0.01]$ | $-3.6\%$ | 219/216 |
| | hybrid | $-0.15$ $[-0.28, -0.01]$ | $-4.9\%$ | $-0.09$ $[-0.22, 0.03]$ | $-3.0\%$ | 225/230 |
| | graph-hop | $-0.10$ $[-0.26, 0.05]$ | $-3.6\%$ | $-0.14$ $[-0.29, -0.00]$ | $-4.9\%$ | 233/226 |
| Qwen-27B / DeepSeek-Flash | all four | $-0.08$ $[-0.19, 0.02]$ | $-2.8\%$ | $-0.07$ $[-0.18, 0.02]$ | $-2.4\%$ | 240/240 |
| | hybrid | $-0.06$ $[-0.18, 0.04]$ | $-2.1\%$ | $-0.05$ $[-0.17, 0.06]$ | $-1.6\%$ | 240/240 |
| | graph-hop | $-0.10$ $[-0.23, 0.01]$ | $-3.6\%$ | $-0.10$ $[-0.22, 0.01]$ | $-3.3\%$ | 240/240 |
| Qwen-Flash / DeepSeek-Flash | all four | $-0.05$ $[-0.19, 0.06]$ | $-1.8\%$ | $-0.09$ $[-0.21, 0.01]$ | $-3.0\%$ | 240/240 |
| | hybrid | $-0.02$ $[-0.16, 0.11]$ | $-0.7\%$ | $-0.06$ $[-0.19, 0.05]$ | $-2.0\%$ | 240/240 |
| | graph-hop | $-0.08$ $[-0.23, 0.04]$ | $-2.9\%$ | $-0.12$ $[-0.26, 0.00]$ | $-4.0\%$ | 240/240 |
| DeepSeek-Flash / Qwen-Flash | all four | $-0.08$ $[-0.23, 0.05]$ | $-2.6\%$ | $-0.08$ $[-0.20, 0.02]$ | $-2.6\%$ | 217/220 |
| | hybrid | $-0.04$ $[-0.21, 0.11]$ | $-1.2\%$ | $-0.04$ $[-0.16, 0.07]$ | $-1.3\%$ | 230/229 |
| | graph-hop | $-0.12$ $[-0.28, 0.03]$ | $-3.9\%$ | $-0.15$ $[-0.32, 0.00]$ | $-4.8\%$ | 226/229 |

graph-hop access leaves unused at 6k (72–73%, against 26% at 2k). Relation removal was run at 6k, where each task carries about 40 relation statements, and shuffled order at 75 records.

Table 6: Auxiliary static contrasts per generator/judge pairing (abbreviated as in Table 1), one seed, linear presentation, in points with unadjusted 95% intervals. Each contrast is listed at the settings where it was run.

