Autonomous AI Research and Automated Paper Writing:

Systems Taxonomy, Open-Source Frameworks, and Empirical Failure Modes Across 5,932 Controlled Runs

Shanglin Yang
Learn LLM Research
learn-llm.com · github.com/shangliy

November 2026 (Data-Chained Preprint Edition)

Abstract

Between 2024 and late 2026, autonomous scientific research (*AI Auto-Research*) expanded from fixed-template scripts (`The AI Scientist v1`) into an ecosystem of over fifty multi-agent discovery pipelines, survey compilers, code-space tree-search engines, and automated peer-review models. Evaluating this stack requires separating *generating a plausible-looking manuscript* from *producing true, non-obvious scientific knowledge*. We survey 50+ papers, open-source repositories, and benchmarks across the six-stage research lifecycle, and audit them against **5,932 controlled empirical runs** spanning four studies: (1) `ideation-goodhart` (2,640 runs across enumerable `NAS-Bench-201`, `HPOBench`, and private CTR hypothesis spaces), (2) *Lineage Is Memory* (1,750 preregistered runs across nine prompt and code optimization tasks), (3) `kvleak` (945 priced Claude Code transcripts reconciled with dedicated-GPU context compaction), and (4) two new supplementary audits reported here—597 multi-hop agent republication documents (2,894 blocks across three model tiers) testing *citogenesis*, plus scope-typed *Data-Chaining* on an 80-venue OpenStreetMap dataset. Five findings emerge: (i) optimizing an LLM judge for ideation raises judge scores while collapsing ground-truth precision from 87% to 28% (40% → 3% on private CTR transfer); (ii) greedy parent inheritance ($\Delta_{\text{lineage}} = +0.238$) outweighs full-set error clustering and proposer scaling, catching oracle clustering by Round 4 at half the tokens; (iii) self-improving harnesses suffer from trajectory–validity decoupling and public rubric leakage unless gated by network-isolated private splits; (iv) form-only citation validation fails to stop opinion-to-fact citogenesis (0.50 Hop-4 fact share on all-opinion seeds, scaling from 0.09 to 0.22 upgrades/hop with model size), whereas our *origin-aware primary-source gate* eliminates 100% of pure citogenesis (0.00) while retaining 94.1% of primary-grounded claims; and (v) cutting raw tokens by 39%–60% via premature `/compact` raises actual API dollar bills by +14% to +36% under prompt caching. Every quantitative statistic in this manuscript is programmatically compiled from run logs via scope-checked *Data-Chaining* macros.

1 Introduction and Lifecycle Taxonomy

Autonomous scientific research systems aim to close the loop from literature comprehension and hypothesis generation to code implementation, experiment execution, and manuscript preparation [21, 29, 31, 36, 45, 53]. Open-source frameworks released between 2024 and 2026 can compile formatted eight-page LaTeX manuscripts for $2–$20 in API calls, and individual agent-generated papers have passed workshop or conference peer review [53, 60].

Yet surface plausibility is an unreliable proxy for scientific progress. Most claims in the autonomous-research literature rely on three fragile evaluation signals: (i) LLM-as-a-judge novelty or review scores [39, 52], (ii) unisolated container rollouts on public benchmarks [5], or (iii) raw token counts rather than billed dollar costs [56]. When an optimizer loops over any of these signals without executable ground truth, Goodhart's law dominates.

This paper makes three contributions:

1. **Comprehensive Six-Stage Ecosystem Taxonomy:** We organize 50+ papers, open-source frameworks, and evaluation benchmarks (2023–2026) across six stages: (1) literature synthesis, (2) ideation, (3) code/experiment search, (4) self-improving harnesses, (5) manuscript compilation, and (6) peer review and reproduction (Tables 1–3).

2. **Controlled Empirical Audits Across 5,932 Runs:** We synthesize controlled experiments spanning `ideation-goodhart` (2,640 calls) [55], *Lineage Is Memory* (1,750 runs) [57], and `kvleak` (945 priced transcripts) [56], alongside two new supplementary experiments: a 597-document (2,894-block) multi-hop provenance audit quantifying *citogenesis* and testing an *origin-aware primary-source gate*, plus scope-typed *Data-Chaining* verification on an 80-venue OpenStreetMap corpus.

3. **Executable Engineering Blueprint:** Practicing our own recommendations, this preprint is compiled via programmatic *Data-Chaining* (`code/run_supplementary_audit.py` and `code/test_invariants.py`), binding every empirical metric macro directly to SHA-256 verified logs with compile-time scope invariants.

1.1 Stage-Specialized Modules and Benchmarks

Rather than running monolithic topic-to-PDF scripts, engineering teams increasingly compose stage-isolated modules (Table 2) and evaluate them on executable benchmarks (Table 3). Across all 12 benchmarks in Table 3, a sharp divide separates *text-output benchmarks*—where fluent prose masks $> 50\%$ analytical rubric failures on `DeepResearch Bench II` [4]—from *from-scratch executable reproduction* such as `PaperBench` [43], where early frontier runs scored $\sim 21\%$ across 8,316 author-verified grading nodes.

2 Stage I: Scientific Ideation, Shared Priors, and Judge Goodhart

In a blind evaluation by 100+ NLP researchers over 4,000 proposals, Si et al. [39] showed that human reviewers rate LLM-generated ideas higher on novelty than human ideas ($p < 0.05$). Subsequent closed-loop experiments exposed why perceived novelty diverges from scientific truth: training ideation models with execution RL collapses diversity toward conservative tweaks [40], and `HindSight` [6] found a negative correlation ($\rho = -0.29, p < .01$) between LLM-judged novelty and post-cutoff scientific impact.

