Bounded Returns to Orchestration:
A Synthesis Ceiling Beneath Eleven Controlled Deep-Research Architectures

Peter McCann Strain
peter.mccann.strain@gmail.com
petermccannstrain.com

July 29, 2026

Abstract

Deep-research agents pairing LLMs with web search are usually compared while orchestration logic, base model, and retrieval stack all change at once, so no difference is cleanly attributable to architecture. We hold the tool layer and base model fixed and run eleven architectures (a single-pass baseline, eight GPT-4o pipelines, two 7B agents) across 90 queries, scored by a cross-family three-judge panel on a nine-dimension rubric. We reliability-audit the panel *before* use. Four findings follow, their nulls load-bearing. (i) *Orchestration has bounded, task-conditional returns*: the gain is largest where the single-pass baseline is weakest and vanishes where the baseline already suffices. No architecture dominates on the judge-robust criterion, and the five strongest pipelines are judge-robustly inseparable. The largest isolable component moves scores by ≈ 0.06 raw (0.04 length-adjusted, still significant). Clamping the highest-mean pipeline's token budget towards the baseline's leaves its advantage undetectable, a power-limited null. A length-adjustment control separately bounds the baseline-to-cluster premium to roughly 0–0.05 across three functional-form specifications. We therefore attribute the gain to architecture bundled with the compute it spends. One lever outside this taxonomy, inference-time verification, does move scores significantly (Section 10.5). (ii) A 7B agent on the same scaffold scores 0.23 below its GPT-4o twin (0.14 on decontaminated queries), and 0.38 (0.34) below the top-cluster pipeline P4. That comparison uses P4. The highest-mean pipeline of (i) is P1, and substituting it would widen the gap. Orchestrating the 7B backbone adds only 0.01–0.07 (n.s.). We do not attribute the gap to scale alone. (iii) The one varying signal, citation quality, is partly judge-made: citation *density* earns a provenance-independent bonus from the two Claude judges but a precise null from GPT-5.2. The panel carries only 1.65 effective votes. (iv) Serving every architecture identical pooled oracle sources raises the six-pattern top cluster's citation quality ($+0.162$, 95% CI $[0.06, 0.26]$), yet leaves its factual accuracy statistically equivalent to zero within ± 0.05. A claim-level entailment measure attributes the shortfall the same way: pooled near-ideal retrieval closes only ≈ 0.09 of it. The binding limit is *synthesis*, beneath the retrieval ceiling. Findings (i) and (iv) reproduce on a second, same-lineage frontier backbone, a reduced contrast with n=32 queries for (i) and n=17 for (iv); finding (i) additionally survives a search-success control. We release the full analysis apparatus now, with the underlying 248,536 criterion-level verdicts staged for a forthcoming archive version.

1 Introduction

Ask two deep-research systems the same question, and you receive two multi-page reports. Ask which *architecture* produced the better one, and the published literature cannot tell you. The obstacle is structural. Systems are released as whole stacks: a new agent ships a new orchestration strategy, but also a different base model and a different search-and-extraction toolchain. Its reported gains are measured against whatever the previous stack happened to be. An observed quality difference can therefore be charged to any of the three, or to their interaction. The field has produced dozens of deep-research systems [14, 55, 84, 125] but very few

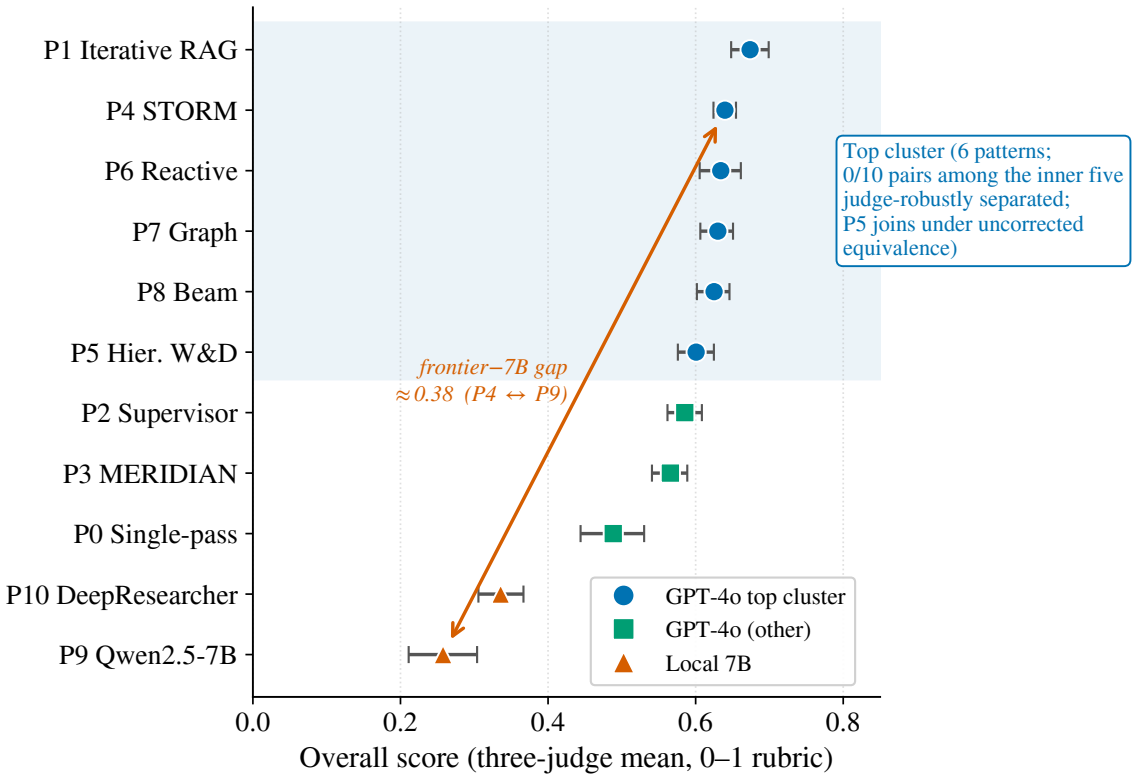

Figure 1: No architecture dominates; a sub-frontier 7B sits far below GPT-4o. Per-pattern overall score (three-judge mean on a 0–1 rubric, N=90 queries, 87–89 for three patterns; error bars are query-clustered bootstrap 95% confidence intervals of the mean, 5,000 resamples). Six GPT-4o pipelines (dark, shaded band) form the top band (*the cluster*, defined in Section 5.3); no pair among the inner five (the cluster minus P5, also defined in Section 5.3) is separated by the per-judge robust-separation test in any of the three judges, and a sixth (P5) joins under uncorrected equivalence testing at the ±0.05 margin. The single-pass baseline (P0) and the two local-7B agents (P9/P10, orange) sit well below. The orange arrow marks the frontier–7B gap (P4↔P9, ≈ 0.38), which the contrast cannot isolate to model scale (it also bundles pre-training data, alignment, and 4-bit quantisation). Grand means average over per-source leaders that rotate (Table 8).

controlled comparisons. The controlled studies that do exist [97, 103] consistently report smaller architectural effects than the single-system papers imply. The qualitative MAST failure taxonomy [8] corroborates this pattern from case analysis. When the confound is removed and architecture is measured cleanly, the striking result is what does *not* move. Most of the quality signal the field attributes to orchestration flattens out. A null result is only informative if the measuring instrument can be trusted. Establishing that these nulls are *real*, and not the instrument simply failing to resolve a difference, is as much of the contribution as the nulls themselves.