| Contrast | Setting | Q27/QF | Q27/DS | QF/DS | DS/QF |
|---|---|---|---|---|---|
| Hybrid − sequential | 25/2k | $+0.74$ $[0.50, 0.97]$ | $+0.52$ $[0.35, 0.69]$ | $+0.79$ $[0.60, 0.99]$ | $+0.93$ $[0.70, 1.16]$ |
| | 75/2k | $+0.81$ $[0.57, 1.06]$ | $+0.65$ $[0.46, 0.83]$ | $+1.01$ $[0.78, 1.24]$ | $+1.05$ $[0.76, 1.34]$ |
| | 25/6k | $+0.28$ $[0.07, 0.48]$ | $+0.17$ $[0.03, 0.32]$ | $+0.33$ $[0.17, 0.48]$ | $+0.19$ $[0.01, 0.37]$ |
| | 75/6k | $+0.52$ $[0.29, 0.75]$ | $+0.42$ $[0.23, 0.61]$ | $+0.33$ $[0.17, 0.48]$ | $+0.43$ $[0.22, 0.66]$ |
| Graph-hop − sequential | 25/2k | $+0.70$ $[0.47, 0.94]$ | $+0.47$ $[0.30, 0.64]$ | $+0.82$ $[0.63, 1.01]$ | $+0.85$ $[0.63, 1.07]$ |
| | 75/2k | $+0.74$ $[0.51, 0.97]$ | $+0.62$ $[0.45, 0.79]$ | $+0.91$ $[0.70, 1.12]$ | $+0.98$ $[0.69, 1.28]$ |
| | 25/6k | $+0.13$ $[-0.06, 0.32]$ | $+0.09$ $[-0.07, 0.24]$ | $+0.15$ $[-0.00, 0.31]$ | $+0.13$ $[-0.06, 0.32]$ |
| | 75/6k | $+0.29$ $[0.05, 0.53]$ | $+0.32$ $[0.13, 0.50]$ | $+0.17$ $[0.00, 0.34]$ | $+0.13$ $[-0.14, 0.40]$ |
| Backfill − graph-hop | 25/6k | $+0.13$ $[-0.02, 0.28]$ | $+0.19$ $[0.06, 0.31]$ | $+0.13$ $[0.00, 0.27]$ | $+0.13$ $[-0.05, 0.31]$ |
| | 75/6k | $+0.37$ $[0.19, 0.55]$ | $+0.20$ $[0.07, 0.35]$ | $+0.30$ $[0.16, 0.44]$ | $+0.41$ $[0.20, 0.62]$ |
| Backfill − hybrid | 25/6k | $+0.00$ $[-0.14, 0.15]$ | $+0.10$ $[0.00, 0.20]$ | $-0.04$ $[-0.14, 0.06]$ | $+0.09$ $[-0.06, 0.23]$ |
| | 75/6k | $+0.14$ $[-0.02, 0.30]$ | $+0.10$ $[-0.02, 0.23]$ | $+0.15$ $[0.03, 0.26]$ | $+0.11$ $[-0.07, 0.28]$ |
| Inspiration papers − question only | 75/2k | $+1.34$ $[1.05, 1.62]$ | $+1.06$ $[0.82, 1.30]$ | $+0.48$ $[0.28, 0.68]$ | $+3.27$ $[3.05, 3.47]$ |

Table 7: Presentation controls of the base pairing (Qwen-27B with the Qwen-Flash judge), one seed, in points. The relation-removal intervals are at the 97.5% level of their two-contrast family.

| Contrast | Setting | Tasks | Difference [interval] | Level |
|---|---|---|---|---|
| Shuffled − linear (hybrid) | 75/2k | 234 | $+0.09$ $[-0.05, 0.23]$ | 95% |
| | 75/6k | 240 | $+0.14$ $[-0.01, 0.28]$ | 95% |
| Shuffled − grouped (hybrid) | 75/2k | 238 | $+0.03$ $[-0.12, 0.18]$ | 95% |
| | 75/6k | 238 | $+0.01$ $[-0.13, 0.15]$ | 95% |
| With − without relation statements (hybrid) | 25/6k | 237 | $+0.02$ $[-0.16, 0.20]$ | 97.5% |
| | 75/6k | 238 | $-0.12$ $[-0.30, 0.06]$ | 97.5% |

The relation-removal control keeps the admitted records identical on both sides (11.4 and 10.8 records per task at 75 and 25 records) and deletes the 43 and 42 relation statements per task, which shortens the evidence by about a third, from about 5,700 to 3,750 tokens at 75 records. The freed budget is deliberately left unfilled: refilling it would admit new records and change the record set, whereas leaving it empty isolates the presence of the statements at the cost of also shortening the prompt. Deleting the statements produced no detectable change in either pool (Table 7), which is consistent with the admitted records, rather than the relation statements, carrying the information that the matching score rewards. Because the control changes the statements and the prompt length together, it bounds the statements' own contribution rather than settling it.

# D PROTOCOL DETAILS

## D.1 STATIC STUDY

**Records and relation statements.** A static record is a candidate paper's title and abstract, shown with its index in the benchmark's candidate list; the benchmark's labels are never shown. Relation statements were extracted by Qwen-Flash (temperature 0, reasoning disabled) from each record separately, with at most eight statements per record. Each statement has one of four types (a mechanism *addresses* a problem, the paper *uses* a mechanism, a mechanism *extends* another, two findings *contrast*) and must quote a verbatim evidence span from the title or abstract, so every statement is traceable to a quoted span of its source text; it is rendered as "`[#12] <source> --type--> <target> ("<span>")`". Because every statement must quote its evidence, the type distribution follows what abstracts state explicitly: most of the 17,111 statements are *uses* (10,753) or *addresses* (5,884), and *contrasts* (319) and *extends* (155) are rarer. Because extraction is per record, statements never link two records. Cross-record structure is built at access time: graph-hop access merges identically named entities across records, so every edge of its graph is an extracted statement and every cross-record link an exact entity match. Richer cross-record relations are outside this study (Section 6).