Even structured RL frameworks illustrate this boundary. `IdeaScientist` [7] decomposes ideation into a gap finder, cross-domain innovator, and report writer on `Qwen3.6-27B` across 277 post-cutoff test papers. While its headline score rises by 14.0 points over `DeepScientist` ($60.6 \rightarrow 74.6$), the untrained stage-decomposed harness (`IdeaScientist-Base`) accounts for +12.9 points (73.5), whereas asynchronous GRPO LoRA training adds +1.1 points (statistically indistinguishable from Base in 50-paper blind human evaluation), and pairwise preference rewards collapsed into verbosity.

2.1 Controlled Audit: `ideation-goodhart` (2,640 Runs)

To test ideation optimization against exact ground truth, `ideation-goodhart` [55] constructed fully enumerable hypothesis spaces: (1) `NAS-Bench-201` (`NB201`, 1,740 directional operator-interaction claims across 15,625 architecture cells; 198 true claims with sign consistency across CIFAR-10/100 and ImageNet16-120 and $|\Delta| \geq 2.0$ pp), (2) `HPOBench` (296 tabular MLP hyperparameter claims; 41 true), and (3) an unseen private 6-dataset industrial CTR configuration benchmark (430 claims; 21 true) for zero-shot transfer. Given leave-one-model-out shared prior $\bar{P}_{-m}(c) \in [0, 100]$ across `Haiku`, `Sonnet`, and `Opus`, a claim is *Surprising-Valid* (SV) iff:

$$\mathrm{SV}(c) = \mathbb{I}\big[\mathrm{Valid}_{\mathrm{GT}}(c) = 1 \ \wedge \ \bar{P}_{-m}(c) < 50\big]. \quad (1)$$

Across 2,640 CLI runs over 12 pre-registered stages, four empirical findings hold:

1. **Shared Prior Ceiling:** Inter-model agreement across `Opus`, `Haiku`, and `GPT` families is ≈ 0.50, versus ≈ 0.20 agreement with ground truth. Introspective self-challenge flips 20% of `Sonnet`'s judgments, of which only 42% were originally wrong (matching its 41% base error rate).

2. **Judge Goodhart Collapse (Table 4):** Optimizing an ideation harness to maximize an `Opus` judge raises judge scores ($7.5 \rightarrow 8.2$ on `Sonnet`; $8.2 \rightarrow 8.9$ on `Opus`, $p < .001$) while collapsing ground-truth precision from 79% to 43% on `Sonnet`, from 87% to 28% on `Opus`, and from 40% $\rightarrow$ 3% on zero-shot private CTR transfer. Inspection of the evolved prompts shows the optimizer explicitly bans common, true mechanism classes that appear "obvious" to the judge.

3. **Pure-Reasoning Frontier Invariance:** Across 19 pure-reasoning harnesses (L_9 orthogonal array of retrieval, role count, flaw/negate challenge, and selection rules, plus designing 100 experiments without reading outcomes), *none* significantly beat single-call direct generation (SV = 1.25 at 5K tokens vs. 1.00–1.50 at 22K–37K tokens; cross-domain retrieval $p = .73$).

4. **The Stage-12 Binary-Complement Measurement Trap:** In Stages 5–9, letting `Opus` design and read 100 real experiments appeared to double `NB201` SV ($1.08 \rightarrow 2.12$, Holm $p = .003$). Stage 12 uncovered an instrumentation artifact: asking the prior panel only $P(\textsc{higher})$ and setting $P(\textsc{lower}) = 100 - P(\textsc{higher})$ folded $P(\textsc{no_interaction})$ into LOWER, inflating LOWER priors so 0% of SV candidates were LOWER (vs. 54–57% of true claims). Re-evaluating with a symmetric 3-way prior ($P(\textsc{higher}), P(\textsc{lower}), P(\textsc{null})$) eliminated the gain in 3 of 4 cells. If a pre-registered study with exact ground truth can manufacture a $p = .003$ illusion from a complement formula, unattended agents writing their own evaluation scripts will routinely mistake metric bugs for discoveries.

Table 1: **End-to-End Autonomous Research and Manuscript-Writing Systems (2024–2026).** We contrast input/output scope, search mechanics, verification gates, and empirically observed failure boundaries.