The closest prior work exposes the confound from two directions. DeepScholar-Bench [74] restricts web access to a single arXiv retrieval API across every system it evaluates. It also runs a gold-reference oracle *ablation*, an experiment that removes or replaces one component of a system and re-measures the result to see how much that piece was contributing, that nearly saturates retrieval-quality and verifiability scores while nugget coverage stays low. That is the clearest existing evidence that retrieval, not orchestration, is the binding constraint on academic synthesis. But it is limited to academic synthesis: it uses no cross-family judge panel and reports no inter-rater-reliability or equivalence statistics. A study of five agent architectures instantiated across three model families and six long-horizon, environment-interactive benchmarks (260 configurations) standardises tools, prompts, and compute [40]. It finds that coordination delivers diminishing returns past a

capability threshold, convergent with what we report. Its task family differs from ours, though: agentic coding and tool-use tasks scored by execution-based checks and LLM judges, versus our long-form judged research-report synthesis. Removing the confound among orchestration, model, and retrieval is a *precondition*: only once architecture is isolated can the flat signal be read as a fact about orchestration rather than an artefact of the comparison. What remains missing is a comparison that fixes a single tool layer, spans general-purpose deep-research queries, places frontier pipelines and local 7B agents on one axis, and reports the reliability and equivalence statistics a null-heavy result demands.

This paper supplies it. We fix one Bing / Semantic-Scholar / arXiv / trafilatura tool layer and one retry policy. We run eleven architectures at full budget on a shared 90-query manifest, scored by a cross-family three-judge panel (GPT-5.2, Claude Opus 4.1, Claude Sonnet 4.5) on a nine-dimension rubric. Holding everything but one factor constant yields three clean contrasts. The first is orchestration (P0 vs. P1–P8; model and tools fixed). The second is the model bundle (P9 vs. P0; architecture and tools fixed, while scale, data, alignment, and quantisation move together). The third is RL training (P9 vs. P10; same backbone and tools), though the RL agent also adds a learned multi-search loop, so this contrast bundles training with scaffold. We additionally reliability-audit the judge panel itself before comparing any patterns. The audit is a finding in its own right, because on a null-heavy result the trustworthiness of the instrument is what licenses the nulls. The pre-use audit establishes that the panel resolves relative orderings, even though the three-judge panel collapses to only 1.65 effective votes (its same-lab pair to 1.32). No prior controlled study carries all of this at once: one frozen tool layer underneath an orchestration, a model-capability, and an RL contrast; a reliability-audited cross-family panel; an architecture-controlled oracle-retrieval intervention that the panel re-scores; and a compute-matched best-of-N control. Three contemporaneous 2025–2026 studies reach adjacent conclusions from single angles: within-system removal attribution [61], a capability-saturation law [40], and panel redundancy [41]. What is new here is their conjunction on one fixed model and tool layer, plus two interventions none of them runs. The first is a direct oracle test that separates the retrieval ceiling from the synthesis ceiling by feeding every architecture identical pooled sources. The second is a judge-localised audit showing that the one varying signal, citation quality, is partly a citation-density artefact of two same-lab judges.

Our contribution is an instrument trustworthy enough to make "no effect" a defensible measurement, together with what that instrument then locates: the effect lives in capability and synthesis. We report four findings below. The fifth item, the released and reliability-audited apparatus, we claim as foundation: the other four findings rest on it. An architecture-controlled oracle-retrieval intervention ties findings 1 and 2 together, demonstrating directly the retrieval-and-synthesis ceiling that sits beneath the orchestration layer (Section 7, Section 10).

Contributions.

1. **Orchestration has bounded, task-conditional returns.** No architecture dominates across the query distribution: the per-source leader rotates among P1, P6, and P7 (Table 8). That rotation is descriptive; the certified fact behind it is the pattern×source interaction. The five strongest GPT-4o pipelines are judge-robustly inseparable on the grand mean (a sixth, P5, joins only under uncorrected equivalence testing at the ±0.05 margin). A seven-component ablation registry puts the largest isolable component lever at ≈ 0.06 raw (0.04 length-adjusted), a small fraction of the capability gap reported below (Section 5.3, Section 8). A matched-budget probe bounds the premium further: clamping the highest-mean pipeline's (P1's) token budget down towards the single-pass baseline's own budget leaves P1's advantage no longer distinguishable from zero, a power-limited null (Section 5.8). We therefore attribute the gain to architecture bundled with the compute it spends: the tool interface is held fixed, but token budget is not equalised. A length-adjustment control likewise bounds the length-robust share of the baseline-to-cluster step at roughly 0–0.05 (Section 5.9). A frozen-evidence defence isolates the mechanism further: holding

one identical evidence set fixed and varying only the synthesis scaffold, one of four alternative scaffolds, draft-then-revise, does separate from a single pass on factual accuracy (+0.079, replicated by an independent judge family at +0.125); a matched-compute check rules out call count as the driver. The other three scaffolds, and an injected-noise dose ladder, find no significant evidence that orchestration's edge over the baseline otherwise grows as evidence quality degrades (Section 7.1). A second positive lever survives outside this design too: rubric-guided verification beats a budget-matched unguided-revision control on both the weakest and the strongest pipeline (Section 10.5). The paper's bounded-returns reading is therefore compatible with two specific, replicated mechanisms doing real work, draft-then-revise (Section 7.1) and rubric-guided verification, against a broader field of components and architectures that do not. Bounded returns also reproduces externally. Across six independent, publicly released deep-research codebases (plus one run under a second search backend), the top four of the seven arms land within 0.018 of one another on the primary judge. The leaderboard order itself is judge-family-dependent, though, and none of the pairwise rank agreements reaches significance at this sample size (Section 5.11).

2. **The frontier–7B capability gap dwarfs every architectural effect.** A sub-frontier 7B agent on the same scaffold scores 0.23 below its GPT-4o twin and 0.38 below the top-cluster pipeline P4 on the three-judge mean, larger than the GPT-4o patterns' entire span. The gap is present and same-signed on every individual judge (per-judge 0.29–0.46 against P4), and it survives decontamination (0.14 and 0.34 on the contamination-clean residual, both significant) (Section 5.4, Figure 1). Orchestration shows no sign of closing it from either side: running the P1 and P4 architectures on the 7B backbone yields premiums of +0.01 and +0.07, neither significant on the 12-query probe. Their confidence intervals exclude closure beyond +0.14, a fraction of the 0.38 gap (Section 5.4). We do *not* attribute the gap to parameter count alone: the contrast also varies pre-training data, alignment, and quantisation. A contemporaneous task-matched 8B agent reaches near-frontier quality elsewhere [83]. The binding axis is model capability and its match to the task, which orchestration does not supply. At the frontier, capability also looks like the *cheaper* axis. On an uncorrected, coverage-limited comparison, one GPT-4o-to-GPT-4.1 generation upgrade buys at least as much quality as the entire architecture stack, from the single pass up to the top pipeline, at roughly $1/17.5$ the token spend (Section 5.4).