**Access methods.** *Hybrid retrieval* queries with the research question. It retrieves the top 25 items by BM25 and the top 25 by bge-m3 dense similarity from an index containing each record's text and each of its relation statements, maps items to their records, and reranks the union of records (at most 50) with bge-reranker-v2-m3 on the question and the record text. Each record carries all its statements. *Graph-hop access* forms a graph whose nodes are records and typed entity nodes and whose edges come from the relation statements. It starts from the three records with the highest bge-m3 similarity to the question, traverses two hops breadth-first, keeps at most 50 nodes ordered by hop count and similarity, and maps nodes back to records. Each record carries only the statements on the traversal. *Sequential admission* takes records in the benchmark's list order, without a query. *Dense backfill* is triggered only when graph-hop admission stops because its candidates are exhausted; it appends the pool's unreached records ordered by bge-m3 similarity to the question, keeping the graph-hop prefix unchanged.

**Admission and presentation.** All methods admit whole records in their own order and stop at the first record that does not fit; records are never skipped or truncated. The budget covers the evidence block and is measured with the Qwen-27B tokenizer. A record is admitted only if the longest of the linear, relation-grouped, and shuffled renderings still fits, which guarantees identical admitted records and relation statements across presentations. The *linear* presentation lists each record followed by its statements. The *relation-grouped* presentation lists all record texts under a records heading, followed by the statements grouped under one heading per relation type. The *shuffled* control permutes records and their statements within the linear layout.

**Controls.** The inspiration-paper condition supplies the benchmark's designated inspiration papers for the task, with their relation statements, in list order under the 75-record, 2k setting; the budget occasionally excludes one (recall 0.96). The question-only condition sends the same instruction with an empty evidence block. The relation-removal control re-renders the records admitted by hybrid retrieval with linear presentation at 6k, deleting all relation statements without re-retrieval or refilling the freed budget. Sequential-admission, shuffled, backfill, inspiration-paper, question-only, and relation-removal conditions each use one seed.

**Generation and judging.** The generator is instructed to propose one concrete hypothesis using only the evidence records and to cite their identifiers, returning JSON with the hypothesis and cited identifiers; only the hypothesis is judged. Generation uses temperature 0.3 and at most 1,024 output tokens. The judges use the official ResearchBench matching prompt and parser unchanged, at temperature 0 with a 4,096-token output limit and one retry, so that scores stay comparable with the benchmark's protocol. The score (0–5) counts how many key methodological points of the reference hypothesis the generated hypothesis covers and whether it uses them similarly, and penalizes clearly flawed additions at the top of the scale. We keep two properties of the official implementation as they are: the benchmark data provide no separate key-point summary, so the reference hypothesis itself

is passed as the key points, and the official prompt refers to chemistry as the discipline for every task (the generator, by contrast, is told the task's discipline; Appendix G). Both properties are constant across conditions, so the two sides of every paired contrast are judged under the same prompt, and the comparison with human annotators in Appendix F was made with the judges configured in this way.

### D.2 Dynamic study

**Tasks, records, and branches.** The 150 tasks are held-out recombinations from the CHIMERA recombination-prediction test split without target leakage, stratified by relation type and by whether the recombination crosses domains. A record is a knowledge-base recombination instance, shown as its source abstract, the recombination's context, and the two entities involved, preceded by a short identifier and its branch label. The large pool is a fixed sample of 8,000 records from years up to 2023, stratified by year and relation; the small pool is a nested sample of 2,000. For each task, records from the anchor's year or later, records from the anchor's paper, and records mentioning the held-out recombination partner are removed. A record's branches are its two fine-grained domains; the task's primary branch is the anchor's fine-grained domain, with a fallback to the coarse domain when fewer than five eligible records carry the fine label. A record is shown as primary-branch if either of its domains is the primary branch, and as another branch otherwise.