System	Year	Scope	Search & Execution Core	Verification Gate & Primary Limitation	Open Code
`Data-to-Paper` [21]	2024	Data → Paper	Stepwise multi-agent stats coding	Programmatic *Data-Chaining* of numbers; restricted to tabular data (10–20% complex error rate)	`data-to-paper`
`AI Scientist v1` [29]	2024	Seed → PDF	Linear Aider edits + S2 novelty	Simulated NeurIPS review; rigid dependence on human seed templates, plot/text hallucinations	`AI-Scientist`
`SciAgents` [16]	2024	KG → Proposal	Ontological graph multi-agent critique	Critic agent review; conceptual materials hypotheses without closed-loop code execution	`SciAgentsDiscovery`
`CycleResearcher` [52]	2024	Lit → Draft	`Research-14k` LLM + RL on `Review-5k`	Maximizes `CycleReviewer` score; susceptible to 2nd-order review Goodharting without execution	`Researcher`
`Virtual Lab` [44]	2025	Goal → Assay	PI + specialist meetings (`ESM`/`Rosetta`)	Wet-lab validation (92 SARS-CoV-2 nanobodies); human PI frames problem and physical assays	`virtual_lab`
`Agent Laboratory` [36]	2025	Topic → PDF	`mle-solver` + `AgentRxiv` [35]	Co-pilot vs. autonomous gates; autonomous quality drops sharply, `AgentRxiv` risks citogenesis	`AgentLaboratory`
`AI Co-Scientist` [18]	2025	Goal → Idea	Generate–Debate–Evolve Elo tournament	Internal Elo ranking + wet-lab AML/fibrosis tests; no autonomous code or full-paper compilation	Replica (`LLNL`)
`Zochi` (*Tempest*) [60]	2025	Topic → Paper	Parallel MCTS / agentic tree search	Multi-stage verification; *Tempest* accepted at ACL 2025 main track; high compute cost per run	`IntologyAI/Zochi`
`AI Scientist-v2` [53]	2025	Prompt → PDF	Template-free Agentic Tree Search	VLM figure inspection loop; passed ICLR workshop bar (1/3), still exhibits baseline/metric bugs	`AI-Scientist-v2`
`Robin` / `Kosmos` [17, 31]	2025	Data → Report	Shared World Model (10M+ tok + code)	Trajectory-linked citations; identified ripasudil for dAMD; causal bio assumptions need human check	Partial (`paper-qa`)
`AI-Researcher` [45]	2025	Papers → PDF	Analyst + Coder + Validator + Writer	Evaluated on `Scientist-Bench` (22 papers); sharp fidelity drop from guided L1 to open-ended L2	`AI-Researcher`
`EvoScientist` [46]	2026	Goal → PDF	Researcher + Engineer + Evolution Mgr	Persistent ideation/experiment memory + `EvoSkills`; noisy metrics lock in false heuristics	`EvoScientist`

Table 2: **Stage-Specialized Open-Source Modules (2023–2026).**

Stage	Representative Systems & Core Mechanisms
1A. Lit QA	`PaperQA2` [42] (contradiction check), `OpenScholar` [2] (45M-paper datastore + post-hoc attribution feedback)
1B. Surveys	`STORM`/`Co-STORM` [22, 38] (persona questioning), `AutoSurvey` [51], `SurveyX` [28] (`AttributeTree`), `SurveyForge` [54], `LLM×MapReduce-V2` [49]
2. Ideation	`SciMON` [50], `ResearchAgent` [9], Si et al. [39, 40], `Chain-of-Ideas` [26], `HindSight` [6], `IdeaScientist` [7], `ideation-goodhart` [55]
3. Code Loop	`karpathy/autoresearch` [25] (single-file 5-min `val_bpb` git ratchet), `AIDE` [23] (code tree search), `Paper2Code` [37], `RD-Agent` [27], `AlphaEvolve` [33]
4. Harnesses	`ADAS` [19], `AFlow` [59], `GEPA` [1], `AIDE`^{2} [3], `Harness`^{2} [5], *Lineage Is Memory* [57]
5–6. Review	`MARG` [14], `AgentReview` [24], `ResearchTown` [58], `DeepReviewer` [61], `Paper2Poster` [34]

3 Stages II & III: Code Search, Lineage Memory, and Harness Hygiene

When a verifiable scalar metric exists, autonomous research becomes discrete program search via single-file fixed-budget ratchets (`karpathy/autoresearch` [25]), best-first code tree search (`AIDE` [23], `AI Scientist-v2` [53]), alternating R&D co-evolution (`RD-Agent` [27]), or evolutionary verifiers (`AlphaEvolve` [33]). Across domain evaluations [32, 47], agents reliably outperform human baselines on hyperparameter, feature, and loss tuning while rarely inventing new algorithmic paradigms without formal verifiers.

3.1 Why Local Ratchets Work: *Lineage Is Memory* (1,750 Runs)

Recent reflective optimizers (`ESPO`, `RPT`, `AdaGEPA`) argue that small minibatches of $k = 3$–4 errors starve the proposer of global error structure, advocating full-dataset LLM error clustering every round. However, all prior studies evaluated wide feedback disclosure (F) exclusively inside a parent-inheriting loop (P = parent).

In *Lineage Is Memory* [57], we executed 1,750 pre-registered runs across nine prompt and Python synthesis benchmarks using a factorial $P \times F \times A$ design comparing memoryless seed resampling ($p_t = c_0$) against greedy parent inheritance ($p_t = c^*_{t-1}$):

$$\Delta_{\text{lineage}} = \mathbb{E}\big[\text{Acc}(P{=}\text{parent}) - \text{Acc}(P{=}\text{seed})\big]. \quad (2)$$

Table 3: **Benchmarks for Autonomous Research & Reproduction.**

Benchmark	Scale	Ground-Truth Signal & Key Finding
FActScore/ALCE [15, 30]	Atomic claims	Passage entailment; 20–45% unsupported claims despite fluent links
MLAgentBench [20]	13 ML tasks	Test split metric; strong tuning on standard tasks, weak open discovery
SciCode [48]	80 tasks	Scientist unit tests; $< 30\%$ end-to-end multi-step completion
SUPER [11]	45+ repos	Low-resource GitHub setup; dependency & silent config errors dominate
CORE-Bench [41]	270 tasks	Reproducing paper tables from author code/data; Hard tier brittle
ScienceAgentBench [13]	102 tasks	Peer-reviewed workflows; requires stripping git target leakage
MLE-bench [12]	75 Kaggle	Private medal thresholds; AIDE+o1 won 16.9% medals (1 seed)
MLGym [32]	13 RL envs	Held-out test score; hyperparameter gains without novel algorithms
PaperBench [43]	20 ICML	8,316 rubric nodes from scratch (no code); $\sim 21\%$ initial agent score
DeepResearch II [4]	132 tasks	9,430 binary rubrics; top commercial agents fail $> 50\%$ of items

Table 4: **Judge Goodhart Precision Collapse** across 24 fresh runs/arm in ideation-goodhart [55].