3. **Citation scoring adds a density bonus on top of grounding.** A report-level regression of the `factual_accuracy` score shows grounded citations raise it ($\beta_{\text{provenance}}$=+0.108), while citation *density* adds an independent, provenance-independent bonus ($\beta_{\log \text{count}}$=+0.040). The effect is robust under the query-clustered test, but it is not certifiable at the coarser eleven-pattern grouping, where too few clusters remain (Section 6). The bonus is carried by the two Claude judges, and it is a precise null under GPT-5.2 (β=−0.001). The bonus therefore belongs to a citation-*presence* rubric read by length-sensitive judges. Two concurrent audits [71, 79] corroborate it on deployed systems.

4. **A pooled-source oracle intervention isolates a dual ceiling, confirmed across the full three-judge panel.** Serving every pattern an identical ideal source set raises `citation_quality` by +0.162 pooled across the six-pattern cluster (defined in Section 5.3; block-bootstrap 95% CI $[0.06, 0.26]$, resampled by pattern), while leaving `factual_accuracy` statistically equivalent to zero within ±0.05. On the GPT-5.2 single-judge 30-query arm, the same intervention shrinks the baseline-to-cluster gap from 0.044 to 0.010. It does not re-order the cluster, and the change's bootstrap CI spans zero (Section 7). The split reproduces under both Claude judges as well, so it does not turn on a single judge. An injected-gold dose ladder makes the mechanism legible: as the gold available to the pipelines rises, they cite and paraphrase it almost one-to-one yet convert none of it to verified accuracy (Section 7). Citation quality is retrieval-bound, and factual accuracy is synthesis-bound. Neither ceiling is the orchestration layer the field optimises.

5. **A released, reliability-audited apparatus.** We release the full analysis apparatus: every build and rec-

onciliation script, the canonical number store, the manifest and nine-dimension rubric, and the ablation registry specification. We also release a judge panel audited for inter-rater reliability and length/citation-density bias. The underlying 248,536 criterion-level verdicts and the Bing-vs-Tavily tool-layer intervention set are staged for a forthcoming archive version, with a human validation pilot named as the primary open item (Section 4, Appendix A).

Two negative by-products, a distilled 7B judge that misses its agreement target and an in-house RL agent that fails to beat its untrained base, are reported as methods findings (Section 5.12).

2 Related Work

Deep-research systems supply our reference architectures. Stanford's STORM [84] and Co-STORM [35] introduced perspective-grounded multi-agent conversation (our P4). WebThinker [55] introduced a reactive model-driven control loop (P6). MindSearch [14] introduced dynamic sub-query graph expansion (P7). Search-o1 [54] introduced reason-in-documents. A second wave trains the agent end-to-end: WebDancer [107], WebShaper [95], WebWeaver [57], Tongyi DeepResearch [96], DR Tulu [83], Pangu DeepDiver [87], and GAIR's DeepResearcher [125] (our P10, used unmodified). The AI-Scientist line [60, 113] inspires our beam search (P8). Two surveys [33, 122] catalogue the space. Neither runs a controlled head-to-head. Commercial systems [23, 72, 75, 120] define user expectations. They publish neither a fixed model nor a fixed tool layer, so they are poor material for controlled comparison. One commercial write-up [2] is a partial exception: it discloses an orchestrator-worker design and model assignment. But it still lacks the query manifest or reliability statistics a controlled study needs. The comparison gap itself is confirmed by benchmark-track studies [20, 86]. Leading agents score under 68% rubric compliance under LLM-judge grading, and a smaller human-expert validation sub-study reports comparable disagreement.

Controlled comparisons, and the confound we attack. Our patterns' primitives come from the 2022–2024 reasoning literature: ReAct [116], Reflexion [88], Tree- and Graph-of-Thoughts [6, 115], and plan-and-solve and least-to-most decomposition [101, 127]. The multi-agent frameworks [30, 110] and ensembling arguments [50, 100] motivate orchestration, though the returns to more calls are themselves non-monotone [12]. A countervailing line is empirically sceptical, converging from several independent angles. Cost-controlled evaluation shows agent gains often shrink or vanish at matched spend [37], and single agents match or beat multi-agent systems at equal token budgets [97]. Strong single-prompt agents likewise match multi-agent debate [103], and multi-agent debate itself reduces to majority voting with no systematic gain beyond it [16]. Outside the debate setting, in-context procedural prompting outperforms external orchestration frameworks on procedural tasks [19]. The most direct test yet comes from a pre-registered study of domain-expert authors judging AI-generated feedback on their own research: it finds that single-pass frontier models beat multi-agent debate, despite the debate tool spending roughly 30× the tokens [28]. Two accounts explain why: an entropy-dynamics analysis shows strong reasoning models can make *worse* orchestrators through context-squeezing [128], and the MAST taxonomy [8] catalogues the coordination failures behind these results. The design motivation closest to ours is empirical: Nechepurenko and Shuvalov [69] treat coordination as a configurable *architectural layer*, holding the base model, tools, and prompt fixed while varying only the coordination configuration. That is precisely the isolation our study demands, reached from the multi-agent side. The confound argument we make below stands on its own; we do not attribute it to any impossibility result. Concurrently, and closest in spirit to our headline, Lu et al. [61] apply leave-one-out / removal-based attribution to the components of one agent system. They report that much orchestration is redundant, a partial scoop of the "bounded returns" vocabulary. They decline full combinatorial Shapley-style attribution, showing removal-based attribution matches the combinatorial methods at lower cost. We claim no primacy on the phrase: they attribute *within* a single system, whereas we compare *across* eleven architectures at a matched tool budget with run-noise controls. The two are complementary readings of the same diminishing-

returns picture. Closest to our method, Kim et al. [40] standardise tools, prompts, and compute across five architectures, three model families, and six long-horizon, environment-interactive benchmarks (scored by a mix of execution-based checks and LLM judges), and find diminishing coordination returns past a capability threshold. Their capability-saturation law is, in effect, the mechanism behind our flat cluster. Established on agentic coding and tool-use tasks, it leaves our long-form, cross-family design for judged research-report synthesis, which pairs a rubric with IRR, outside its evidence base, as it does our model-capability and RL axes. Closest to our task, DeepScholar-Bench [74] fixes the retrieval API for academic synthesis and runs the oracle-retrieval ablation we lean on in Section 10.1. It is academic-only and reports no cross-family panel or equivalence statistics. We position our work as complementary to both: general-purpose queries, frontier and 7B agents on one axis, and the reliability and equivalence discipline a null-heavy result demands.

RL for deep research. The line of work applying RL to search, Search-R1 [36], R1-Searcher [91], ReSearch [13], ZeroSearch [93], and the GRPO recipe [18, 85], trains 7–72B models to interleave reasoning with search, generally reporting gains over base-model RAG. A recent survey [53] systematises the data, reward-design, and credit-assignment choices this line turns on. DeepResearcher (our P10) sits at the intersection of this line and the deep-research task. A counter-current argues less is more. SimpleDeepSearcher [94] matches RL with 871 curated supervised fine-tuning (SFT) trajectories, and LIMO/LIMR/s1 [56, 67, 119] recover much of the RL lift with tiny high-quality sets. Our P9-vs-P10 contrast isolates the RL effect ($\Delta = 0.078$). Our P12 negative result, read alongside DR Tulu's RL against a long-form rubric [83], supports a reward-matching reading (Section 10); we make no blanket claim about RL.

