STRUCTURE AND DYNAMICS OF AGENTIC INNOVATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study scientific innovation as a dynamical system: the *state* is a citation network of research ideas distilled from published papers, and the *dynamics* are LLM agents that navigate the network and write new ideas (and reference links) back into it. Agents propose ideas freely; each new idea is then scored by whether a similar paper appears in a later conference, published *after* both the papers in the network and the agent model's training cutoff. We create multiple agents to navigate a 16,208-node network built from NeurIPS, ICLR, ICML, and AAAI papers (2020–2024) and change the number of corpus papers an agent is allowed to read, from 100 up to the whole corpus. We find three things. First, more reading is not better: the share of ideas that later papers go on to realize rises and then falls, peaking at 800 papers, while both extremes sink to the same floor, agents given no papers doing no better than agents free to read all 16,208. The number of ideas produced follows the same curve. Second, team size buys only throughput: the hit rate is flat from 10 to 150 agents while the total scales linearly, because the agents hardly interact—in 10,300 actions, one agent cited another's idea exactly once. Third, the human citation graph is not what helps them: deleting all 80,962 edges makes generalists better, while removing semantic search makes them worse. Agents move through the network by meaning rather than by links, and the links they write back ignore a paper's popularity and favor recent frontier work. Reading scope—not team size, not citation structure—is the lever on the quality of machine-generated research ideas.

1 INTRODUCTION

Can a system of LLM agents, given the recent scientific literature as a navigable network, anticipate where science is going next? And if so, what features of the system—the structure of the network, the behavioral constraints on the agents, the size of the team—determine how well it does? These questions matter both for the emerging practice of automated research ideation (Lu et al., 2024; Si et al., 2025) and for the science of science, where the interplay between knowledge structure and discovery has long been studied on human data (Fortunato et al., 2018; Uzzi et al., 2013; Wu et al., 2019).

We formalize the setting as a dynamical system. The **state** is an *idea network*: each node is a one-paragraph distillation of a published paper, and directed edges are citations. The **dynamics** are LLM agents that act on the state through a small action set—semantic search, browsing a node and its neighbors, random jumps, adding or removing links, and *generating* a new idea node that cites the ideas it builds on. State and dynamics are inseparable: every generated node and link becomes part of the network that all agents subsequently navigate (a stigmergic channel (Theraulaz & Bonabeau, 1999)).

Evaluation exploits the arrow of time. The corpus contains only papers published before the agent model's training cutoff (September 2024). A generated idea scores only if a paper published *after* the cutoff realizes it—*anticipation*, verified by live literature search (Kinney et al., 2023; Priem et al., 2022) and graded by an LLM judge on a six-level realization scale, reported across three cumulative venue-recognition tiers (top venues, top+second-tier venues, any published paper). Matches to pre-cutoff papers or to corpus members are excluded, and—because multiple-choice probes show the agent model's parametric memory in the AI domain extends through April 2025, seven months

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.

past its official cutoff (Appendix A.10)—only realizations published on or after 2025-06-01 count. Neither pretraining memorization nor post-training memory can score.

Our central manipulation is deliberately minimal: a single *specialization dial*. Each "specialist" agent draws one topic from a 50-topic pool and may only *read* (search, browse, jump within, and hence cite) the m corpus papers nearest its topic anchor in embedding space, with $m \in \{100, 400, 800, 1600, 3200, 6400\}$; writing is always unconstrained. Unrestricted generalists provide the $m = \infty$ endpoint. A second manipulation varies team size $N \in \{10, 50, 100, 150\}$ at fixed per-agent interaction depth.

Findings.

- **More reading is not better (Sec. 5).** Accuracy rises and then falls as the dial widens, peaking at m=800—a coherent sub-area, not a single topic—at 31% tier-1 acc@$\geq$2, with the whole narrow-to-mid band $m \in \{100, 400, 800, 1600\}$ (24–31%) ahead of broad scopes (15–19%). Both ends of the dial sit on the same *parametric floor*: agents given nothing but a topic name (13%) do no better than agents free to read all 16,208 papers (14%). Reading pays only when it is narrow. Output follows the same curve—83 ideas at m=400 against 50 for generalists over the same 80 rounds—so narrow reading costs nothing in quantity. We had expected very narrow scopes to starve agents of material to recombine; they did not, though the narrow end's edge over the peak survives only under memory-contaminated scoring, not under the guarded metric.
- **Team size buys throughput, not quality (Sec. 6).** From N=10 to N=150 generalists at fixed depth, per-idea accuracy is flat (12–19%) while absolute hits grow roughly linearly. The reason is blunt: the agents barely interact. Across all scaling runs (575 ideas, 10,300 actions) not one agent cited another agent's generated idea.
- **The human citation graph is not what helps them (Sec. 7).** Deleting every initial citation edge makes generalists *better* (60 ideas at 27% tier-1 accuracy against 50 at 14%): with no neighbors to wander through, agents commit to semantic search, and even rebuild structure by their own logic. Disabling random jumps changes nothing. Disabling search quadruples production but *halves* accuracy (10%)—search is where genuine anticipation comes from; without it agents backfill from parametric memory, which the realization window discounts.
- **The dynamics leave a characteristic handwriting (Sec. 8).** How far agent-written links reach is almost invariant to scope (median cosine distance 0.144–0.153, slightly *below* the 0.179 of real citations): the dial moves *where* agents write, not *how far* they link. Their citations also ignore a paper's popularity—the median cited paper has in-degree 0–2, against 23 for the target of a real citation edge—and chase recency: 51% point at the corpus's final two years, against 8% for real edges. LLM dynamics grow the network by a different law than the preferential attachment (Barabási & Albert, 1999) that grew it.

Together these results suggest that in the current regime, the quality of machine-generated research ideas is governed by how each agent's reading is situated in the knowledge structure, while multi-agent organization contributes parallelism but not yet collaboration. Our full event logs, evaluation verdicts, and code are available at `https://anonymous.4open.science/r/structure-and-dynamics-of-innovation-D607/`.

2 Related work

LLM research ideation. Recent systems generate research ideas or entire papers with LLMs (Lu et al., 2024), and human studies find LLM-generated ideas can be judged more novel than expert ideas (Si et al., 2025). Evaluation in this literature relies mostly on human or LLM preference judgments of the ideas themselves. We instead score *anticipation*: an idea counts only if the future literature realized it, an outcome measure that needs no human panel and cannot be gamed by fluent writing alone.

Science of science. Large-scale studies of human science link discovery to network structure: atypical combinations of prior work predict impact (Uzzi et al., 2013), small teams disrupt while large teams develop (Wu et al., 2019), and citation networks grow by preferential attachment (Barabási & Albert, 1999; Fortunato et al., 2018). We port these questions to a controlled substrate where the

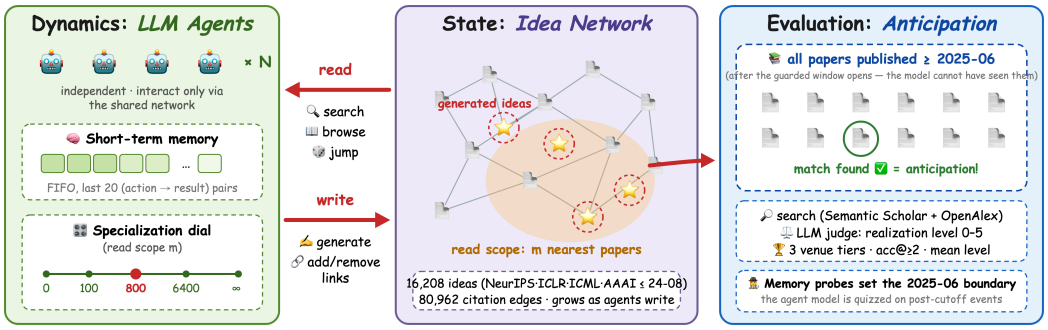

Figure 1: Overview. **Dynamics** (left): N LLM agents that interact only through the shared network, each with a 20-entry FIFO memory and a position on the specialization dial; their five actions live on the read/write channels to the state. **State** (middle): a citation network of 16,208 one-paragraph ideas distilled from four venues' papers; a specialist reads only the m corpus papers nearest its topic anchor, and generated ideas ($\star$) join the network permanently. **Evaluation** (right): each generated idea is matched against the set of papers published on/after 2025-06—after the guarded window opens, so the agent model cannot have seen them—via live literature search and an LLM judge; memory probes on post-cutoff events set the window.