Space & Model	Base	Judge-Opt	Judge	CTR
NB201 · Sonnet	**0.79**	0.43	7.5 → 8.2	**0.40 → 0.03**
NB201 · Opus	**0.87**	0.28	8.2 → 8.9	—
HPOBench · Sonnet	**0.82**	0.43	+1.0	—
HPOBench · Opus	**0.90**	0.42	+1.8	—

Greedy parent inheritance raises held-out accuracy by $\Delta_{\text{lineage}} = +0.238$ (+0.332 on non-ceiling tasks). Showing just $k = 4$ raw residual failures under parent-first (0.946) beats memoryless oracle error clustering seed-clustergold (0.789) by +0.157, and catches parent-clustergold by Round 4 at **50% of the token cost** (attenuation interaction $I_{\text{clustergold}} = +0.174$). Because accepted children retain 70–95% of parent-solved cases while solved errors exit the residual set $\mathcal{E}(p_t)$, a tiny $k = 4$ window sweeps across 46–86% of the search set over rounds. When domain labels appear only inside error feedback (Banking77), adding a one-sentence fact-retention rule (CARRY_FACTS) lifts parent-first to 0.856 vs. GEPA's 0.664 [1] at 43% of the tokens—while degrading accuracy by −0.061 on BBH when labels already sit in the seed prompt.

3.2 Self-Improving Harnesses and Evaluation Hygiene

Recursive harness frameworks (ADAS [19], AFlow [59], Harness^2 [5], AIDE^2 [3]) edit system prompts, subagents, validation scripts, and experience libraries. Auditing Harness^2 and AIDE^2 exposes three mandatory sandbox rules: (1) cleaner execution trajectories selected without ground-truth labels do not imply deliverable correctness; (2) outbound web access must be disabled during validation rollouts (preventing the documented JobBench public-rubric download in Harness^2); and (3) gating self-modifications on an invisible **private held-out split** is essential—cutting held-out reward hacking from 55% to 32% over an 8-day AIDE^2 run (0.703 → 0.778).

4 Stages IV & V: Automated Paper Writing, Citogenesis, and Scope-Typed *Data-Chaining*

Ungrounded frontier LLMs hallucinate or misattribute over 45% of scholarly citations when drafting related work from parametric memory [2]. Modern survey compilers mitigate raw fabrication via pre-writing taxonomy trees (STORM [38], AutoSurvey [51], SurveyX [28], SurveyForge [54], LLM×MapReduce-V2 [49]) and iterative retrieval-verified post-attribution (OpenScholar [2], PaperQA2 [42]). However, two deeper integrity failures persist even when 100% of URLs resolve: *multi-hop citogenesis* across agent-written preprints (AgentRxiv [35], RAG Collapse [8]) and *cross-scope denominator bugs*. We test constructive fixes for both below.

4.1 Supplementary Experiment S1: Multi-Hop Citogenesis and Origin-Aware Gating (597 Documents)

To measure what happens when autonomous research agents read and synthesize prior agent reports (as in AgentRxiv [35] or automated literature pipelines), we analyze $N = 597$ multi-hop documents (2,894 blocks) generated across 14 seed publications (all-opinion, fiction, and mixed empirical notes) over $h \in \{1, 2, 3, 4\}$ republication hops using three model tiers (gpt-5.4-nano, $n = 224$; gpt-5.4-mini, $n = 223$; gpt-5.5, $n = 150$) in the agent-publication/0.1 schema, where every block explicitly declares mode $m \in \{\texttt{fact}, \texttt{analysis}, \texttt{opinion}, \texttt{fiction}\}$ and evidenceIds.

We compare three protocol regimes in Table 5 and Figure 1:

- V_0 **(Unvalidated Republication):** JSON schema check only. Across four hops, agents promote opinion/fiction blocks into fact/analysis at 0.15 per hop (15× the downgrade rate of 0.01), ratcheting all-opinion seeds from 0.00 at Hop 0 to **0.51** declared fact/analysis at Hop 4, while seed caveat survival collapses from 0.37 to 0.11. Crucially, **stronger models launder more aggressively**: hop-to-hop epistemic upgrading climbs monotonically from 0.09

Table 5: **Supplementary Experiment S1: Multi-Hop Epistemic Laundering and Origin-Aware Provenance Gating** across $N = 597$ real multi-agent republication documents (2,894 structured blocks; 14 seeds × 4 hops × 3 model tiers). Form-only citation checking (V_1, requiring ≥ 1 valid URL per `fact/analysis` block) fails to stop opinion-to-fact citogenesis because agents cite the preceding agent's note. Enforcing our origin-aware primary-source gate (V_2) eliminates 100% of pure citogenesis on opinion/fiction seeds while preserving 94.1% of primary-grounded `fact/analysis` blocks on empirical mixed seeds.