Benchmarks and the retrieval-ceiling evidence. Our 90-query manifest draws from DRACO [126], DeepSearchQA [25], ResearchQA [52], and LitQA2 [90], chosen because no single benchmark covers the breadth a deep-research agent faces. BrowseComp [105], WebWalker [108], GAIA [65], FRAMES [44], and HLE [77] stress adjacent sub-capabilities, and ReportBench scores survey-style reports for cited-literature quality and statement faithfulness [51]. The reading of retrieval as the binding constraint is anticipated by DeepScholar-Bench's oracle ablation [74] and by retrieval-gated gains on FRAMES [44]. Three concurrent 2026 results reach the same ceiling from different angles and frame our oracle test (Section 7) as its architecture-controlled instance. TaxoBench [121] reports that the best deep-research agent recalls only 20.9% of the papers experts cite. Even when structure is supplied, organisation tops out near 28–29% semantic-path similarity against a human 47–58%, a synthesis gap that sits downstream of retrieval. BrowseComp-Plus [15] makes agent comparison fair by handing every system an identical frozen corpus, the community fixed-corpus form of the same condition our pooled oracle imposes. Ours is comparable but per-query and superset in one respect: it freezes an identical document set for *every* pattern on *every* query, so the fixed-evidence control holds within the architecture contrast as well as across systems on a shared corpus. A diagnostic study finds that on multi-hop questions a training-free iterative retriever can beat a one-shot ideal-evidence oracle, raising aggregate accuracy from 69% to 81% [4]. Staged retrieval reduces late-hop failures, mitigates context overload, and corrects early hypothesis drift, yet high composition-failure rates persist even under perfect retrieval. That is the dual ceiling our Section 7 demonstrates across eleven oracle-scored architectures (quantified on the six-pattern top cluster), beyond the single-system setting.

LLM-as-judge methodology, and citation grounding. Our panel builds on MT-Bench [124], G-Eval [59], Prometheus [38], criterion-level rubric evaluation [3, 39, 118], and the argument for a panel of juries [98]. The bias literature motivates our design: verbosity [21, 80], self-preference [42, 73], style-over-substance [109], evaluator inconsistency [92], and the rating-indeterminacy framing [24] we adopt for substantive dimensions. Large human-comparison studies [5] bound what any judge can certify. A concurrent audit built specifically for research agents, REFLECT [102], finds the best LLM judges score below 55% at assessing

agent reasoning and evidence use, with evidence verification the weakest sub-task. This is the reliability ceiling our pre-comparison audit (Section 5.1) measures directly, and the reason we lead on cross-judge ordinal claims, not absolute substantive levels. A parallel result on LLM-judge self-inconsistency, where the same judge gives inconsistent scores across runs [26], reinforces the same caution; this is why we report a within-judge re-test alongside cross-judge agreement. A long-form-specific meta-evaluation likewise finds current judges unstable across scenarios, with rubrics and references helpful but not sufficient [11]. A cross-judge companion to this within-judge instability shows scores can move even when the responses being judged are held fixed [114]. The same paper reports that repeated-sample juries add little once judge errors are correlated, the mechanism our own effective-judge count formalises below. A nine-judge panel is shown to carry only about two effective independent votes [41], the redundancy problem our own effective-judge audit (Section 5.4) instantiates independently on a cross-family three-judge panel. Authority bias is already documented at the judgement level. Probes that insert fabricated references show LLM judges reward authoritative-looking citations [10], and the CALM framework quantifies authority bias as one of twelve systematic judge biases [117]. Our citation-density finding (Section 6) extends the style-over-substance and authority-bias lines to citation grounding. The citation-support audits of deployed generative search [58] and the ALCE citation-quality benchmark [22] established the measurement problem. Retrieval-verified factuality methods [66, 106] are the right instrument but, as we show, do not transfer to the synthesised-research register without a partial-support rubric, whereas a grounding rubric that supplies the source, FACTS Grounding [34], discriminates well. Two concurrent audits of deployed deep-research agents [71, 79] corroborate the density-over-grounding effect in the wild.

3 Architectures

Against that literature, our contribution is a controlled comparison; we introduce no new system. We study eleven deep-research architectures spanning two paradigm families and a deliberately wide orchestration-complexity axis. The design goal is to vary one factor at a time. Family A holds the base model (GPT-4o) and the tool layer fixed and varies only the orchestration logic, from a single forward pass (P0) through eight increasingly elaborate pipelines (P1–P8). Family B holds the architecture fixed (the P0 scaffold) and varies the model, substituting a local 7B backbone (P9) and its descendant trained with reinforcement learning (P10). Two post-hoc probes complete the picture: a verbatim ReAct controller (P11) and an in-house 7B agent trained with Group Relative Policy Optimization (GRPO) (P12), both reported on a single-judge basis (Section 5.12).

Table 1 lists the set. All Family A patterns share one Python package: the same `LLMCaller`, retry and rate-limiting logic, JSON-mode parsing, and identical tool wrappers around Bing Web Search, Semantic Scholar, arXiv, and `trafilatura`. Family B substitutes a `LocalLLMCaller` (`transformers` with 4-bit quantisation) and reuses the Family A control flow and tool wrappers verbatim. Several constants differ to fit the 7B backbone's 16.6GB VRAM budget: P9/P10 halve the generation cap (4096 vs. P0/P1/P4's 8192 tokens), cut the per-call source-extraction budget (15K vs. 50K characters), and truncate assembled evidence to 6000 words before generation, a step Family A has no equivalent of. The length-adjustment control of Section 5.9 bounds their main visible symptom (shorter P9/P10 output), and the frozen-evidence arm of Section 5.4 sidesteps the extraction-budget asymmetry entirely by serving every backbone a byte-identical, already-truncated context. The factor is cleanest for the orchestration contrast: P0 vs. P1–P8 changes only the wiring. For the model contrast the "factor" is the model *as a bundle*: scale, pre-training data, alignment, and quantisation move together, a confound we are explicit about in Section 5.4. The taxonomy supports three structured contrasts:

- **P0 vs. P1–P8** isolates *orchestration* (model and tools fixed);
- **P9 vs. P0** measures the *frontier-vs-7B capability gap* (architecture and tools fixed; the model changes as a bundle);

Table 1: The eleven architectures. Family A varies orchestration at fixed model and tools; Family B varies the model at fixed architecture. Inspirations name the published system whose control structure each pattern reconstructs.

Family	Pattern	Inspiration	Control structure
A. GPT-4o pipelines	P0 Single-pass	none	One retrieval round, one synthesis call
	P1 Iterative RAG	classical agentic RAG	Retrieve→reason→retrieve until reflection is satisfied or 3 rounds elapse
	P2 Supervisor	supervisor–worker	Plan, fan out parallel sub-agents, fuse
	P3 MERIDIAN	none	Topic mining + quality-evaluator self-loop
	P4 STORM	STORM / Co-STORM	Multi-perspective conversations + triangulation
	P5 Hier. W&D	none	Width-and-depth schedule + meta-evaluation
	P6 Reactive	WebThinker	Model decides search/read/write each step
	P7 Graph	MindSearch	Dynamic sub-query graph expansion
	P8 Beam	DeepDiver / AI-Scientist	Explore K research-direction hypotheses, prune to a surviving beam
B. Local 7B	P9 Qwen2.5-7B	none	P0 scaffold, 7B backbone
	P10 DeepResearcher	GAIR DeepResearcher-7b	RL-trained end-to-end tool-use agent

- **P9 vs. P10** varies *RL training* (same backbone and tools; P10 additionally runs a learned multi-search loop, so the contrast bundles RL with that scaffold).