"researchers" are LLM agents whose information diet can be intervened on directly—an experiment impossible with human scientists.

Multi-agent LLM systems. Most LLM multi-agent frameworks coordinate through explicit messaging. Our agents interact only stigmergically (Theraulaz & Bonabeau, 1999), through traces left in a shared artifact—the mechanism by which real science coordinates through the literature itself. Our finding that this channel stays essentially unused at present scales (Sec. 8) quantifies how far current agent teams are from literature-mediated collaboration.

3 The Environment: State and Dynamics

3.1 State: An Idea Network from Four Venues

We collect the 16,208 papers published at NeurIPS, ICLR, ICML, and AAAI between January 2020 and August 2024 with at least 25 citations, via the Semantic Scholar Open Data Platform (Kinney et al., 2023). Each paper's title and abstract are distilled by an LLM into a self-contained one-paragraph *idea*. Citation edges among corpus members are recovered from Semantic Scholar (references and citations arms) augmented with OpenAlex (Priem et al., 2022), yielding 80,962 edges (6.4% of nodes isolated). Ideas are embedded with a sentence encoder (Xiao et al., 2023); embeddings define semantic search, agent scopes, and the 2-D maps below (PCA + t-SNE (van der Maaten & Hinton, 2008)).

3.2 Dynamics: Agents and Actions

Agents share one action interface: `search` (semantic query, top-k over readable nodes), `browse` (read a node and its in/out neighbors), `sample_frontier` (random jump within readable scope), `generate` (add a new idea node with citations to the ideas it builds on), `add_links`/`remove_links` (curate existing edges). Each agent is a single LLM (`gpt-5`, medium reasoning) with a rolling memory of its last 20 action–result pairs and no other channel to teammates. Agents act in round-robin; one *round* = every agent acts once. There is no generation budget: whether and when to write is the agent's choice, so idea count is an outcome variable. Generated nodes and edges immediately join the shared state.

Table 1: The scoring scheme. **Top:** the judge's six-point realization scale—how close a retrieved paper is to the generated idea. Rows on or below the middle rule count as a hit for our headline metric acc@$\geq$2. **Bottom:** the three cumulative recognition tiers, which decide *which* retrieved papers are eligible to be scored at all. Levels and tiers are independent: an idea has a best level within each tier.

Level	Relation between the matched paper and the generated idea
0	different problem entirely
1	same broad area, different goal
2	same specific problem, unrelated approach
3	same problem, approaches in the same method family
4	same problem and core mechanism, differing in secondary components
5	essentially the same idea (problem, mechanism, and key design choices)

Tier	Realizations admitted (cumulative)
1	60 top venues (58 CCF-A, 2026 edition, plus COLM and RLC), or $\geq$ 50 citations
2	tier 1, plus CCF-B venues or $\geq$ 10 citations
3	any published paper

3.3 THE SPECIALIZATION DIAL

A specialist draws one topic from a pool of 50 corpus-derived subfields (clustering + LLM naming; seeded draws, distinct within a run). Its *readable region* is the m corpus papers nearest the topic anchor in embedding space (an equal-mass region; the radius adapts per topic). Reading, searching, jumping, *and citing* are confined to this region—citing means having read—while the content of generated ideas is unconstrained. We sweep $m \in \{100, 400, 800, 1600, 3200, 6400\}$ (conditions C2–C7, numbered by ascending scope); unrestricted generalists are C1 ($m = \infty$), and a zero endpoint C0 gives each agent ONLY its topic name—no reading, no navigation, direct topic-conditioned generation (7 ideas per agent, each prompted to differ from the agent's previous ones)—measuring what parametric knowledge alone achieves. All conditions use N=10 agents, 80 rounds ($\leq$7 for C0), and the same seed, so topic draws are identical across the sweep and only the radius varies.

Because topics vary in size (natural partition: median 228 papers, range 44–1096), an $m = 800$ region is *not* one topic: measured own-topic purity is 5–60% (median $\sim$35%), i.e., a coherent subarea neighborhood containing the topic plus its semantic neighbors. This is why we treat m as a continuous specialization dial rather than a one-topic/many-topic dichotomy.

4 EVALUATION: SEARCH-VERIFIED ANTICIPATION

For each generated idea we extract three literature-search queries (problem, method, combination), retrieve candidates from Semantic Scholar and OpenAlex, and have an LLM judge grade each candidate on the six-point realization scale of Table 1, which also defines the three cumulative recognition tiers we report. Candidates published on or before the training cutoff (2024-09-30) and candidates whose titles match corpus members are excluded before scoring, so neither memorization nor retrieval of the corpus can score. Near-duplicates of corpus ideas (embedding cosine $\geq$ 0.95) are floored to level 0.

Realization window. The official knowledge cutoff of the agent model is 2024-09-30, but official cutoffs are not hard data boundaries: multiple-choice probes (Appendix A.10) show the model reliably answers questions about AI-domain events through April 2025 and approaches chance from May–June 2025 onward. A "realization" the model may simply remember is not an anticipation. We therefore count only realizations published on or after **2025-06-01**—the earliest month at which probe accuracy is indistinguishable from chance—eight months past the official cutoff. Per tier we report the mean best realization level and acc@$\geq$2, the share of ideas whose best in-window match reaches level 2 (same specific problem)—our headline accuracy; higher thresholds ($\geq$3, $\geq$4, =5) are recorded in the released verdicts. Verbatim evaluation prompts, the full pipeline (retrieval, gating, caching), and sample generated ideas with their audit trails appear in Appendices A.8, A.4, and A.11.

Table 2: The specialization dial: N=10 agents, 80 rounds, one seed. Read scope m is the number of corpus papers each agent may read; writes are always free. C0 is the zero endpoint: agents receive ONLY their topic name and generate with no reading of any kind—the parametric-ideation floor. Only realizations published on/after 2025-06-01 count (Sec. 4). Each tier cell reports acc@$\geq$2, the number of realized ideas (level $\geq$ 2, in parentheses), and the mean best realization level.

	scope m	ideas	tier-1 acc (hits) / mean	tier-2 acc / mean	tier-3 acc / mean
C0	0 (topic name only)	70	13% (9) / 0.30	19% (13) / 0.43	67% (47) / 1.64
C2	100	67	24% (16) / 0.57	33% (22) / 0.78	73% (49) / 1.99
C3	400	**83**	24% (20) / 0.57	30% (25) / 0.73	**88% (73) / 2.40**
C4	800	74	**31% (23) / 0.77**	**41% (30) / 1.00**	80% (59) / 2.24
C5	1600	58	28% (16) / 0.76	34% (20) / 0.93	69% (40) / 1.88
C6	3200	55	15% (8) / 0.36	24% (13) / 0.65	62% (34) / 1.69
C7	6400	54	19% (10) / 0.43	28% (15) / 0.67	78% (42) / 2.20
C1	∞	50	14% (7) / 0.32	16% (8) / 0.38	56% (28) / 1.46

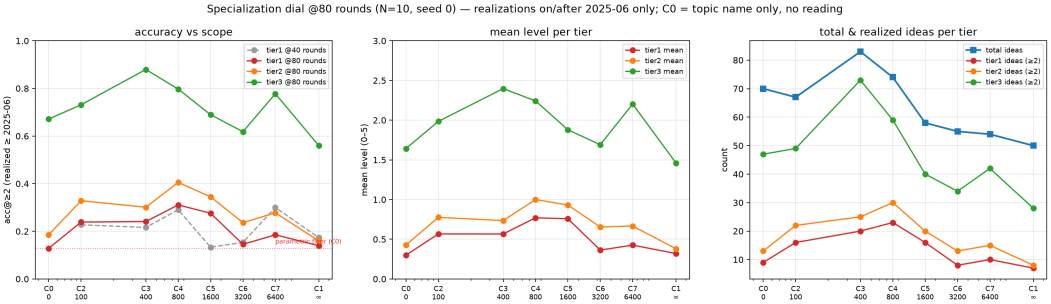

Figure 2: The specialization dial at 80 rounds. Left: acc@$\geq$2 per tier vs. read scope (log axis); the gray dashed line is the tier-1 curve at the 40-round midpoint, whose deviations from the 80-round curve illustrate small-n noise. Middle: mean realization level per tier. Right: total idea production and the number of realized ideas (level $\geq$ 2) per tier.