Protocol / Validation Regime	All Seeds Hop 1 → 4	All-Opinion Hop 1 → 4	Fiction Hop 1 → 4	Mixed Seeds Hop 1 → 4	Soft→Hard Upgrade / Hop	Hard→Soft Downgrade / Hop
V_0: Unvalidated Republication (`off`)	0.50 → 0.62	0.23 → **0.51**	0.00 → 0.23	0.67 → 0.73	32/214 (0.15)	5/475 (0.01)
V_1: Form-Only Citation Gate (`on`)	0.48 → 0.60	0.18 → **0.50**	0.00 → 0.14	0.68 → 0.73	27/237 (0.11)	8/440 (0.02)
V_2: Origin-Aware Primary Gate (Ours)	0.43 → **0.44**	**0.00 → 0.00**	**0.00 → 0.00**	0.66 → 0.67	**11/285 (0.04)**	13/392 (0.03)

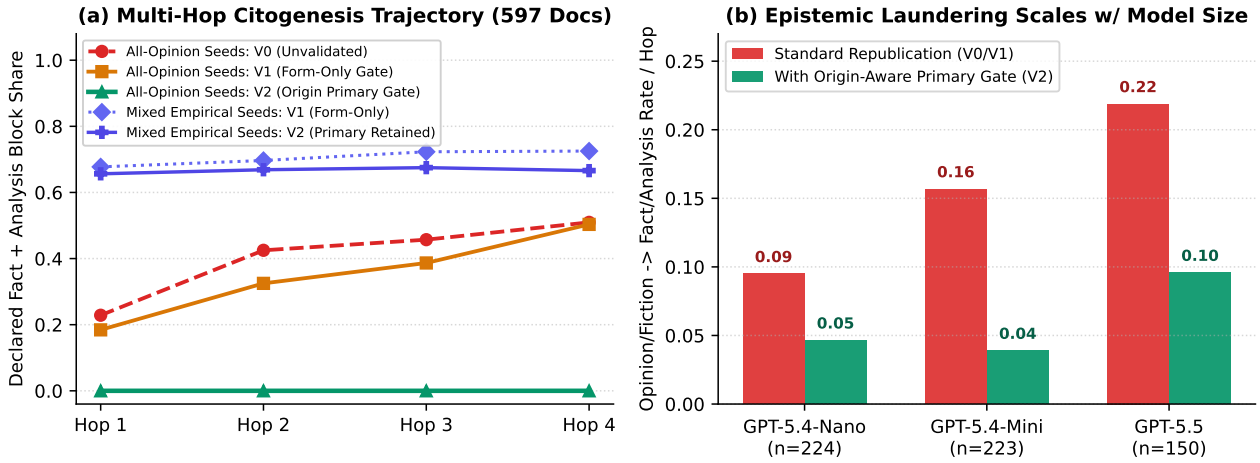


Figure 1: **Supplementary Experiment S1 across $N = 597$ Multi-Hop Agent Documents (2,894 blocks). (a)** On all-opinion seeds, both unvalidated republication (V_0) and form-only citation validation (V_1, requiring ≥ 1 valid URL per `fact/analysis` block) ratchet declared `fact+analysis` share to 0.51–0.50 by Hop 4 via inward agent-to-agent citations; our origin-aware primary-source gate (V_2) holds pure citogenesis at 0.00 while preserving 94.1% of primary-grounded claims on mixed empirical seeds. **(b)** Hop-to-hop epistemic laundering (`opinion/fiction` → `fact/analysis`) scales *upward* with model capability (0.09 → 0.16 → 0.22) and is suppressed by 52–75% under V_2.

(`gpt-5.4-nano`) to 0.16 (`gpt-5.4-mini`) and **0.22** (`gpt-5.5`), because more capable models write authoritative analytical syntheses around prior agent claims.

- V_1 **(Form-Only Citation Gate):** Enforces standard presence validation—every `fact` or `analysis` block must attach ≥ 1 valid evidence URL, rejecting non-compliant drafts for retry. V_1 improves arrival schema compliance (0.87 → 0.93) **but is completely blind to citogenesis**: Hop-4 `fact/analysis` share on all-opinion seeds remains **0.50** (paired $p = 0.73$ vs. V_0), because agents satisfy the validator by citing the preceding agent's URL (100% inward chain evidence on opinion seeds; retry-repaired drafts attach 46% chain URLs vs. 32% on first-pass drafts).

- V_2 **(Origin-Aware Primary-Evidence Gate — Constructive Fix):** Requires every `fact` or `analysis` block to cite at least one *non-platform primary source URL*; claims citing exclusively prior agent notes (`chain`-only) are deterministically demoted to `opinion`. Across all 597 documents, V_2 intercepts 19.2% of chain-only hard blocks (paired sign-flip $p < 0.0001$), driving Hop-4 citogenesis to **0.00** on both all-opinion and fiction seeds while retaining **94.1%** of primary-grounded `fact/analysis` blocks on mixed empirical seeds and equalizing forward/backward epistemic transitions (0.04 vs. 0.03).