P3 (MERIDIAN) and P5 (Hierarchical Width-and-Depth) replicate no single published system: they recombine topic-mining, self-evaluation, and width/depth-scheduling motifs that recur across the prompt-engineered family, and serve as deliberately over-engineered points on the complexity axis.

4 Experimental Design

Four design elements carry the study: the query manifest, the shared tool layer, the rubric, and the judge panel, followed by the pre-specified inference plan that governs every comparison.

4.1 Query manifest

The evaluation manifest contains 90 queries drawn from five sources and stratified by difficulty. The released manifest comprises 40 DRACO, 20 DeepSearchQA, 15 ResearchQA, 10 LitQA2, and 5 custom queries, labelled `simple` (5), `moderate` (25), and `complex` (60).[1] Each query carries a natural-language prompt, an optional gold answer, and a list of expected coverage topics consumed by the rubric. Domains span law, biomedicine, finance, the arts, and consumer decision-making.

4.2 Tool layer

Every one of the eleven patterns calls an identical retrieval *interface*: Bing Web Search v7 (top-10 results), the Semantic Scholar Graph API and the arXiv API for academic retrieval, and `trafilatura` for page extraction under a 200 KB content cap. All patterns share one `httpx.AsyncClient` singleton and an identical eight-attempt exponential backoff. GPT-4o patterns share a global rate gate (semaphore N=12, 200 rpm). The local

[1]All pattern means, dimension scores, regression coefficients, and verdict counts in this paper are regenerated directly from the underlying parquets by the accompanying analysis pipeline (see "Reproducibility and Artefact Availability" below for what is and is not yet in the public archive). Two judge-reliability probes, the within-judge re-test (κ=0.91) and the self-preference permutation (p=0.82), are carried from the evaluation phase that produced them and are the only two figures in the paper that sit outside the guarantee of regeneration and reconciliation.

7B patterns serialise on a single RTX 5080 (16.6 GB). Two things are held constant across patterns: the *model* and the tool *code*. The evidence pool is free to vary. Search-query strategy is intrinsic to each architecture (P0 searches the verbatim query; P1 issues up to 25 decomposed sub-queries; P4 issues per-perspective plans), and the search cache is keyed by query string, not by pattern. So each architecture *induces* its own retrieved and cited evidence set. The cross-pattern overlap is near-zero: across the nine GPT-4o patterns and 90 queries the cited-URL sets are predominantly disjoint, with no query approaching a half-shared set. We therefore read pattern-score differences as reflecting *architecture together with the retrieval it induces*, not as isolating reasoning over a shared corpus. Where a score must be credited to architecture alone, the canonical fixed-evidence control is the oracle backend of Section 7, which freezes one identical evidence corpus per query for every pattern. The dominant web channel (Azure `web_search_preview` via Bing, $\approx 3.5\times$ larger than the academic channel) was healthy throughout corpus generation. The academic sub-channel ran with Semantic Scholar down (429 rate-limited throughout), so academic retrieval was effectively arXiv-only ($\approx 97\%$ arXiv). This condition was uniform across all patterns and queries, so it confounds no cross-pattern contrast and corrupts no report (Section 11).

4.3 Rubric

Reports are scored on nine dimensions under a binary-criterion decomposition in the style of FLASK, BiGGen-Bench, and HealthBench. The dimensions and their default weights are `information_recall` (0.20), `factual_accuracy` (0.20), `coverage` (0.10), `analytical_depth` (0.15), `citation_quality` (0.10), `logical_coherence` (0.05), `organization` (0.05), `instruction_following` (0.10), and `attribution_quality` (0.05). Each dimension carries two to eight general binary criteria (38 in total) plus task-specific `coverage` criteria derived per query from the manifest's expected topics. Each judge returns a chain-of-thought verdict per criterion and a per-dimension score in $[0, 1]$. The overall score is the weighted sum. Weights are query- and source-dependent (the canonical weight hash is fixed by the released manifest). Section 5 confirms that rankings are robust to the weighting scheme.

4.4 Three-judge panel

We score every report with three frontier judges chosen from distinct model families: GPT-5.2 (Azure OpenAI, an endpoint that is not on provisioned throughput),[2] Claude Opus 4.1, and Claude Sonnet 4.5. Each (pattern $\times$ query) cell is judged by all three where compute permitted, and coverage is uneven by design: GPT-5.2 scores every cell, the two Claude judges fewer. Across the full scored set, which includes ablation, oracle, and variance arms that are single-judge by design, 39% of cells carry all three verdicts. Restricted to the eleven base patterns that every headline rests on, 99.8% carry all three (983 of 985). We report panel-averaged scores only where the panel is present. (The post-hoc P11/P12 probes are single-judge by design; including their cells drops the coverage figure to 84%.) A fourth scorer, a version-bumped Claude run as a reproducibility probe, also appears in the underlying verdicts but is excluded from the banked panel average and every certified headline separation. Current-generation Claude models are used elsewhere only as explicitly labeled, standalone crosscheck analyses (e.g., Section 7.1, Section 6), never folded into a headline verdict. A family-of-judges independence constraint runs through the study: no judge may share a pre-training family with a pattern it scores. This is why Qwen- and DeepSeek-family models are excluded as candidate judges (they are the backbones of P9/P10). It is also why the distilled DR-Judge-7B of Section 5.12 is barred from scoring P9 and P10. The panel's inter-rater reliability is reported in Table 5 and audited in Section 5.

[2]The judge is a specific, dated Azure model deployment pinned for the duration of the study; we never call the rolling ChatGPT-facing endpoint. Azure deployments do not silently upgrade unless explicitly configured to. GPT-5.2 is no longer OpenAI's current flagship model as of this paper's submission window, but every verdict in the underlying corpus was scored against the same fixed instance. So reproducibility does not depend on the model's continued availability as a frontier system.

4.5 Pre-specified analysis

Inference follows a four-gate hierarchy designed to bound multiple-comparison risk and to remain robust to distributional violations. "Pre-specified" throughout this paper means fixed in the internal analysis plan before the results in this manuscript were written up. We filed no timestamped registration with an external registry (e.g., OSF or AsPredicted). Several downstream specification choices (equivalence margins, length-adjustment functional forms, rubric-weighting variants) were made during analysis. Each of the four gates asks the same question, whether two architectures' scores really differ, at a progressively stricter standard, so a claim that survives all four is on firmer ground than one that only clears the first. **Gate 1** fits a crossed random-effects model overall $\sim$ pattern + (1 | query) + (1 | judge) on the un-aggregated data and tests the pattern fixed effect by likelihood ratio: a single omnibus test asking whether the eleven patterns differ *at all*, once query-to-query and judge-to-judge variation are accounted for. **Gate 2** refits Gate 1's test separately on each of the nine rubric dimensions, with a Holm correction, an adjustment that keeps the false-positive rate honest when many tests are run at once, applied across the nine. **Gate 3** runs paired Wilcoxon tests for all 55 pattern pairs *separately per judge*, Holm-corrected within judge. (A Wilcoxon test is a standard non-parametric test for whether two paired sets of scores differ.) A pair is "judge-robust" only if it is Holm-significant with consistent direction in all three judges. **Gate 4** reports Bayesian signed-rank contrasts for the headline comparisons, with regions of practical equivalence in the sense of Kruschke [45]. This Bayesian version of the same paired test can certify two patterns as equivalent, where a plain significance test can only fail to find a difference. We additionally fit two one-sided tests (TOST), which ask whether two things are the same within a stated tolerance [47, 82], and report Spearman ranking concordance with Fisher-z intervals, the standard confidence-interval construction for a correlation coefficient. Holm's step-down procedure [29] controls each declared test family. The equivalence margin, discussed in Section 5, is matched to the minimum detectable effect of the design, with no substantive threshold behind it; we therefore label it *power-justified*. It was not part of the pre-specified plan. We report results at both ± 0.05 and the stricter ± 0.02 region of practical equivalence (ROPE, the tolerance band a TOST test checks against).