**Development step.** Each step queries with the task text (the anchor concept, its branch, and the context) and the current proposal. Literature is retrieved first under a 4,000-token cap, then history under a 1,536-token cap, and the whole prompt is limited to 6,000 tokens; the current proposal is truncated to 1,024 tokens and the output to 1,024. Hybrid retrieval uses separate literature and history indexes (25 literature candidates per retriever, 3 history candidates), each reranked. Graph-hop access uses three literature and three history anchors, traverses two hops, and backfills literature only. History contains all earlier candidates and selections of the same trajectory; a selection identical to the current proposal is not shown again. The instruction for both operators asks for a proposal of at most 250 words with motivation, method, and expected result sections, requires the anchor to remain central, and requires citations by record identifier to the evidence and history shown. Refinement allows auxiliary mechanisms only from primary-branch records; recombination requires at least one mechanism from another branch, its citation, and an explanation of how it connects to the anchor.

**Selection.** The selector (DeepSeek-Flash, temperature 0) rates each candidate from 0 to 10 for scientific soundness, novelty, and specificity, seeing only the candidate text. The highest score wins; ties are broken by a seeded hash of task and step. Tied top scores occurred in 61–81% of selections across conditions and systems.

**Citation measure.** Each identifier cited in the selected proposal is classified as a displayed literature record of the primary branch, a displayed literature record of another branch, a pool record never displayed in the trajectory, a history entry, or unresolved. The outcome is 1 if at least one cited identifier is a displayed literature record of another branch; otherwise it is 0, and the proposal remains in the denominator. Across 5,500 selected outputs of the base system, four identifiers were unresolved and none referred to an undisplayed record.

**Controls and analysis sets.** Independent sampling retrieves once for the initial proposal and sends the same prompt eight times before one selection. History-off trajectories show no history to either access or generation, keep the literature cap unchanged, and still pass each selected proposal to the next step. Twenty tasks, chosen by stratified sampling, were used for protocol development and judge calibration; all primary analyses use the other 130. The history-off subset of 50 tasks overlaps the development tasks in 11, leaving 39 primary tasks.

### D.3 Timing of analysis decisions

The static analysis plan, including the equivalence margin, was fixed before any condition-level result of the pilot was read and before the main collection. Dense backfill was added as a secondary diagnostic after the pilot results and before the main collection. The dynamic protocol, including graph-hop plus backfill and separate literature and history anchors, was chosen after the static results

and a development pilot. Its primary outcome was originally judge-rated proposal quality; after the candidate judge configurations missed the pre-set calibration threshold on development items, the citation measure became the primary outcome, and its analysis plan was fixed before formal collection. That plan had nine secondary slots at Bonferroni-9 (99.44%). Because one slot holds two separately tested history-off contrasts, all ten secondary contrasts were re-reported at 99.5% after the results were read, and the interaction interval was set at the analysis stage. No secondary contrast is rejected at either level. The replications with further generators and a second judge were planned, with their own contrast families, after the base results and before their runs. The relation-removal control was added after the static and dynamic results were known. The replication of the auxiliary controls in the three other pairings was arranged after all results were known and is descriptive. The paired pool-size contrast, the operator contrasts, and the small-pool access contrasts of the dynamic study were computed while preparing this paper. None of these plans was registered externally.

## E Additional Dynamic Results

Table 8 gives the paired step-four access and operator contrasts with their intervals, and Figure 7 the step-wise counts behind Table 2, including steps two and three, and the counts of the control conditions for all three systems.

Table 8: Paired step-four contrasts in multi-step development, percentage points, 130 primary tasks. Access: graph-hop plus backfill minus hybrid retrieval; large-pool intervals ($^\dagger$) are pre-specified (97.5% for Qwen-27B, 99.375% for the replication generators), small-pool intervals are descriptive 95% score intervals. Operator: recombination minus refinement, with descriptive 95% task-bootstrap intervals.