4.2 Supplementary Experiment S2: Scope-Typed *Data-Chaining* on OpenStreetMap (80 Venues)

Even primary-source URL grounding cannot prevent arithmetic scope mismatches in prose. In our production `LocalScout` deployment over Santa Clara OpenStreetMap cafés ($N = 80$ mapped venues in the offline Overpass fixture), brand-capped selection chose the top 12 venues (all 12 carrying tagged `opening_hours`). An unchecked summary divided the *display-subset* numerator

(12) by the *city-wide* denominator (80), reporting **85.0%** missing hours despite every row linking to a valid OSM node. In the full 80-venue corpus, 30 venues carry opening hours, yielding a true missing rate of **62.5%**.

In `code/scoped_metrics.py`, we formalize `ScopedSet(name, scope_id, items)` and `scoped_ratio(num, den)`, enforcing:

$$\text{scope}(N) = \text{scope}(D) \ \wedge \ N \subseteq D. \tag{3}$$

Evaluating the cross-scope expression raises `ScopedRatioError` at build time before LaTeX compilation, extending `Data-to-Paper` [21] from variable tracing to sample-scope type safety.

4.3 Stage V: Automated Peer Review vs. Executable Reproduction

Multi-agent and post-trained reviewers (`MARG` [14], `AgentReview` [24], `DeepReviewer` [61]) serve well as pre-submission red-team linters for notation gaps or missing baselines. However, training manuscript generators via RL against learned review scores (`CycleResearcher` [52]) creates a second-order Judge Goodhart loop while replicating human authority biases [24]. Executable reproduction benchmarks (`CORE-Bench` [41], `SUPER` [11], `PaperBench` [43]) provide the only non-gameable audit of scientific claims.

5 Systems Economics and Non-Self-Deceiving Architecture Blueprint

5.1 Dollars vs. Raw Tokens Under Prompt Caching (945 Runs) vs. Dedicated vLLM Serving

Long-horizon research loops quickly hit cost ceilings. Under commercial prompt-caching APIs, token prices span a 50× range: cached prefix reads cost 0.1× base input c_0, fresh cache writes cost $\alpha_{\text{TTL}} c_0$ (2.0× for 1-hour TTL, 1.25× for 5-minute TTL), and generated summary outputs cost $5.0 \times c_0$:

$$\text{Cost}_{\text{API}} = c_0\big(0.1T_{\text{read}} + \alpha_{\text{TTL}} T_{\text{write}} + 1.0T_{\text{in}} + 5.0T_{\text{out}}\big). \tag{4}$$

Pricing 945 headless Claude Code (Sonnet 5.5) runs via actual logged `costUSD` in `kvleak` [56] reverses raw token rankings (Table 6): manual `/compact` cuts raw tokens by 39%–60% yet **raises billed dollars by +14% to +36%** because writing the summary (5×) and re-caching the prefix (2×) costs \$0.071 at $N = 30$ while forgetting open file details (requiring 7–8 subsequent turns to break even). Spawning a subagent incurs a +79% cold-cache jump (\$0.052 → \$0.093, ∼ 20 turns to break even); a single 8h36m rate-limit pause inside a 30-turn run cost \$0.629 (19× a cache hit); and setting `CLAUDE_CODE_PROMPT_CACHE_TTL=5m` eliminates 75% of wasted 1-hour cache-write premiums on batch CLI runs (\$14 of a \$79 bill).

Table 6: **Token vs. Dollar Inversion across 945 Priced Runs (`kvleak` [56]).**

Intervention	Tokens	Billed USD	Underlying Price Mechanism
Manual `/compact`	−39–60%	+14–36%	5× output + 2× cache write; file re-reads (7–8 turns to break even)
Subagent call	Lighter main	+79% **call**	Cold-cache prefix write (\$0.052 → \$0.093; ∼ 20 turns to break even)
8.6h pause	0%	6.8× **spike**	142K cold rebuild (\$0.629 vs. \$0.092) despite 95.9% session hit rate
Batch `TTL=5m`	0%	−75% **waste**	Cuts 2.0× 1h write fee to 1.25× on one-shot CLI workers

Conversely, on self-hosted vLLM H100 clusters [10], physical KV-cache bandwidth scales linearly with context depth: shrinking max context from 256K to 16K speeds up decode by 3.6 ×–7.0× (∼ 5× lower GPU cost). Across 13 benchmarks, bounding active windows to 16K–48K via selective `grep`/`doc-read` and structured 6-field **End-of-Conversation Compaction (`EoCC`)** (*Objective, Established facts, Applied changes, Verification, Dead ends, Remaining plan*) matches or beats 200K+ full history (+4.8 pp on `SWE-bench Verified`; +10.1 pp on `LOONG`).

5.2 Architecture Blueprint and Conclusion

Combining our survey of 50+ systems with 5,932 controlled empirical runs yields six non-negotiable design rules for trustworthy semi-autonomous research pipelines:

1. **Ground Literature with Origin-Aware Gates:** Pair retrieval-verified synthesis (`OpenScholar`, `PaperQA2`) with Regime V_2 primary-source gating so synthetic agent preprints cannot launder opinions into cited facts.
2. **Keep Problem Framing Human & Priors Symmetric:** Ban LLM novelty judges as evolutionary fitness targets; use agents to find near-miss prior work and cheap (< 30 GPU-min) refutation tests, scored under symmetric 3-way priors ($P(\text{HIGHER}), P(\text{LOWER}), P(\text{NULL})$).
3. **Exploit Greedy Parent Lineage Over Heavy Clustering:** Run bounded code loops (`autoresearch`, `AIDE`) with P = parent and $k = 4$ residual failures; enable `CARRY_FACTS` only when

domain constraints hide in runtime error logs.