A measurement caveat that propagates everywhere: Claude Sonnet's *stored* overall score is corrupted upstream, so all analyses recompute Sonnet's overall from its per-dimension verdicts. The discrepancy this resolves is documented in the released data dictionary.

5 Results

We present results in the order the inference depends on them. The instrument comes first, then the pattern comparisons it licenses, then three robustness controls on the one gap that matters, then the replications, and finally two apparatus by-products.

5.1 Judge reliability-audit precedes pattern comparison

We reliability-audit the measuring instrument before using it, using two standard inter-rater-agreement statistics that both run roughly from 0 (no better than chance) to 1 (perfect agreement): Krippendorff's α and the intraclass correlation coefficient (ICC). On the overall score the three-judge panel attains Krippendorff $\alpha = 0.42$ [43] and a single-rater ICC(A,1) $= 0.49$ [89] (95% CI $[0.08, 0.72]$). This rises to ICC(A,k=3) $= 0.74$ ($[0.21, 0.88]$) for the panel-averaged score we actually analyse (Table 5). These are point estimates in the conventional "good" band, but the intervals are too wide to certify it. ICC(A,k) also inflates mechanically with k and should not be compared to single-rater values elsewhere. These statistics are computed on the 983 complete-case cells and reproduce to four decimals under an independent implementation (`pingouin`). The overall α is unchanged to four decimals on the wider 985-unit available-case set, and the query-vs-judge variance split holds under a single fixed REML estimator (restricted maximum likelihood, a standard way of fitting variance components without also having to search over multiple candidate model fits). The agreement is sharply dimension-dependent: structural dimensions agree well (`organization` $\alpha = 0.84$) while substantive ones agree weakly or not at all (`factual_accuracy` 0.13, `information_recall` 0.11,

attribution_quality −0.10). Following Guerdan et al. [24] we read the weak substantive agreement as *rating indeterminacy*: even a perfect judge would disagree here because the construct itself is under-defined; judge noise is not the driver. A within-judge re-test reinforces this reading, GPT-5.2 re-scoring a stratified subset at Cohen's $\kappa = 0.91$. The judges are individually *reproducible* yet *disagree* on substance. They can certify relative orderings, not absolute substantive levels. This boundary governs every claim below. A concurrent, much larger-scale audit lands on the identical distinction: across 21 judges and ≈541,000 judgments, agreement metrics overstate true reliability by 33–41 percentage points and judge rankings shift by up to 14 positions across benchmarks [70]. This is the same gap our own panel shows at a much smaller scale: a panel can be reliable (consistent) without being valid (correct). Bias regressions show small, non-zero length and citation-count effects ($\beta(\log \text{wc}) \approx 0.04$–$0.11$) and a consistent local-7B penalty, all small relative to the pattern effects. A self-preference permutation test is null ($p = 0.82$), and it is structurally able to detect only GPT-family preference. We therefore analyse the panel average and additionally require cross-judge robustness for every headline separation.

Reproducibility is not validity, so we also test whether any judge's factual verdicts track ground truth. Against a mechanically checkable verifiable-answer gold slice (LitQA2 + DeepSearchQA; judge coverage differs by design, Section 4: GPT-5.2 n=322, Claude Sonnet n=290, Claude Opus n=259), GPT-5.2 is the only judge whose answer-sensitivity excludes chance (Figure 2). AUC here is the area under the ROC curve, a standard 0–1 score for how well a judge separates correct from incorrect answers, where 0.5 is chance. Its factual-accuracy AUC is 0.66 (correct-minus-incorrect delta 95% CI $[0.024, 0.107]$). Both Claude judges sit near chance (0.54 for Opus, 0.50 for Sonnet), both intervals spanning zero, a null we read as absence of evidence for their sensitivity rather than evidence of blindness. On the matched common report set a paired DeLong test, the standard test for comparing two correlated AUC values, confirms the gap: GPT-5.2 over Sonnet +0.157 (p=0.0003) and over Opus +0.107 (p=0.031). Only the GPT-5.2-over-Sonnet asymmetry survives the more conservative query-clustered bootstrap. The Opus arm rests on DeLong alone. Three caveats keep this an anchor rather than a certificate. First, the answer signal tracks whole-report completeness, within which the factual dimension is only one component. Second, the positive class clusters on only 11 of the 27 verifiable queries, so the AUC's effective evidence base is small. Third, even the answer-sensitive judge shows only poor chance-corrected agreement with the gold labels ($\kappa = 0.17$), the familiar gap between ranking and agreement. The operative consequence is the cross-family validity asymmetry: the one judge we lean on for the independent headline is the one whose factual verdicts demonstrably move with truth.

5.2 The pattern effect is real, and unevenly distributed

A pre-specified crossed random-effects model (Gate 1) finds an overwhelming pattern effect (likelihood-ratio p far below any threshold). We do not lean on the exact p-value: it is inflated by within-cell judge correlation (pseudo-replication). We rest the headline separations on the query-paired, per-judge Gate-3 tests. The variance decomposition is the substantive output: the largest *non*-pattern variance source is query difficulty (ICC(query) ≈ 0.45, stable across optimisers); judge stringency is smaller still (ICC(judge) = 0.18). Which query is asked therefore matters more than which judge scores it. This justifies analysing judge-averaged scores while reporting per-judge robustness.

5.3 On the grand mean, orchestration is inseparable, but the leader is task-conditional

Table 3 and Figure 1 give the headline. The single-pass baseline and eight orchestration pipelines span 0.49 (P0) to 0.67 (P1), and six of them, P1, P4, P5, P6, P7, P8, occupy a band (0.60–0.67) at the top of that range. Five of the six are judge-robustly inseparable, and the sixth, P5, joins only under uncorrected equivalence testing at the ±0.05 margin. Throughout the paper, *the cluster* names these six patterns and *the inner five* the cluster minus P5. This architecture sense is unrelated to the statistical sense of "cluster" used later for standard-error corrections (cluster-robust, wild-cluster bootstrap, query-clustered); the two never refer to the same thing, and we flag the distinction again where the two senses fall close together. This paper reuses a