5 Experiment 1: The Specialization Dial

Table 2 and Figure 2 give the core result.

An inverted-U peaking at a sub-area scope. Tier-1 accuracy runs 24% $\rightarrow$ 24% $\rightarrow$ **31%** $\rightarrow$ 28% $\rightarrow$ 15% $\rightarrow$ 19% $\rightarrow$ 14% as scope widens from 100 papers to the whole corpus; mean realization level tracks the same shape in every tier. The m=800 sub-area neighborhood leads on accuracy, absolute hits, and mean level in tiers 1–2; the whole narrow-to-mid band (100–1600, 24–31%) clears every broad scope (14–19%); and even the extreme m=100 core—smaller than the median natural topic of 228 papers—holds 24%, refuting the pre-registered fear of material starvation at the narrow end. (Without the realization window, m=100 appears to win outright at 51%; its excess hits concentrate exactly in the memory-risk months, a caution documented in Appendix A.10.) Broad scopes underperform throughout (unfocused wandering; a semantically diffuse topic region produced an agent with zero output).

Both ends of the dial sit on the parametric floor. The zero endpoint C0—agents given only their topic name, generating from parametric knowledge alone—scores 13%, statistically identical to the unrestricted generalists' 14%. Reading the *entire* corpus adds nothing over reading nothing; only situated, narrow reading does, roughly doubling the floor at the m=800 peak. Every condition in this paper should be read against this 13% zero point: it is what the underlying model gives away for free.

Constraint beats freedom everywhere. Every scoped condition matches or beats the unrestricted generalists on every tier, by rate and by absolute hits, and the narrow band does so by a factor of about two.

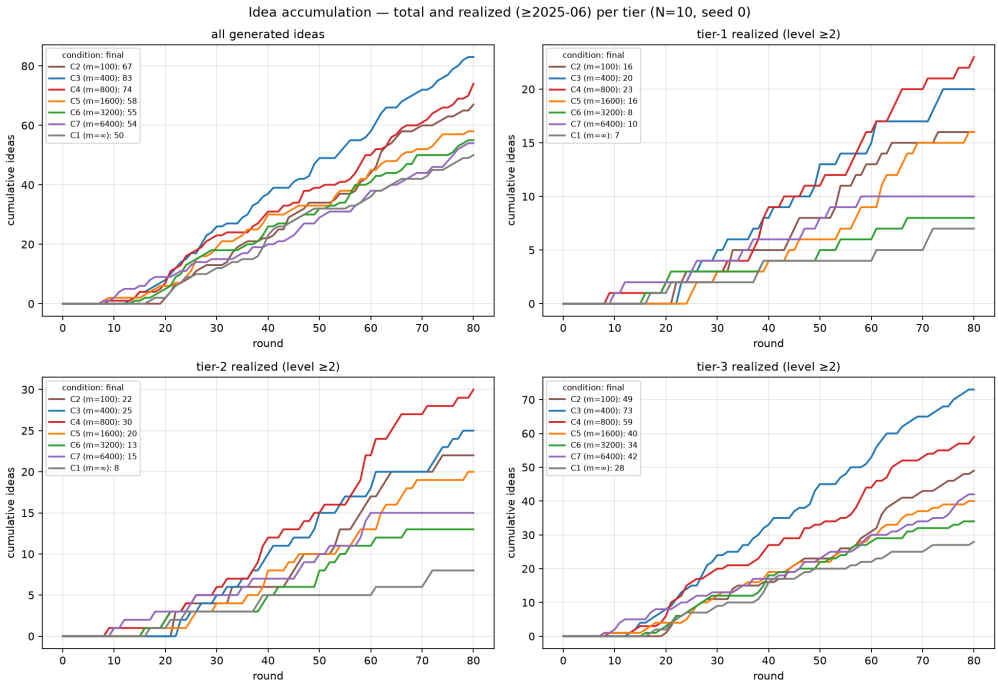

Figure 3: Cumulative idea production over rounds: all generated ideas (top left) and the subsets realized at level ≥ 2 in each recognition tier. All conditions start producing between rounds 8–17; the production gap is a *steady-state rate* effect, not a head start: narrow scopes settle into $\sim$1 idea/round while broad scopes run at $\sim$0.65/round with recurring plateaus. The realized subsets accumulate in near-constant proportion to production—no condition's quality drifts over the run—and no curve saturates by round 80.

Narrow readers ship; broad readers browse. Idea production peaks at m=400 and falls as scope widens (67, 83, 74, 58, 55, 54, 50 for m = 100 through ∞), and Figure 3 shows why: after a comparable exploration onset (first team idea between rounds 8 and 17; median personal onset round 21–24), narrow scopes settle into a faster and steadier production rate, while broad scopes stall in recurring exploration plateaus. Generalists redirect effort into curation: they performed 71 `add_links` actions versus 23–38 for scoped conditions.

What feeds a realized idea. The event logs let us trace each idea's provenance: we take the generating agent's preceding 20 actions (exactly its FIFO memory window) and classify every read in that window by the navigation channel that produced it—the node was surfaced by an earlier *search*, exposed as a neighbor *link* in an earlier browse, delivered by a random *jump*, or is the agent's *own* earlier idea. Table 3 aggregates all 1,328 evaluated ideas across our experiments. Realized and unrealized ideas are preceded by the *same volume* of reading ($\sim$10.6 reads); what differs is its composition. Ideas realized in recognized venues (tier 1) grew out of windows with more search-driven reads (5.78 vs. 4.55), fewer link-driven reads (3.69 vs. 4.64), fewer jumps, and almost no re-reading of the agent's own output. Search-fed material predicts realization while link-fed material predicts its absence—fully consistent with the channel ablations of Sec. 7, where removing links helps and removing search hurts.

6 Experiment 2: Team Size

Holding per-agent depth fixed at 30 rounds and growing the team from 10 to 150 generalists (Table 4, Figure 4), per-idea accuracy stays in the 12–19% band on tier 1—the N=10 value rests on 2 hits and is within noise of the rest—while absolute tier-1 hits grow roughly linearly (2, 12, 34, 46). Team size buys throughput, not per-idea quality.

Table 3: Provenance of the reading that precedes each generated idea: average composition of the generating agent's 20-action lookback window (= its memory), grouped by realization outcome (tiers are cumulative; "unrealized" = no in-window tier-3 match at level ≥ 2). Pooled over all 1,328 evaluated ideas from every run in the paper.

	n	**reads**	**searches**	**search-fed**	**link-fed**	**jump-fed**	**own-idea**
tier-1 realized	237	10.57	5.85	**5.78**	**3.69**	1.02	0.08
tier-2 realized	317	10.56	5.66	5.61	3.79	1.09	0.07
tier-3 realized	888	10.68	5.28	5.18	4.12	1.29	0.09
unrealized	440	10.77	5.07	4.55	4.64	1.41	0.16

Table 4: Scaling generalist teams at fixed per-agent depth (30 rounds each, one seed). The $N{=}10$ point is the 30-round prefix of C1. Per-idea accuracy is flat; absolute hits scale with output.

N	**ideas (per agent)**	**tier-1 acc (hits) / mean**	**tier-2 acc / mean**	**tier-3 acc / mean**
10	12 (1.2)	17% (2) / 0.42	25% (3) / 0.67	75% (9) / 2.00
50	103 (2.1)	12% (12) / 0.31	17% (18) / 0.45	65% (67) / 1.77
100	180 (1.8)	19% (34) / 0.47	23% (41) / 0.57	64% (115) / 1.70
150	280 (1.9)	16% (46) / 0.38	24% (66) / 0.60	62% (173) / 1.69

The mechanism is that the agents are, in practice, independent samplers. Across all four scaling runs (575 generated ideas, 10,300 actions), **zero** citations of another agent's generated idea occurred, and only two self-citations; in the dial conditions of Sec. 5 (which run 80 rounds), 113 of 114 citations of generated nodes were *self*-citations—agents building private research agendas—and exactly one crossed agents. The stigmergic write channel is active (new nodes join the shared graph), but the read-back channel is essentially silent: 383 new nodes among 16,208 corpus nodes are simply too rare for semantic search to surface. Making teammates' contributions discoverable (e.g., recency-weighted search or frontier sampling) is, on this evidence, the necessary condition for genuine collaboration effects—and the natural next intervention.