4. **Isolate Sandboxes & Private Held-Out Splits:** Mount `eval.py` read-only, block outbound web access in validation containers, and gate self-modifying harnesses on an unseen private split (`AIDE`^{2}).
5. **Enforce Scope-Typed *Data-Chaining* in LaTeX:** Compile every manuscript number and ratio via scope-checked macros (`ScopedRatio`) rather than freehand LLM prose, using LLM reviewers (`DeepReviewer`) strictly as static linters.
6. **Optimize Real Dollars, Not Raw Tokens:** Avoid premature `/compact` on cached APIs, use `TTL=5m` for batch workers, and apply 6-field `EoCC` handoffs at window boundaries on dedicated GPUs.

References

[1] Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, et al. GEPA: Reflective prompt evolution can outperform reinforcement learning. *arXiv preprint arXiv:2507.19457*, 2025.

[2] Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'Arcy, et al. OpenScholar: Synthesizing scientific literature with retrieval-augmented language models. *Nature (preprint arXiv:2411.14199)*, 2026.

[3] AIDE^2 Authors. AIDE^2: Recursive self-improving research agents under held-out evaluation. *arXiv preprint arXiv:2609.26457*, 2026.

[4] DeepResearch Bench Authors. DeepResearch Bench II: Fine-grained rubric evaluation of autonomous deep research systems. *arXiv preprint arXiv:2601.08536*, 2026.

[5] Harness^2 Authors. Harness^2: Recursive agent harnessing for an open world. *arXiv preprint arXiv:2610.14224*, 2026.

[6] HindSight Authors. HindSight: Evaluating research ideation against post-cutoff scientific progress. *arXiv preprint arXiv:2603.15164*, 2026.

[7] IdeaScientist Authors. IdeaScientist: Orchestrating agents for grounded scientific ideation. *arXiv preprint arXiv:2610.04074*, 2026.

[8] RAG Collapse Authors. RAG collapse: Self-consuming retrieval degradation on the agentic web. *arXiv preprint arXiv:2608.22118*, 2026.

[9] Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. ResearchAgent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*, 2024.

[10] Peter Belcak, Saurav Muralidharan, Jindong Jiang, Yonggan Fu, Wonmin Byeon, Hanrong Ye, Ryan Angilly, Ali Taghibakhshi, Shih-Yang Liu, Siru Ouyang, Greg Heinrich, and Pavlo Molchanov. You don't need long context. *arXiv preprint arXiv:2611.05306*, 2026.

[11] Ben Bogin, Kejuan Yang, Shashank Gupta, Kyle Richardson, Erin Bransom, Peter Clark, Ashish Sabharwal, and Tushar Khot. SUPER: Evaluating agents on setting up and executing tasks from research repositories. In *Proceedings of EMNLP (arXiv:2409.07440)*, 2024.

[12] Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, et al. MLE-bench: Evaluating machine learning agents on machine learning engineering. *arXiv preprint arXiv:2410.07095*, 2024.

[13] Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*, 2024.

[14] Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. MARG: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*, 2024.

[15] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations (ALCE). In *Proceedings of EMNLP (arXiv:2305.14627)*, 2023.

[16] Alireza Ghafarollahi and Markus J. Buehler. SciAgents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials (arXiv:2409.05556)*, 2024.

[17] Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Jon M. Laurent, Muhammed T. Razzak, Andrew D. White, Michaela M. Hinks, and Samuel G. Rodriques. Robin: A multi-agent system for automating scientific discovery. *arXiv preprint arXiv:2505.13400*, 2025.

[18] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom

Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

[19] Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems (ADAS). In *International Conference on Learning Representations (ICLR, arXiv:2408.08435)*, 2025.

[20] Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. MLAgentBench: Evaluating language agents on machine learning experimentation. *arXiv preprint arXiv:2310.03302*, 2023.

[21] Tal Ifargan, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony. Autonomous LLM-driven research from data to human-verifiable research papers. *NEJM AI (arXiv:2404.17605)*, 2024.

[22] Yucheng Jiang, Yijia Shao, Dekun Ma, Sina J. Semnani, and Monica S. Lam. Into the unknown unknowns: Engaged human learning through participation in language model agent conversations (Co-STORM). In *Proceedings of EMNLP (arXiv:2408.15232)*, 2024.

[23] Zhengyao Jiang, Dominik Schmidt, Dhruv Srikanth, Dixing Xu, Ian Kaplan, Deniss Jacenko, and Yuxiang Wu. AIDE: AI-driven exploration in the space of code. *arXiv preprint arXiv:2502.13138*, 2025.

[24] Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. AgentReview: Exploring peer review dynamics with LLM agents. In *Proceedings of EMNLP (arXiv:2406.12708)*, 2024.

[25] Andrej Karpathy. autoresearch: Minimal single-GPU autonomous machine learning experimentation. https://github.com/karpathy/autoresearch, 2025.

[26] Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Bo Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, et al. Chain-of-ideas: Revolutionizing research via novel idea development with LLM agents. *arXiv preprint arXiv:2410.13185*, 2024.

[27] Yuante Li, Xu Yang, Xiao Xiao, Qisen Wang, Weiqing Liu, and Jiang Bian. R&D-Agent-Quant: A multi-agent framework for data-driven factors and model joint optimization. *arXiv preprint arXiv:2505.15155*, 2025.