| | | Access difference | | Operator difference | |
|---|---|---|---|---|---|
| Generator | Pool | Refinement | Recombination | Hybrid | Graph-hop + backfill |
| Qwen-27B | small | +3.8 [−2.9, 11.0] | −2.3 [−11.0, 6.3] | +62.3 [53.8, 70.8] | +56.2 [47.7, 64.6] |
| | large | −3.8 [−12.4, 4.4]$^\dagger$ | +6.2 [−4.1, 16.5]$^\dagger$ | +52.3 [43.1, 60.8] | +62.3 [53.8, 70.8] |
| Qwen-Flash | small | −3.1 [−10.0, 3.6] | +1.5 [−7.2, 10.3] | +53.1 [44.6, 62.3] | +57.7 [49.2, 66.2] |
| | large | −3.1 [−11.3, 4.1]$^\dagger$ | −0.8 [−13.1, 11.6]$^\dagger$ | +60.0 [50.8, 68.5] | +62.3 [53.8, 70.8] |
| DeepSeek-Flash | small | −0.8 [−8.7, 7.1] | −5.4 [−12.8, 1.6] | +58.5 [49.2, 67.7] | +53.8 [45.4, 62.3] |
| | large | +7.7 [−0.9, 17.7]$^\dagger$ | 0.0 [−12.2, 12.2]$^\dagger$ | +62.3 [53.8, 70.8] | +54.6 [46.2, 63.1] |

**Secondary family of the base system.** Figure 8 shows the ten secondary contrasts of the base system at 99.5% (Bonferroni over ten hypotheses, a level set after the results were read; the plan's nine slots at 99.44% give the same empty rejection set). None is rejected; two pool-size contrasts have unadjusted 95% intervals that exclude zero (refinement with graph-hop plus backfill, $[-15.5, -0.3]$; recombination with hybrid retrieval, $[-18.8, -1.4]$), both favoring the small pool. The interaction is the task-level mean of the access difference under recombination minus that under refinement; no test statistic was specified for it, so only its interval is reported.

**Replication systems.** In the pre-specified replication family, no access difference is rejected. Of the four changes relative to Qwen-27B, three range from $-6.9$ to $+0.8$; the fourth, the large-pool refinement access difference of DeepSeek-Flash, is $+11.5$ with an interval $[0.0, 24.6]$ that reaches zero. For Qwen-Flash and DeepSeek-Flash, the same secondary contrasts are descriptive (95% intervals); no confirmatory family was added. Qwen-Flash resembles the base system throughout. DeepSeek-Flash does so as well, except that under refinement with hybrid retrieval its step-four rate falls below both the independent best-of-eight rate ($-15.4$, $[-22.6, -10.2]$) and the small-pool rate ($-10.8$, $[-18.0, -4.5]$). This is the largest instance of the step-wise decline of refinement rates in Table 2, from 20.8% at step one to 9.2% at step four in this condition, and it widens the operator difference. No generator shows an access-by-operator interaction: the estimate is $+10.0$ for Qwen-27B, $+2.3$ for Qwen-Flash, and $-7.7$ for DeepSeek-Flash, all with 95% intervals containing zero.

**Evidence admitted.** Averaged over steps, both access methods admitted 6.7–7.2 literature records per step, 40–50% of them from branches other than the primary branch, in all three systems. Backfill thus served its purpose: the exposure gap of the static study does not arise, and the dynamic access