7 Experiment 3: is the citation structure a scaffold?

If agents used the citation network as a navigational scaffold, deleting its edges should hurt. It does the opposite (Table 5): with an edgeless initial graph, the same generalist team produces 20% more ideas at nearly double the tier-1 accuracy (27% vs. 14%)— approaching the dial's $m{=}800$ peak of 31%—and starts generating at team round 8 instead of 17.

The mechanism follows from Sec. 8: agents discover papers by semantic search, not by traversing references. With edges present, `browse` surfaces each node's citation neighborhood, and agents wander along the *human* citation topology—which pulls toward older, hub-heavy regions that their own attachment behavior demonstrably avoids. Removing the edges removes the distractor: browsing degenerates to reading, navigation collapses onto semantic search, and exploration converges faster on writeable niches. Notably, the agents do not leave the graph bare: they spontaneously performed 65 curation actions, adding 53 links of their own to the edgeless network—they rebuild structure, but by their own semantic logic rather than the corpus's citation logic.

Two companion ablations remove the other channels. Disabling random jumps is neutral (14% vs. 14%)—unsurprising, since C1 agents only jumped $\sim$20 times in 800 actions. Disabling *semantic search* produces the most dramatic behavioral shift of any intervention in this paper: production quadruples (201 ideas, written from citation links, jumps, and parametric knowledge) while guarded accuracy falls to 10%, the lowest of any condition. Notably, under memory-contaminated scoring the no-search condition *appears* unharmed (24%); its hits evaporate fastest when the realization window is applied—the fingerprint of parametric recall standing in for reading. The three channels thus play three distinct roles: citation links are a distractor (removal helps), random jumps are irrelevant, and semantic search is where genuine anticipation comes from. We caution that these are

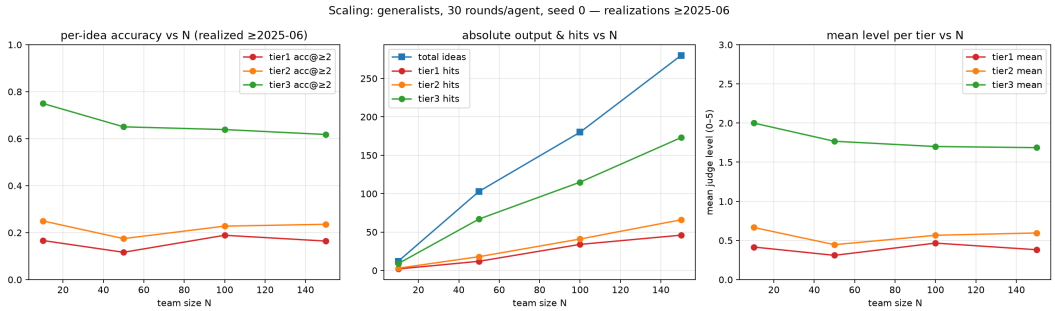

Figure 4: Team-size scaling at fixed per-agent depth. Left: per-idea accuracy is flat in N. Middle: total output and absolute hits grow linearly. Right: mean realization level per tier is likewise flat in N; stigmergic coupling (citations of teammates' generated ideas) is zero throughout (Sec. 6).

Table 5: Navigation-channel ablations: identical to C1 (10 generalists, 80 rounds, same seed and corpus) except one channel is removed—the initial citation edges, random jumps, or semantic search. The three channels play three roles: links are a distractor (removal helps), jumps are neutral, and search is the source of genuine anticipation (removal quadruples output but halves accuracy).

	ideas	**tier-1 acc (hits) / mean**	**tier-2 acc / mean**	**tier-3 acc / mean**
no edges	60	**27% (16) / 0.67**	**30% (18) / 0.73**	63% (38) / 1.63
no jumps	63	14% (9) / 0.38	17% (11) / 0.46	**71% (45) / 1.87**
no search	**201**	10% (20) / 0.27	15% (30) / 0.38	62% (125) / 1.63
C1 (full)	50	14% (7) / 0.32	16% (8) / 0.38	56% (28) / 1.46

single seeded runs per arm; the directions, the mean-level gaps, and the behavioral signatures are mutually consistent, but replication is needed before the strongest reading.

8 Analysis: How the Dynamics Rewrite the State

We compare the links written by each condition (generate citations plus curation edges) against the 80,962 human citation edges that built the corpus.

Invariant handwriting. The median semantic distance between the endpoints of agent-written links is 0.144–0.153 across all six conditions—a band of width 0.009—versus 0.179 for real citations (Figure 5). Cross-topic rates hug the corpus baseline (58–75% vs. 62%). Reading constraints move *where* an agent writes, not *how far* its links reach; the local radius of citation appears to be an intrinsic property of the LLM dynamics.

Degree-decoupled, recency-seeking attachment. Real citation edges concentrate on hubs: the median target of a corpus edge has in-degree 23, and a preferential-attachment null model (Barabási & Albert, 1999) run on the same graph (800 template nodes, citations drawn $\propto$ in-degree) reproduces this concentration (median target in-degree 17). Agent-written links sit at the opposite end of this scale: median target in-degree 0–2, against a uniform-selection baseline of 1 (mean 11.6 vs. uniform 5.0, i.e., only a mild hub tail). The reason is the discovery channel: agents find papers by semantic search, which is degree-agnostic, whereas human discovery historically flows through reference lists, the visibility cascade that produces preferential attachment (Barabási & Albert, 1999). Instead of heat, agents seek freshness: 51% of the m=800 condition's citations target papers from 2023 onward, versus 8% of real citation edges. Two growth laws, one state: the human dynamics that built the network and the LLM dynamics extending it obey different attachment rules.

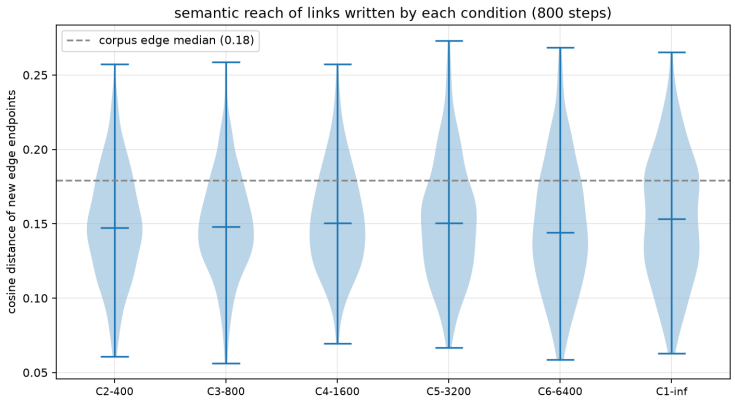

Figure 5: Semantic reach of newly written links (cosine distance between endpoints), per condition. The dynamics' "handwriting" is nearly invariant to the reading constraint, and slightly tighter than real citation edges (dashed).

9 LIMITATIONS

Single seed. Every condition rests on one seeded run (50–280 ideas each). Ideas cluster by agent thread, and a replicate of the m=1600 condition (identical topics, fresh sampling) moved tier-1 accuracy from 15% to 28%—roughly $\pm$7 points of thread-level noise per run; we report the replicate, and release both. The narrow-band-vs-broad contrast and the floor comparisons exceed this noise band; the fine structure within each band does not, and the exact peak location needs replication across seeds (and topic draws) before strong claims. The clearest findings (scoped $>$ unrestricted on all tiers; monotone production; flat scaling; zero coupling) are consistent across many runs and both experiments.

One agent model, LLM judge. Dynamics are those of a single model (`gpt-5`); the judge is a smaller LLM whose six-level rubric we audited but did not calibrate against human experts at scale. The realization window (post-2025-06) is young; scores will only grow as the literature fills in. The window is set by behavioral probes rather than by a guarantee from the model provider, and a paper whose official publication date is in-window may have circulated as a preprint earlier; both residual risks are documented in Appendix A.10.

Scope mechanism. Equal-mass regions leak across topics by design (Sec. 3.3); we quantified purity rather than enforcing single-topic scopes, and treat this as the definition of the dial rather than a confound.