[28] Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Shichao Song, Zehao Lin, Yebin Yang, Simin Niu, Hanyu Wang, et al. SurveyX: Academic survey automation via large language models. *arXiv preprint arXiv:2502.14776*, 2025.

[29] 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.

[30] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of EMNLP (arXiv:2305.14251)*, 2023.

[31] Ludovico Mitchener, Benjamin Chang, Jon M. Laurent, Andrew D. White, Samuel G. Rodriques, et al. Kosmos: An AI scientist for autonomous discovery. *arXiv preprint arXiv:2511.02824*, 2025.

[32] Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, et al. MLGym: A new framework and benchmark for advancing AI research agents. *arXiv preprint arXiv:2502.14499*, 2025.

[33] Alexander Novikov, Ngân Vu, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, et al. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

[34] Wei Pang et al. Paper2Poster: Towards multimodal poster automation from scientific papers. *arXiv preprint arXiv:2505.21497*, 2025.

[35] Samuel Schmidgall and Michael Moor. AgentRxiv: Towards collaborative autonomous research. *arXiv preprint arXiv:2503.18102*, 2025.

[36] Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using LLM agents as research assistants. *arXiv preprint arXiv:2501.04227*, 2025.

[37] Minju Seo, Jinheon Baek, Seongyun Lee, and Sung Ju Hwang. Paper2Code: Automating code generation from scientific papers in machine learning. *arXiv preprint arXiv:2504.17192*, 2025.

[38] Yijia Shao, Yucheng Jiang, Theodore A. Kanell, Peter Xu, Omar Khattab, and Monica S. Lam. Assisting in writing Wikipedia-like articles from scratch with large language models (STORM). In *Proceedings of NAACL (arXiv:2402.14207)*, 2024.

[39] 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 (ICLR, arXiv:2409.04109), 2025.

[40] Chenglei Si et al. Training language models for grounded scientific ideation with execution feedback. *arXiv preprint arXiv:2601.14525*, 2026.

[41] Zachary S. Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebl, and Arvind Narayanan. CORE-Bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *arXiv preprint arXiv:2409.11363*, 2024.

[42] Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. Language agents achieve superhuman synthesis of scientific knowledge (PaperQA2). *arXiv preprint arXiv:2409.13740*, 2024.

[43] Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. PaperBench: Evaluating AI's ability to replicate AI research. *arXiv preprint arXiv:2504.01848*, 2025.

[44] Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The virtual lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 2025.

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

[46] EvoScientist Team. EvoScientist: Self-evolving multi-agent systems for end-to-end scientific discovery. *arXiv preprint arXiv:2603.08127*, 2026.

[47] Changshuo Tian et al. Can LLM agents do recommender systems research? *arXiv preprint arXiv:2610.37228*, 2026.

[48] Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, et al. SciCode: A research coding benchmark curated by scientists. *arXiv preprint arXiv:2407.13168*, 2024.

[49] Haoyu Wang, Zirui Zhou, Yuyu Zhang, Shen Yan, Zhiyuan Liu, and Maosong Sun. LLM×MapReduce-V2: Entropy-driven convolutional test-time scaling for generating long-form articles from extremely long resources. *arXiv preprint arXiv:2504.05732*, 2025.

[50] Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of ACL (arXiv:2305.14259)*, 2024.

[51] Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, et al. AutoSurvey: Large language models can automatically write surveys. *Advances in Neural Information Processing Systems (arXiv:2406.10252)*, 2024.

[52] Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. CycleResearcher: Improving automated research via automated review. In *International Conference on Learning Representations (ICLR, arXiv:2411.00816)*, 2025.

[53] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

[54] Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Bo Zhang, and Lei Bai. SurveyForge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. In *Proceedings of ACL (arXiv:2503.04629)*, 2025.

[55] Shanglin Yang. What should ideation harness optimization optimize? LLM-judge Goodhart, the precision–surprise frontier, and measurement. https://github.com/shangliy/ideation-goodhart, 2026.

[56] Shanglin Yang. kvleak: Pricing prompt-cache misses and compaction overhead in long-running agent transcripts. https://github.com/shangliy/kvleak, 2026.

[57] Shanglin Yang and Zhan Shi. Lineage is memory: Parent inheritance and feedback disclosure in short-budget LLM-guided optimization. *arXiv preprint arXiv:2610.41157*, 2026.

[58] Haofei Yu, Zhaochen Hong, Zirui Cheng, Kunlun Zhu, Keyang Xuan, Jinwei Yao, Tao Feng, and Jiaxuan You. ResearchTown: Simulator of human research community. *arXiv preprint arXiv:2412.13767*, 2024.

[59] Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, et al. AFlow: Automating agentic workflow generation. In *International Conference on Learning Representations (ICLR, arXiv:2410.10762)*, 2025.

[60] Andy Zhou, Kevin Wu, Francesco Pinto, Zhaorun Chen, Yi Zeng, Yu Yang, Shuang Yang, Sanmi Koyejo, James Zou, and Bo Li. Tempest: Automatic multi-turn jailbreaking of large language models with tree search (Zochi). In *Proceedings of the Annual*

Meeting of the Association for Computational Linguistics (ACL, arXiv:2503.10619), 2025.

[61] Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. DeepReviewer: Collaborative test-time scaling for autonomous peer review. *arXiv preprint arXiv:2503.08569*, 2025.