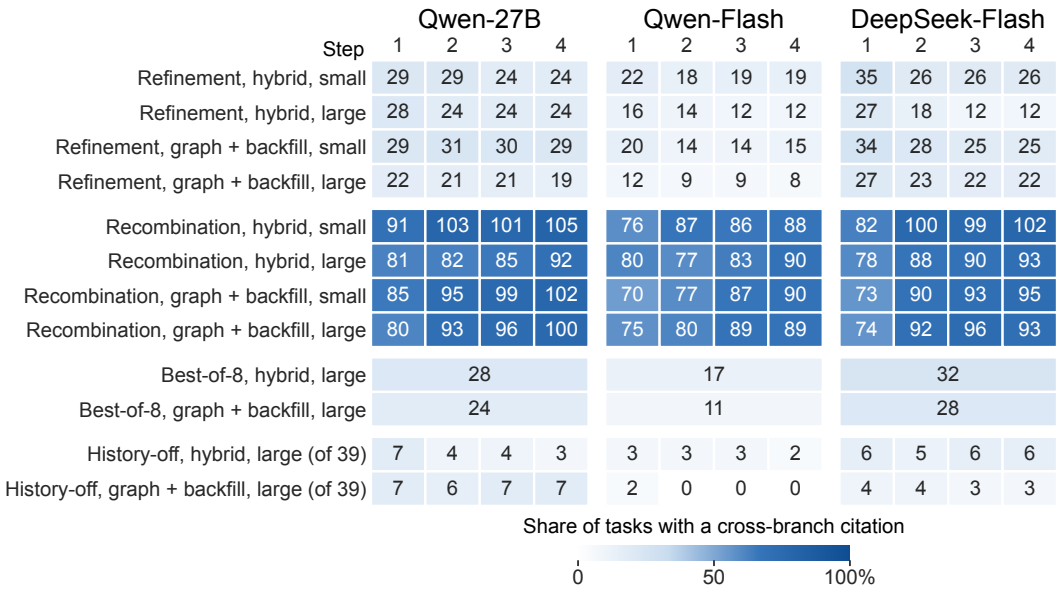


Figure 7: Selected proposals with explicit cross-branch citation at development steps 1–4 (counts; independent sampling has a single count, spanning the four steps). Shading gives the count as a share of the tasks. Trajectory and independent-sampling rows use the 130 primary tasks; history-off rows use the 39 primary tasks of the history-off subset.

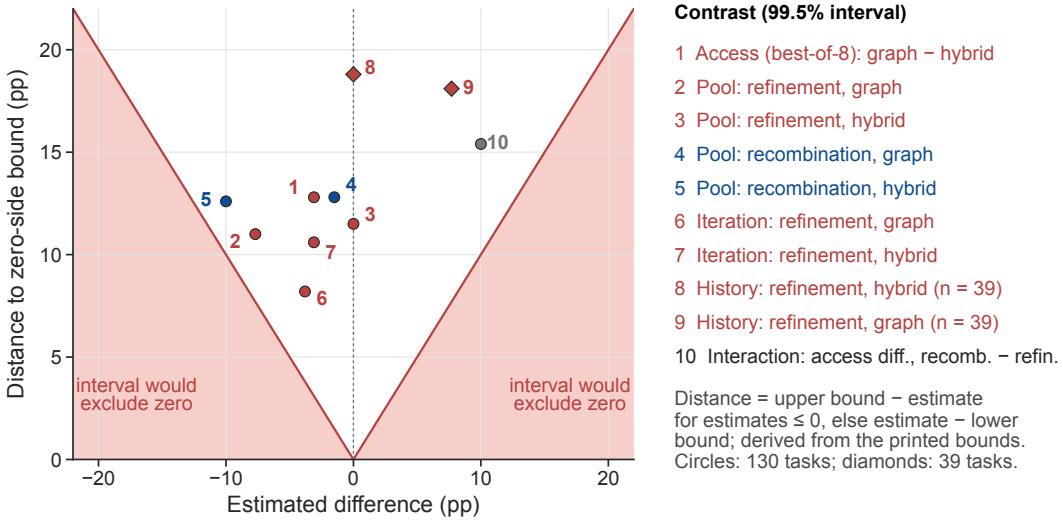


Figure 8: Dynamic secondary family of the base system (Qwen-27B generator), percentage points, 99.5% intervals. Each point is one contrast, numbered as in the legend; the vertical axis is the distance from the estimate to the interval bound on the side of zero, so an interval would exclude zero only for a point in the shaded region. "Graph" denotes graph-hop plus backfill. The first nine contrasts use paired score intervals and exact McNemar tests; the interaction uses a task-level bootstrap.

contrast compares the two organizations at similar exposure. In the base system, the graph-hop traversal alone exhausted its candidates before filling the literature budget at 91–99% of retrievals in the four graph-hop trajectory conditions, and dense backfill then added about 3.5 records, so about half of the literature in those prompts came from the traversal and half from backfill; without backfill, the static bottleneck would have recurred at nearly every step. History-on trajectories admitted about two history entries per step after the first. Selected proposals cited literature records only: no selected

proposal in any system cited a history entry; one of the base system's 12,800 candidates did, and it was not selected. The exclusion of history entries from the outcome (Appendix A) therefore never changed a value, and the outcome rests on displayed literature records alone, in line with the history-off control, which changed the rate by 0.0 and +7.7 points for the base generator (Section 5.4).