10 CONCLUSION

Framing machine-assisted innovation as a dynamical system on an idea network lets us intervene on the coupling between researcher and knowledge structure directly. On this substrate, under an evaluation guarded against the agent model's post-cutoff parametric memory, the operative lever on idea quality is *how reading is situated*—a sub-area scope of $\sim$800 papers beats both narrower and broader diets, and any scope beats none—while team size sets only throughput, because today's agents do not yet read each other. The human citation structure itself, scaffold of every account of cumulative science, is at best inert and at worst a distractor for these semantically-navigating dynamics: removing it outright improved both output and accuracy. The immediate frontier is switching on the silent agent-to-agent channel—making contributions discoverable to teammates—and measuring whether the scaling curve bends.

AI USE STATEMENT

This work studies LLM agents, and generative AI tools were used in producing it; we disclose all uses here per the ICLR 2027 AI Policy for Authors.

Research ideation and execution. The research questions, experimental designs, evaluation criteria, and every design decision (including the mid-course redesigns reported in the paper) were specified by the human authors. An LLM-based coding agent, operating under continuous human direction, implemented the simulation environment, evaluation pipeline, and analysis code, executed and monitored the runs, and proposed analyses that the authors accepted, rejected, or redirected.

Writing. The manuscript was drafted with LLM assistance from the authors' experimental records and instructions, and LLM assistance was also used to polish prose. The authors reviewed and edited all text.

Retrieval and discovery. LLM assistance was used in locating and organizing related work; all citations were verified by the authors against the cited sources.

As the object of study. Every "idea" analyzed in this paper is by design LLM-generated (`gpt-5`), and the evaluation judge, corpus distiller, and topic namer are LLMs (`gpt-5-mini`); these are the experimental apparatus, not authoring aids, and are documented in Sections 3–4 and the appendices. Generative AI was *not* used to prove mathematical claims.

We have reviewed all AI-assisted work: agent-generated ideas were scored only against independently published papers via the audited pipeline of Section 4, code was reviewed by the authors and is covered by the released test suite, and every number reported in the paper is reproducible from the released event logs and verdicts. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI. LLMs are not authors of this paper.

Ethics statement

This work involves no human subjects and no sensitive data; all corpus material derives from public scholarly metadata (titles, abstracts, citations) accessed through the Semantic Scholar and OpenAlex APIs within their terms of use. The broader practice this paper studies — automated generation of research ideas — raises questions about scientific credit and about flooding venues with machine-generated submissions; our evaluation protocol deliberately scores ideas only by later independent human realization, and we do not submit, publish, or disseminate any agent-generated idea as a research claim.

Reproducibility statement

All runs are driven by seeded configurations; every agent action and its result is logged as an event stream from which the full network state can be replayed. Evaluation verdicts record every candidate paper, judge level, and exclusion reason. Corpus construction uses the public Semantic Scholar and OpenAlex APIs. Code, configurations, the complete event logs of every reported run, and all evaluation verdicts are available anonymously at `https://anonymous.4open.science/r/structure-and-dynamics-of-innovation-D607/`.

References

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999.

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

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. The Semantic Scholar Open Data Platform. *arXiv preprint arXiv:2301.10140*, 2023.

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

Jason Priem, Heather Piwowar, and Richard Orr. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022.

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*, 2025.

Guy Theraulaz and Eric Bonabeau. A brief history of stigmergy. *Artificial Life*, 5(2):97–116, 1999.

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

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Lingfei Wu, Dashun Wang, and James A Evans. Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744):378–382, 2019.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-Pack: Packaged resources to advance general Chinese embedding. *arXiv preprint arXiv:2309.07597*, 2023.

A Appendix

A.1 Corpus construction and idea distillation

Papers are retrieved by bulk venue search on the Semantic Scholar Graph API: all NeurIPS, ICLR, ICML, and AAAI papers with publication dates between 2020-01-01 and 2024-08-31 and at least 25 citations at collection time (16,208 papers). Citation edges among corpus members are recovered by a three-arm pipeline: (i) each paper's reference list; (ii) each paper's *citing* list, which recovers edges whose source reference list was unavailable; (iii) OpenAlex works matched by MAG id, DOI, or arXiv-DataCite DOI. This reduced isolated nodes from 49.7% (references arm alone) to 6.4%.

Each paper's title and abstract are distilled into an idea node by `gpt-5-mini` with the fixed template:

> *Summarize the following paper as ONE paragraph of 3–4 sentences covering: (1) the problem it addresses, (2) the key insight, (3) the method. Write only the paragraph, no preamble.*

Median idea length is $\sim$930 characters. Ideas are embedded with bge-small-en-v1.5 (Xiao et al., 2023) (384-d, unit-normalized); all similarities are cosine.

A.2 Agent system prompt

Every agent shares the system prompt (verbatim):

> *You are a research agent exploring a network of research ideas distilled from published papers. Ideas cite the ideas they build on. Your goal is to find promising unexplored directions and, when you see one, contribute a genuinely new idea to the network. Prefer exploring (search, browse) until you understand a neighborhood well enough that your new idea is specific and well-grounded. If you notice an idea clearly builds on another idea it does not yet reference, you may record that missing link. Some of your actions may be restricted to a region of the network; if an action returns a restriction error, adapt your strategy instead of repeating it.*

Each turn's user message consists of an identity header (`[agent run_id:agent_id]`), the action documentation, and the rolling memory, ending with "Choose your next action (JSON only)". The identity header also keys the LLM response cache, so identical memory states of different agents or runs never share a cached reply.

A.3 Topic pool and equal-mass scopes

The 50-topic pool is built by clustering the corpus embeddings into 50 groups and having an LLM name each cluster from its central papers (e.g., "backdoor attacks and defenses in ml models", "efficient and long-range transformer architectures"). A specialist's scope anchors on its topic name's

embedding; the region radius is set per anchor to the cosine distance of the m-th nearest corpus paper, so every scope holds exactly m corpus papers regardless of how dense its neighborhood is. Natural topic sizes (nearest-anchor partition) range 44–1096 (median 228), which is why even the m=400 region mixes topics; measured own-topic purity of the m=800 regions used in the dial is 5–60% (median ∼35%). Newly generated nodes classify into scopes by the same rule applied to their own embeddings.

A.4 Evaluation pipeline details

For each idea, `gpt-5-mini` extracts three search queries (problem-focused, method-focused, combined). Each query retrieves the top 10 results from the Semantic Scholar paper search API and the top 10 from OpenAlex; per-query results are truncated to the top 5 per channel, deduplicated by title. Every candidate is graded by the judge (`gpt-5-mini`, medium reasoning) with the rubric of Sec. A.9, and then gated in order: (i) title matches a corpus member → excluded (contamination guard); (ii) no publication date → excluded; (iii) published on or before 2024-09-30 → excluded (anticipation only). Surviving candidates are assigned tiers: tier 1 if the venue matches one of 60 recognized top venues (the 58 CCF-A venues plus COLM and RLC, matched by alias lists) or citations ≥ 50; tier 2 additionally if a CCF-B venue or citations ≥ 10; tier 3 for any published venue. Scoring is cumulative: an idea's tier-2 best is the maximum level over tier-1 and tier-2 candidates. Ideas whose embedding matches any corpus idea at cosine ≥ 0.95 are floored to level 0 (near-duplicate guard); no idea in any reported run triggered it. All search responses and judge replies are cached on disk; both search channels degrade independently per query on outage, and failed queries are uncached so re-evaluation refetches only gaps.

A.5 Run protocol and hyperparameters

agent model	`gpt-5`, medium reasoning, 2,000-token visible output
judge / distiller / namer	`gpt-5-mini`
embedding	bge-small-en-v1.5 (384-d)
agent memory	FIFO, 20 (action, result) pairs
search k / eval queries / eval top-k	5 / 3 / 5 per channel
scheduling	round-robin; 1 round = each agent acts once
dial runs	N=10, 80 rounds, seed 0, identical topic draws
scaling runs	30 rounds per agent, seed 0
generation budget	none (idea count is an outcome variable)
training cutoff / anticipation window	2024-09-30 / publications after it

Every action and result is appended to a JSONL event log from which the full network state, agent memories, and random streams can be replayed; runs can be extended from their logs, and all figures and tables derive from the logs plus cached evaluation verdicts.