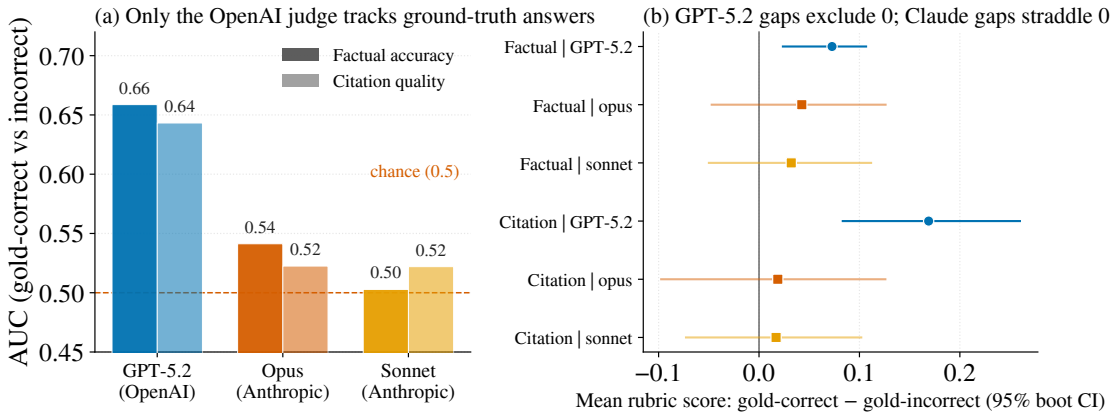

Figure 2: Only GPT-5.2's factual verdicts track verifiable ground truth. Judge factual-accuracy verdicts against the mechanically checkable verifiable-answer gold slice (LitQA2 + DeepSearchQA; judge coverage differs, GPT-5.2 n=322, Claude Sonnet n=290, Claude Opus n=259). GPT-5.2's AUC of 0.66 excludes chance (correct-minus-incorrect delta 95% CI [0.024, 0.107]) while Opus (0.54) and Sonnet (0.50) sit at chance. Even the answer-sensitive judge agrees with gold only weakly (κ = 0.17), the familiar ranking-versus-agreement gap. Panel (a)'s y-axis is truncated below, starting at 0.45 where a full axis would start at 0, to resolve the spread around the drawn 0.5 chance line, which is itself plotted.

small number of words for more than one purpose; the table below is a standing reference, not new content, for the four that recur most:

Term	Sense A	Sense B	Sense C
cluster	architecture: the six top patterns just defined	statistics: clustered/robust standard errors (Sections 6, 9 and 11)	–
family	paradigm: an architecture family (Section 3)	judge lineage: a same-lab judge pair or a scored pattern's model lineage (Sections 4 and 5.4)	statistical: a Holm-corrected multiple-testing family (Section 4)
oracle	the pooled-ideal-source retrieval intervention (Section 7)	the held-out-criteria selector in the best-of-N probe (Section 5.7)	–
arm	one of the eleven base patterns	one of the four local-model-curve configurations (Section 5.4)	one of the seven external bake-off systems (Section 5.11)

We establish the cluster claim two ways, leading with the stricter, method-stable criterion.

Judge-robust separations (primary). Of the 55 pattern pairs, 26 are Holm-significant with a consistent direction in *all three* judges (within-judge counts 30, 40, and 42) — our strict bar for a real difference, and invariant to the correction family: a single joint Holm family over all 165 judge-level tests yields the same 26 pairs. Every one of the 26 crosses a capability tier; none separates peers within one:

- Eighteen involve a local-7B agent: twelve cluster-vs-7B, four mid-tier-vs-7B, two baseline-vs-7B.
- Three cluster members separate from the baseline P0: P1, P4, and P6 clear the tri-judge bar against P0; P5, P7, and P8 do not.
- Five cluster-member pairs clear the bar against the weaker GPT-4o pipelines P2 and P3 combined (not five against each).

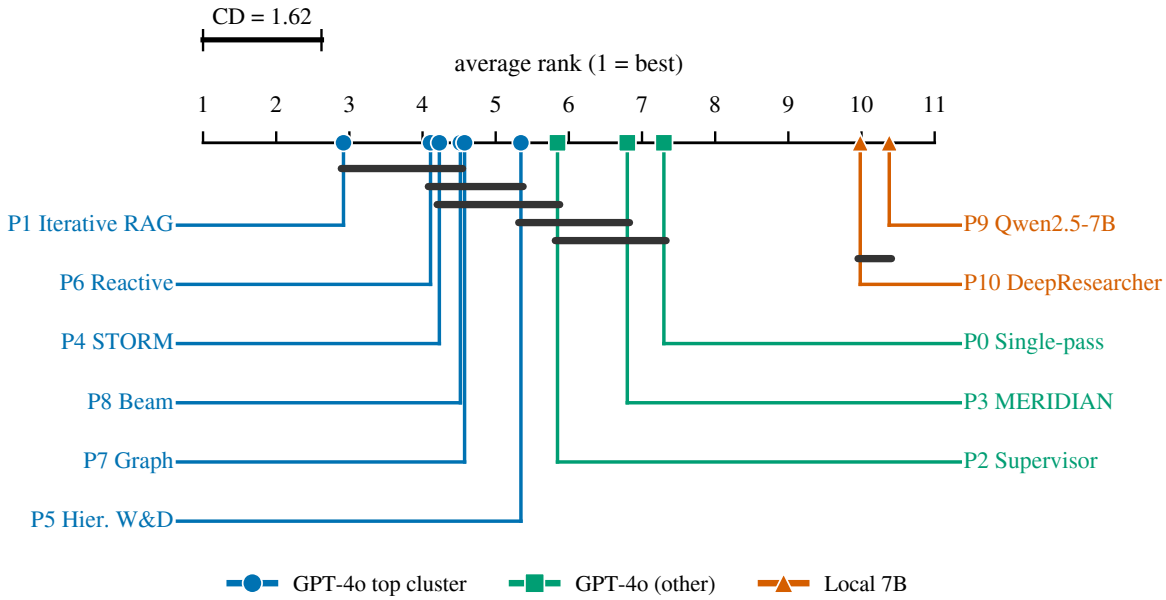

Figure 3: Friedman–Nemenyi critical-difference diagram over the eleven base patterns (average rank of the three-judge mean per query, N=87 queries with complete coverage; lower is better). Patterns joined by a bar differ by less than the critical difference (CD = 1.62) and are not significantly separated. The picture agrees with the stricter Gate-3 judge-robust criterion in substance: the GPT-4o pipelines cluster and the local-7B agents separate cleanly on the right. The coarser rank-based test draws one bar over four of the inner five (P1, P6, P4, P8), with P7's average rank marginally past the critical difference from P1 (1.66 vs. 1.62); the judge-robust criterion, which finds no separable pair among the five, is the one we rest on.

Zero of the ten pairs *within* the inner five (P1, P4, P6, P7, P8) clears the bar, and neither do P2-vs-P3, P0-vs-P2/P3, or P9-vs-P10. The critical-difference diagram (Figure 3) shows the same tiering under a coarser rank-based test.

Equivalence testing (secondary). Adding P5 gives a six-pattern cluster (15 pairs). Under paired-Wilcoxon TOST at a ±0.05 region of practical equivalence, 9 of 15 pairs are formally equivalent (6 of 15 after Holm correction across the fifteen-pair family). Under the more conservative t-based TOST the count is 6 of 15, all six of which survive Holm. The two surviving sets coincide: exactly the pairs within {P4, P6, P7, P8}, with the three P5 pairs the multiplicity casualties. At the tighter, substantively motivated ±0.02 margin the count is 0 of 15 under both tests, corrected or not. The ±0.05 margin is *power-justified*, set at roughly twice our main-study minimum detectable effect (MDE_{80} = 0.025 at n=90, one replicate), without independent substantive motivation. A replicate experiment shows per-run noise (σ_{run} = 0.05) of the same order. A current-Claude crosscheck of the same replicate runs agrees that the replicate-based pattern ordering survives a change of judge family (minimum pairwise rank agreement across the three judge families Spearman ρ = 0.80), though absolute levels differ by judge leniency as elsewhere: the quantity that is robust across judge families here too is the ordering, not the level. This is exactly the regime a concurrent single-run audit warns about: Alvarado Gonzalez et al. [1] report that single-run evaluation on model leaderboards inverts at least one pairwise rank in 83% of evaluation slices relative to a repeated-run reference. This corroborates our decision to lead on run-robust separations over single-run ones. We do not certify any inner-cluster ordering, because Alvarado Gonzalez et al.'s boundary case runs the other way for us. Their significance conclusions were robust to run noise *because* the gaps they tested were large. Our within-cluster gaps are small, of the same order as σ_{run}. We therefore treat equivalence as corroborating, not decisive: the defensible claim is the method-stable one, *no within-cluster pair is judge-robustly separated*.