## F Automatic Judges and Calibration

**Static judges.** Both static judges were compared with a human reference on 40 hypotheses generated by Qwen-27B in the pilot stage (75 records, 2k), drawn across the four main conditions, the inspiration-paper condition, and the question-only condition and covering all 12 disciplines. Two annotators, one of them an author, scored each hypothesis independently on the same 0–5 matching scale, without access to condition, model, or judge score. Scores differing by at most one point were averaged; differences of two or more points were resolved by a third annotator, blind to the first two scores, as the median of the three scores. The annotators reported no model assistance. Agreement between the two annotators was ICC(2,1) 0.66 [0.46, 0.81]. Against the adjudicated reference, the Qwen-Flash judge reaches ICC 0.74 [0.54, 0.87] and Spearman 0.69, scoring 0.31 points higher on average; the DeepSeek-Flash judge reaches ICC 0.84 [0.69, 0.93] and Spearman 0.82, scoring 0.46 points higher. No acceptance threshold was set in advance. The same 40 items also informed the configuration of the Qwen-Flash judge; because they contain only Qwen-27B outputs, this calibration applies to that generator's outputs. On identical Qwen-27B outputs from the main matrix, the two judges agree with ICC 0.68–0.71 across the four pool–budget settings, and their means differ by less than 0.04 points.

**Dynamic judges.** The two dynamic judges share one rubric and prompt and differ only in the model. Each judge sees the anchor, its branch and context, the proposal, and the records the proposal cites as they were displayed, and rates clarity, groundedness, mechanism compatibility, and feasibility from 0 to 3; we average the four ratings. Before the replication runs, these configurations were compared with a human reference on 40 development endpoints of the base system: two annotators (one of them an author) rated each endpoint independently and blind to condition, and their mean served as the reference, with human–human Spearman correlation 0.69 on the averaged rating. A threshold of Spearman 0.4 had been set in advance. The Qwen-Flash configuration reached 0.18, the DeepSeek-Flash configuration 0.30, and an earlier configuration 0.21, so we use neither judge as a measure of proposal quality.

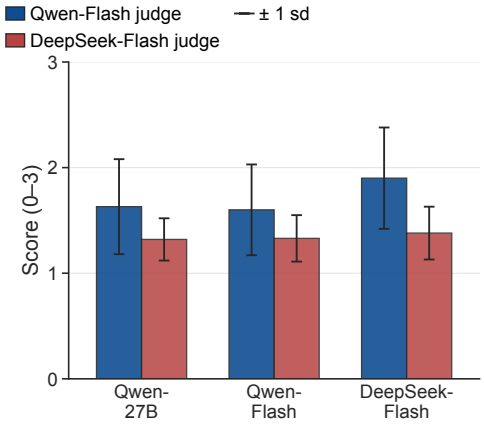


Figure 9: Dynamic endpoint scores (average of four 0–3 ratings) by generator and judge, mean ± 1 sd over 1,600 endpoints each.

The judges differ in more than level (Figure 9). The DeepSeek-Flash judge rated feasibility 1 for at least 1,580 of 1,600 endpoints in every system, and it compresses the other dimensions as well. The two judges also rank generators differently: on common tasks within each of the twelve conditions, DeepSeek-Flash outputs exceed Qwen-27B outputs by 0.12–0.36 points under the Qwen-Flash judge but by 0.00–0.14 under the DeepSeek-Flash judge, which shares its model with the DeepSeek-Flash

generator and with the selector. For access organization, 28 of the 30 graph-minus-hybrid differences (five contrasts, six combinations) have 95% intervals containing zero. The two exceptions occur under the Qwen-Flash judge (Qwen-27B, recombination, large pool: $-0.14$, $[-0.23, -0.05]$; DeepSeek-Flash, recombination, small pool: $-0.10$, $[-0.19, -0.00]$), and the DeepSeek-Flash judge gives $-0.01$ and $-0.02$ for the same comparisons. In the first of these, cross-branch citation was higher under graph-hop plus backfill while the Qwen-Flash judge's score was lower. Given the weak agreement and the missed calibration threshold, we do not interpret these differences as differences in proposal quality.

## G DATASETS

Table 9 summarizes the two sources and how they are used.