A.6 Example trajectory map

Figure 6 shows where the ten m=800 specialists read and where they wrote, on a fixed t-SNE map of the corpus. Writing is never restricted by scope, so the fact that the stars fall inside or beside the tinted read regions is a property of the dynamics rather than of the environment.

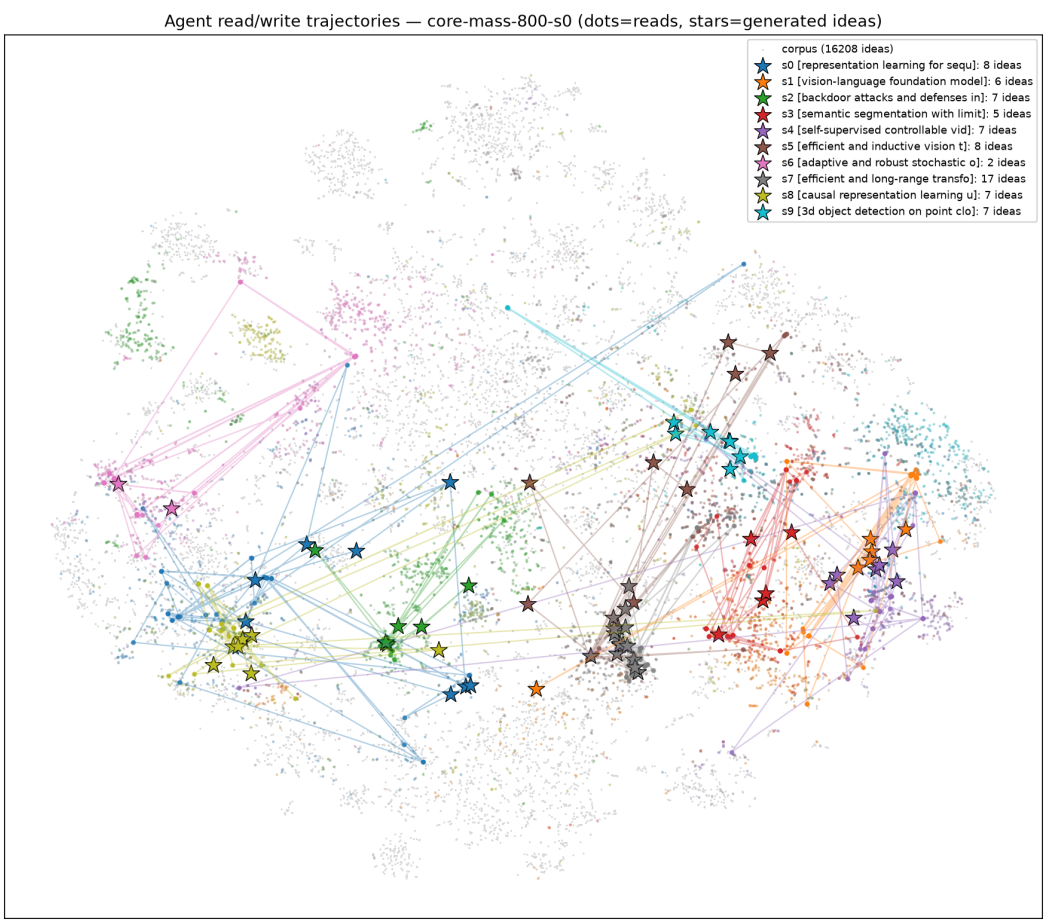

Figure 6: Read/write trajectories of the ten specialists in the m=800 condition on the fixed t-SNE map of the corpus. Tinted areas are each agent's readable region; dots are reads, stars are generated ideas. Although writing is unconstrained, specialists write inside or near the regions they read.

A.7 Agent action interface

Each agent interacts with the environment exclusively through the following six actions, chosen one per turn by emitting a single JSON object ({`"action": ..., "args": {...}`}); malformed replies fall back to `sample_frontier`. Every action returns a JSON result—including scope-restriction errors, which the agent is instructed to adapt to rather than retry.

- `search(query, k=5)` — embed the query and return the top-k nearest ideas *within the agent's readable scope* (over-fetch then filter), each as node id, 300-character preview, and similarity score. This is the agents' primary discovery channel.
- `browse(node_id)` — return the node's full idea text plus up to 10 outgoing (cited) and 10 incoming (citing) neighbors, each with a 200-character preview; neighbors outside the readable scope are filtered out. On an edgeless graph (Sec. 7) this degenerates to reading the node alone.
- `sample_frontier()` — jump to one node drawn uniformly at random from the agent's readable scope; the serendipity channel. Disabled in the jump ablation (`allow_jump=false`), in which case it returns a restriction error.
- `generate(text, cited_ids)` — add a new idea node (a 3–4 sentence paragraph) citing the existing ideas it builds on. Citation admissibility follows the *read* scope (citing means having read): out-of-scope citations are dropped with notice, and the generate pro-

ceeds with the rest. The new node and its edges join the shared graph immediately and permanently.

- `add_links(src_id, dst_ids)` — record missing reference links from an existing idea to ideas it builds on (curation).
- `remove_links(src_id, dst_ids)` — remove reference links that do not actually support the source idea.

Each agent carries a FIFO memory of its last 20 (action, result) pairs, rendered into every prompt together with a stable identity header (run and agent id) and the action documentation; there is no other communication channel between agents. Visible output is capped at 2,000 tokens per turn (reasoning tokens budgeted separately).

A.8 Evaluation prompts

Query extraction (`gpt-5-mini`, verbatim template; {n}=3):

> *Give {n} short keyword search queries (4–8 words each) for finding papers related to this research idea: the FIRST focused on the PROBLEM it addresses, the SECOND on its core METHOD, the THIRD combining both. One query per line, no numbering, nothing else.*
> *Idea: {idea}*

Judge (`gpt-5-mini`, verbatim template; system prompt: "You grade how closely a candidate paper realizes a proposed research idea."):

> *Research idea: {idea}*
> *Candidate paper: Title: {title} Abstract: {abstract}*
> *Grade how closely the paper realizes the idea on this scale:*
> *0 = different problem entirely*
> *1 = same broad area/task, but a different goal*
> *2 = same specific problem/goal, unrelated approach*
> *3 = same problem AND the approaches belong to the same method family*
> *4 = same problem AND the same core mechanism, differing in secondary components*
> *5 = essentially the same idea: problem, core mechanism, and key design choices match*
> *Reply with JSON only: {"level": ¡0-5¿, "evidence": "¡one short sentence¿"}*

A.9 Judge rubric

Level 0: different problem entirely. 1: same broad area, different goal. 2: same specific problem, unrelated approach. 3: same problem, approaches in the same method family. 4: same problem and core mechanism, differing in secondary components. 5: essentially the same idea (problem, mechanism, and key design choices).

A.10 Choosing the realization window: memory probes

The agent model's official knowledge cutoff is 2024-09-30, but official cutoffs are not hard data boundaries: post-training continues until release (August 2025 for our agent model), and any "anticipation" realized by a paper the model has memorized is recall, not foresight. To locate the actual boundary of the model's knowledge we ran three probe batteries against the exact model configuration used by our agents (no tools, no network access), asking about verifiable events binned by when they happened: **A** = 2024-10 to 2024-12, **B** = 2025-01 to 2025-04, **C** = 2025-05 to 2025-08 (up to the model's release), and **D** = 2025-09 onward—events *after* release, which the model cannot know and which therefore measure its guessing ceiling.

	A 24-10..12	**B** 25-01..04	**C** 25-05..08	**D** $\geq$25-09 (control)
yes/no, AI events	1/3	**3/3**	1/4	0/5
4-choice, general events	3/5	1/5	0/5	1/5
4-choice, AI events	3/4	**4/5**	3/4	3/5

Three observations fix the boundary. (i) *Forced choice reveals hidden knowledge*: in free-form probing the model claims ignorance of events (o3, Claude 3.7, the s1 paper) that it answers correctly

under forced choice, so cautious "I don't know" responses cannot be taken as absence of memory. (ii) *Memory is strong through April 2025 in the AI domain*: bin B AI questions are answered near-perfectly in both formats, including content questions that cannot be guessed (the mechanism of the s1 paper, Llama 4's architecture, DeepSeek-R1's training method). General-news knowledge fades much earlier—the post-training refresh is domain-selective, and skewed toward exactly the domain our agents generate in. (iii) *Bin C does not exceed the guessing ceiling*: the control bin D shows that multiple-choice questions about product-release *dates* are guessable at well above chance (3/5 despite being unknowable), and bin C's correct answers are almost exclusively of that style; on yes/no probes and free recall, bin C performance collapses (1/4), and the fabricated-fact controls draw zero false yeses. Knowledge in May–August 2025 is thus indistinguishable from informed guessing.

We therefore set the realization window at **2025-06-01**: the earliest month at which no probe format shows above-guessing knowledge. Without this window, contamination is material—64% of face-value tier-1 hits fall in 2024-10 to 2025-08, and the no-search ablation, which forces agents onto parametric knowledge, loses its hits fastest when the window is applied (Sec. 7). The topic-only condition C0 closes the loop behaviorally: generating from parametric knowledge alone scores 20% at face value but 13% under the window—the residue is the model's genuine ability to propose realizable ideas from priors, and it is exactly this floor that situated reading must (and, when narrow, does) exceed. Two residual risks remain: the window is behavioral evidence, not a provider guarantee, and a paper published in-window may have circulated as a preprint earlier; both bias reported accuracies in opposite directions and neither affects between-condition comparisons, which all use the same window.

A.11 Sample generated ideas

Five verbatim ideas illustrating the range of agent output (realization levels are in-window, tier-3 unless noted; the released verdicts contain every idea with its full audit trail).

C4 (m=800), realized at level 4.

> *We propose a preconditioned kernel-Hopfield in-context learner that unifies linear-time attention with the associative-memory view to implement curvature-aware, one- to few-step optimization from context. A Transformer block with linear attention learns task-adaptive kernel features and uses an MLP to infer a low-rank preconditioner and prior mean from the prompt, so the associative KV accumulators perform a preconditioned GD-β update toward the Bayesian posterior; with non-linear kernels, this becomes a functional-gradient step while retaining linear-time complexity. A meta-objective trains the block to minimize posterior risk over task families, while the Hopfield energy perspective provides a convergence regularizer and layerwise step-size control. This design improves stability on long contexts and sample efficiency over plain linear-attention ICL, and extends naturally to non-linear tasks via learned kernels.*

Matched (level 4): *Transformers Can Implement Preconditioned Richardson Iteration for In-Context Gaussian Kernel Regression* (arXiv, May 2026).

C1 (generalist), realized at level 3 in a tier-1 venue.

> *I propose latency-targeted threshold learning: insert layer-wise learnable pruning thresholds and jointly optimize weights and thresholds with a differentiable latency objective built from device-specific per-layer LUTs. A Lagrangian enforces a user-specified speedup during training, while structured sparsity (e.g., channel and GPU 2:4) is induced via relaxed top-k/straight-through mask parametrizations and then discretized at the end to preserve measured latency. This unifies speed-constrained allocation with end-to-end sparse training, aiming to deliver the hardware-guaranteed speedups of post-training solvers without the accuracy loss from late pruning. Empirically, this should match or surpass post-hoc speed-aware pruning at equivalent targets while eliminating extra retraining passes.*

Matched (level 3): *LightPrune: Latency-Aware Structured Pruning for Efficient Deep Inference on Embedded Devices* (ICCV, October 2025).

C4 (m=800), not (yet) realized.

> *We propose RKHS-Hopfield Linear ICL: parameterize the attention kernel feature map so that linearized attention implements a proximal/gradient step in a reproducing-kernel Hilbert space tailored to the task family, thereby shaping the Hopfield energy landscape to encode desirable attractor basins for in-context learning. Concretely, we learn kernel features and value projections so a single attention update matches a functional gradient descent step on the implicit RKHS loss given the context, with multi-heads covering different step sizes or preconditioners. This yields interpretable, stable associative retrieval aligned with gradient-based learning while retaining linear-time, streaming updates for long contexts. We expect improved non-linear ICL and better extrapolation by aligning the learned kernel with the task NTK and training via a meta-objective that penalizes deviation from the target functional update.*

C0 (topic name only), not realized.

> *Current self-supervised video representations entangle content and dynamics, making them hard to steer along meaningful control dimensions like pose, speed, or viewpoint without labels. Our key insight is to treat controllability as learning a latent operator algebra: mine naturally occurring near-duplicate video segments where only one factor changes, and use their residual differences as pseudo-interventions that reveal control directions. We propose Latent Control Algebra, which retrieves such pairs/triples, estimates dense correspondences to isolate residual changes, factorizes them into a small set of latent generators, and trains an encoder–decoder where content codes plus exponentiated generators reconstruct targets under cycle/inverse consistency and commutator regularization for disentanglement; rolling out generators must also predict short-horizon futures. This yields self-supervised, composable control axes that enable latent steering, controllable editing, and zero-shot policy transfer in downstream tasks.*

No-search ablation, not realized.

> *Adaptive Bias–Variance Compression (ABVC) for federated optimization tunes, per client and round, the bias–variance operating point of a generalized compressor to minimize bits-to-accuracy under heterogeneity and partial participation. ABVC combines EF-BV's error feedback with variance reduction via gradient-difference tracking in the MARINA style, and augments updates with a lightweight Hessian-learning module to form diagonal/low-rank preconditioners that set client-specific compression budgets; once the preconditioner stabilizes, we enable Nesterov-style acceleration as in ACGD. We hypothesize linear convergence under PL/strongly convex objectives with rates that smoothly interpolate between unbiased-VR and biased-contractive regimes, and improved scaling with the number of workers due to independent randomness and preconditioned error dynamics. Empirically, on vision/NLP FL benchmarks with top-k and low-rank compressors, ABVC should reduce communication at fixed accuracy and time-to-ε versus EF-BV, MARINA, and FedNL variants.*

A.12 Case studies: anticipations and their realizations

Case 1 (realization level 5). Agent g11 of the $N{=}50$ scaling run proposed:

> *We propose Conformal Early-Stopping Self-Consistency (CESS), a black-box, distribution-free procedure that turns self-consistency sampling into a risk-controlled stopping rule with finite-sample guarantees. On a calibration set, we run deep self-consistency to learn a mapping from the trajectory of vote-share (and simple diversity features) to correctness, then derive nondecreasing stopping boundaries via conformal calibration so that when the live vote-trajectory crosses the boundary we stop and output the modal answer with $1-\alpha$ coverage; if the boundary is not reached within a budget, we abstain. [...]*

The judge matched it at level 5 to *Black-Box Reliability Certification for AI Agents via Self-Consistency Sampling and Conformal Calibration* (arXiv, February 2026), with the evidence: "the paper implements black-box self-consistency sampling combined with conformal calibration to produce sequential early-stopping rules with finite-sample guarantees."

Case 2 (tier 1, level 4). Agent g37 of the same run proposed *Path-of-Thought Distillation for text-rich graphs*: prompt an LLM with ordered node/edge texts along candidate evidence paths to generate structured relational rationales, then distill those rationales into a hybrid tokenized graph transformer with attention-alignment and counterfactual edge-text robustness losses, so that test-time inference requires no LLM. The judge matched it at level 4 to *Mitigating Adversarial Attacks*

by Transferring LLM-generated Narrative Reasoning for Robust Fake News Detection (SIGIR 2026, a tier-1 venue): both distill LLM-generated reasoning into graph models for robust graph tasks with no runtime LLM, differing in the target domain.

A.13 THE 50-TOPIC POOL

adaptive and robust stochastic optimization methods; efficient and inductive vision transformers; inductive and contextual knowledge graph representation; 3d object detection on point clouds; graph structure learning and generalization; evaluation and augmentation of language models; single-image to 3d generative modeling; continual learning and lifelong adaptation; causal representation learning under interventions; sparse neural network training and pruning; differentially private learning and optimization; self-supervised controllable video representation learning; task-oriented and knowledge-grounded dialogue modeling; efficient and long-range transformer architectures; adversarial robustness and training methods; offline and constrained reinforcement learning with pessimism; accelerated sampling and analysis of diffusion models; 3d generative models for molecular design; federated learning optimization and heterogeneity; learning-based combinatorial optimization methods; efficient model compression and adaptation for llms; semantic segmentation with limited supervision; foundation models and generative modeling for time series; multi-view graph-based clustering methods; language model alignment and calibration; efficient transfer and adaptation of models; backdoor attacks and defenses in ml models; out-of-distribution generalization and robustness; multi-step mathematical reasoning with lm prompts; evaluation and improvement of code generation models; retrieval-augmented multimodal question answering; contrastive representation learning and robustness; neural operator and physics-informed pde solvers; individual fairness and robustness in ML; representation learning for sequential decision making; multimodal representation learning for missing modalities; implicit bias and dynamics of deep learning; multimodal generative and robust audio models; language-models as autonomous decision agents; spatio-temporal traffic and trajectory modeling; vision-language foundation models and benchmarks; recommendation systems and sequential user modeling; offline reinforcement learning methods and theory; robust semi-supervised and noisy-label learning; interpretable and causally-grounded model explanations; few-shot and domain-adaptive object detection; high-resolution diffusion-based image synthesis; online learning and multiagent decision-making; out-of-distribution detection and outlier synthesis; embodied multimodal world-models for agents.