**ResearchBench.** ResearchBench (Liu et al., 2026) decomposes scientific discovery into inspiration retrieval, hypothesis composition, and hypothesis ranking. Its release contains 1,369 paper records and 1,367 retrieval tasks across 12 disciplines: astronomy, biology, business, cell biology, chemistry, earth science, energy science, environmental science, law, material science, mathematics, and physics. Each retrieval task pairs a research question with 75 candidate papers (title and abstract), among which the benchmark designates the inspiration papers of the source paper; the matching composition task provides the source paper's hypothesis, which is the reference for the matching score. We use the first 20 retrieval tasks of each discipline in file order, 240 tasks from 192 source papers with 4,534 distinct candidate records. Models see the discipline, the research question, and the candidates' titles and abstracts with their list index; inspiration labels, the benchmark's background survey, the reference hypothesis, and all other fields are withheld. The data are released for non-commercial research under CC BY-NC 4.0.

**CHIMERA.** CHIMERA (Sternlicht & Hope, 2026) is a knowledge base of more than 28,000 recombination instances extracted automatically from arXiv abstracts published from 2019 to 2024. Each instance links a source and a target concept of one paper by a relation, combination or inspiration, and records the abstract, the coarse (arXiv category) and fine-grained domains of both concepts, and the publication year. Its recombination-prediction version, in which a model predicts the concept that a source concept recombines with, adds the context of each recombination and flags for target leakage in that context. Its train, validation, and test splits contain 19,927, 412, and 7,825 instances, and the test split covers 2024. We draw the 150 dynamic tasks from test instances without target leakage, stratified by relation type and by whether the recombination crosses domains; the held-out target concept is never shown. Literature records are knowledge-base instances from 2019 to 2023: the large pool is a sample of 8,000 stratified by year and relation, and the small pool a nested sample of 2,000; Appendix D describes the per-task filtering.

Table 9: Datasets and their use.

| | ResearchBench | CHIMERA |
|---|---|---|
| Source | Hugging Face `ankilok/ResearchBench` | Hugging Face `noystl/CHIMERA`, `noystl/Recombination-Pred` |
| Release | 1,369 papers; 1,367 retrieval tasks in 12 disciplines | more than 28,000 recombination instances (2019–2024); prediction splits 19,927 / 412 / 7,825 |
| Tasks used | 240 (first 20 per discipline; 192 source papers) | 150 test instances without target leakage (130 primary, 20 development) |
| Records | 4,534 candidate papers; pools of 25 or 75 per task | 8,000 (large) and 2,000 (small, nested) knowledge-base instances, 2019–2023 |
| Record text | title and abstract | source abstract, recombination context, the two entities |
| Relation statements | 17,111 (Qwen-Flash, one pass) | 33,561 (Qwen-Flash, one pass) |
| Reference | benchmark hypothesis of the source paper | primary branch of the anchor; branch labels of records |
| Outcome | official matching score (0–5) | explicit cross-branch citation at step four |

# H MODELS

Table 10 lists the models, which are called by the service identifiers given in Section 4. All models run on self-hosted OpenAI-compatible endpoints from their public weights. The three language models are called with reasoning disabled per request; the two retrieval models serve hybrid retrieval (dense retrieval and reranking), graph-hop seeding, and dense backfill. Token budgets are counted with the Qwen-27B tokenizer for every generator (Appendix B).

Table 10: Models. Parameter counts and architectures are those stated in the model cards and configuration files of the public weights.

| Name in paper | Public weights | Architecture and size | License |
|---|---|---|---|
| Qwen-27B | `Qwen/Qwen3.8-27B` | dense decoder, 64 layers, 27B parameters; 262k context | Apache-2.0 |
| Qwen-Flash | `Qwen/Qwen3.8-Flash-Next-FP8` | mixture of experts (512 experts, 10 active), 125B parameters with 6B activated, plus 51B n-gram embedding; 262k context | Qwen Community 1.0 |
| DeepSeek-Flash | `deepseek-ai/DeepSeek-V4.1-Flash` | mixture of experts (384 routed and 1 shared expert, 6 routed active), 552B backbone parameters with 16B activated per decoded token; 1M context | MIT |
| bge-m3 | `BAAI/bge-m3` | dense text embedding (Chen et al., 2024), XLM-RoBERTa encoder with 24 layers | MIT |
| bge-reranker-v2-m3 | `BAAI/bge-reranker-v2-m3` | cross-encoder reranker, XLM-RoBERTa encoder with 24 layers | Apache-2.0 |