A.14 RECOGNIZED VENUES (TIER 1)

Tier-1 recognition matches 60 venues by alias lists: the 58 venues rated A by the CCF (2026 edition) across all categories, plus COLM and RLC. The full list:

AAAI Conference on Artificial Intelligence; Neural Information Processing Systems; Annual Meeting of the Association for Computational Linguistics; Computer Vision and Pattern Recognition; IEEE International Conference on Computer Vision; International Conference on Machine Learning; International Conference on Learning Representations; Knowledge Discovery and Data Mining; International ACM SIGIR Conference on Research and Development in Information Retrieval; ACM SIGMOD International Conference on Management of Data; IEEE International Conference on Data Engineering; Very Large Data Bases; The Web Conference; ACM Multimedia; Conference on Language Modeling; Reinforcement Learning Conference; ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming; USENIX Conference on File and Storage Technologies; Design Automation Conference; IEEE International Symposium on High Performance Computer Architecture; IEEE/ACM International Symposium on Microarchitecture; International Conference for High Performance Computing, Networking, Storage and Analysis; International Conference on Architectural Support for Programming Languages and Operating Systems; International Symposium on Computer Architecture; USENIX Annual Technical Conference; European Conference on Computer Systems; International Symposium on High-Performance Parallel and Distributed Computing; ACM SIGCOMM Conference; ACM International Conference on Mobile Computing and Networking; IEEE International Conference on Computer Communications; Symposium on Networked Systems Design and Implementation; ACM Conference on Computer and Communications Security; International Conference on the Theory and Applications of Cryptographic Techniques; IEEE Symposium on Security and Privacy; International Cryptology Conference; USENIX Security Symposium; Network and Distributed System Security Symposium; ACM SIGPLAN Conference on Programming Language Design and Implementation; ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages; ACM International Conference on the Foundations of Software Engineering; ACM Symposium on Operating Systems Principles; Conference on Object-Oriented Programming Systems, Languages, and Applications; International Conference on Automated Software Engineering; International Conference on Software Engineering; International Symposium on Software Testing and Analysis;

USENIX Symposium on Operating Systems Design and Implementation; International Symposium on Formal Methods; ACM Symposium on the Theory of Computing; ACM-SIAM Symposium on Discrete Algorithms; International Conference on Computer Aided Verification; IEEE Annual Symposium on Foundations of Computer Science; ACM/IEEE Symposium on Logic in Computer Science; ACM SIGGRAPH; IEEE Conference on Virtual Reality and 3D User Interfaces; IEEE Visualization Conference; ACM Conference on Computer Supported Cooperative Work and Social Computing; ACM Conference on Human Factors in Computing Systems; ACM International Joint Conference on Pervasive and Ubiquitous Computing; ACM Symposium on User Interface Software and Technology; IEEE Real-Time Systems Symposium.

Tier-2 additionally matches the $\sim$90 CCF-B venue aliases; the exact alias lists (including abbreviation variants used for matching) ship with the released configurations.

A.15 Compute and cost

All experiments ran on commodity hardware; the only significant cost is LLM API usage. Approximate totals: corpus distillation (16,208 summaries, `gpt-5-mini`) $\sim$\$15; the six dial runs (4,800 `gpt-5` actions) $\sim$\$90–150; the scaling runs (9,300 actions) $\sim$\$170–280; ablations (1,600 actions) $\sim$\$30–60; evaluation (judging $\sim$1,000 ideas with `gpt-5-mini`, searches free) $\sim$\$40–80. Total on the order of \$350–600. A single dial-style run (10 agents, 80 rounds) costs $\sim$\$15–25 and $\sim$1 hour wall clock; evaluation of its ideas a further $\sim$\$5–10.

A.16 Realization-level distributions

Figure 7 breaks every dial condition down by best realization level, rather than by the acc@$\geq$2 threshold reported in the main text, with one panel per recognition tier.

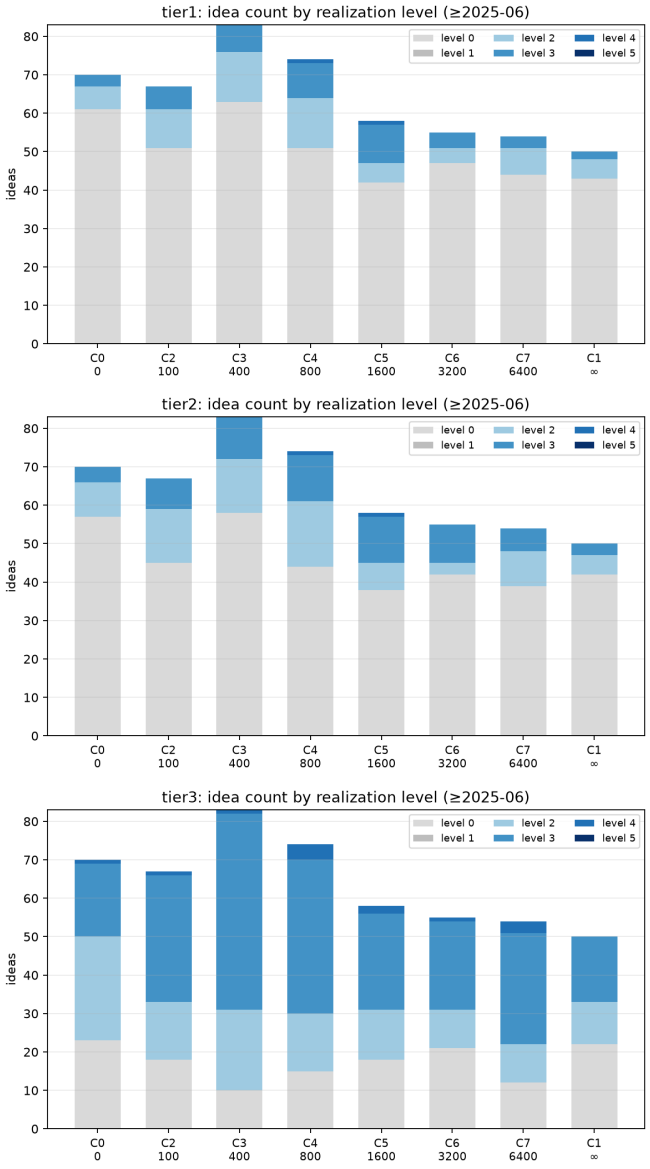

Figure 7: Idea counts by best realization level for each dial condition, one panel per recognition tier (stacked bars; bar height = total ideas generated). Level 1 ("same broad area, different goal") is never the best match in any tier—the judge's effective scale is 0/2/3/4/5.

A.17 Behavioral observations

Onset. Median personal round of an agent's first `generate` action: 21–24 across conditions (minimum 9); the first 40 rounds of a team are dominated by `search`/`browse`. This is why the scaling experiment fixes per-agent depth at 30 rounds: shallower budgets produce (near-)zero output for any team size.

Error-driven adaptation, and its memory horizon. Constraints are never described in the prompt; agents discover them through error results. The jump ablation shows this cleanly: at cold start, 9 of 10 agents' very first action was `sample_frontier` (their standard opening move), each received the restriction error, and all switched to `search`-based cold starts immediately. Across the remaining 790 actions only three further jump attempts occurred (team steps 400, 455, 719)—each

long after the original error had been evicted from the agent's 20-entry FIFO memory. The rolling memory length thus sets the time scale on which learned constraints are forgotten.