What we cannot claim is that the six are *identical*, only that we cannot tell them apart at the power we have.

Where the evidence pushes against flatness. Three signals run the other way. The coarser Friedman–Nemenyi rank test places P7's average rank marginally past the critical difference from P1 (1.66 vs. 1.62; Figure 3). Opus alone puts P1 ahead of P4 by +0.12. And the external-validation battery's pre-specified three-pattern flat-cluster check formally returns FALSE, driven by a P1↔P4 swap of 0.034 (Table 12). One criterion governs all three: a separation counts if it is Holm-significant with a consistent direction in all three judges and stable against run noise ($\sigma_{\text{run}} \approx 0.05$). None of the three meets it, and each is judge- or subset-local. Read directionally, all three favour P1, the same micro-order the naive mean shows and that we decline to certify. Single-judge evidence cannot meet the criterion by construction, which is why the paper's single-judge probes are everywhere labelled uncertified. The strict criterion costs power: the quoted $\text{MDE}_{80} = 0.025$ is the single-judge mean-comparison floor, and the effective minimum detectable effect of the tri-judge Holm criterion is necessarily larger. So "no judge-robust separation" bounds within-cluster differences loosely; 0.025 is too tight a figure to hang on it.

Two redundant judges carry most of the panel's orchestration gain, and their wider score range alone cannot account for it. The baseline-to-cluster gap the three-judge panel reports is 0.145, but the family-clean GPT-5.2 judge alone puts it at 0.066 (Table 14). "Family-clean" here means independent of the Claude judge pair; it is a different family relationship from Section 5.4's concern, where GPT-5.2 instead shares lineage with the GPT-4o patterns it scores. Over half the panel gain is contributed by the two same-lab Claude judges. Because raw cross-judge gaps conflate a judge's effect with its score range, we re-express the gap in scale-free units, using Cohen's d, a standard effect-size measure that expresses a gap in units of its own standard deviation. The paired effect size of cluster-over-P0 is Cohen's $d = 0.90$ for Opus and 0.75 for Sonnet against 0.47 for GPT-5.2. Both Claude judges report a substantially larger effect than GPT-5.2, and the difference excludes zero for both under a within-judge z-scored gap too:

Judge	Cohen's d (cluster-over-P0)	Paired difference from GPT-5.2
Opus	0.901	+0.43 (95% CI $[0.26, 0.63]$)
Sonnet	0.755	+0.29 (CI $[0.09, 0.51]$)
GPT-5.2	0.468	—

The Claude judges therefore genuinely report a 1.6–1.9× larger orchestration effect. This *supports* the judge-redundancy thesis (Section 11), since a correlated same-lab pair inflates the apparent orchestration effect. The conservative reading of the cluster gain is therefore the smaller cross-family one, which is still bounded and still positive. We return to this cross-judge split when the panel disagrees on the bake-off leaderboard (Section 5.11).

Robustness. The ordering is stable across four rubric-weight schemes (minimum Spearman $\rho = 0.94$), and it does not depend on the citation dimension that Section 6 shows to be judge-confounded. Re-weighting the overall score with `citation_quality` removed entirely preserves the tier structure (the same six patterns form the top cluster; Spearman $\rho = 0.97$ against the full rubric), with only the already-inseparable within-cluster micro-order shuffling. One confounder does bite: report length affects the gap's *size*, and we bound it separately in Section 5.9. Nor does the ordering ride on the one source that supplies 40 of the 90 queries. A leave-one-source-out jackknife that drops each of the five source families in turn (DRACO among them, dropping 40 of the 90 queries) leaves P1 rank-1 in every fold, with a maximum rank displacement of 3 places and only inside the already-flat top cluster. The naive panel mean ranks P1 first, but this is an Opus artefact: $\Delta(\text{P1}-\text{P4})$ is +0.12 for Opus versus ≈ 0 for the other two judges. We therefore report P1 $\approx$ P4 inside the cluster; the data do not support the stronger claim "P1 wins." The *identity* of the per-source leader is not pattern-intrinsic either (Table 8). P1 leads outright on three of the five sources (DRACO, LitQA2, ResearchQA), trails P7 by only 0.02 on DeepSearchQA, and trails P6 by 0.08 on Custom. These per-source leads are descriptive, not certified: the subsets are small (Custom holds five queries, too few to separate any pair),

and no rotation is individually significant (the DeepSearchQA P7-over-P1 lead has p=0.78). The table is exploratory, with no family-wise guarantee (Section 11). What is statistically significant is the pattern×source *interaction* itself (likelihood-ratio 187.2, df 40, $p < 10^{-15}$, fit on judge-averaged rows over the eleven base patterns). That is, the orchestration premium is source-conditional: no particular pattern owns a particular source. Individual queries show architecture effects far larger than any mean: on one relocation query P4 beats P0 by 0.71 (0.72 vs. 0.01), while on a narrow factual query P1 beats P4 by 0.36 (0.76 vs. 0.40) (Section 9). The flat grand mean is therefore the signature of large, sign-flipping, task-conditional architecture effects that average out, consistent with the per-task range reported by Kim et al. [40]. We read our result as *no architecture dominates across this query distribution*. The open question it raises is *which* architecture wins *when*. The ordering is also robust to panel missingness: the 983 cells judged by all three panel judges reproduce every per-pattern mean to within 0.002 and preserve the rank order exactly.

The premium is competence-conditional. Reading the flat grand mean as "orchestration is inert" would overstate it, because the gain from orchestrating is not constant across the distribution. Figure 4 plots the gap between P0 and the cluster mean by benchmark source, ordered from the source that is hardest for a single pass to the easiest. Where P0 is already competent (RESEARCH-QA, CUSTOM, with P0 near 0.69), orchestration adds almost nothing, a premium of 0.03 to 0.04. Where the single pass collapses (DEEPSEARCH-QA, P0 at 0.32) orchestration adds 0.23, and on DRACO it adds 0.18. The pattern×source interaction is significant (likelihood-ratio 187 on 40 degrees of freedom). The premium is largest exactly where a single pass is weakest, and vanishes where a single pass already suffices. This is what scopes the thesis. The flat quantity in our headline is the spread *within* the cluster. The P0-to-cluster gap is a separate, sizeable quantity (≈ 0.15) that is itself competence-conditional. Because the manifest is weighted two to one towards the hard sources (60 of 90 queries), the grand-mean premium if anything *under*states the easy-source ceiling. The returns to orchestration are therefore bounded only *once the single-pass baseline is competent on the task*. These per-source premiums are raw-scale quantities: the length channel of Section 5.9 runs through them too (a collapsing P0 writes little as well as badly), and per-source length-adjusted premiums are not separately certified. We read the stratification as locating *where* the raw premium concentrates.

No architecture dominates; blending them reduces risk in-sample, less so out of it. This subsection borrows three cross-field lenses, portfolio theory, ecological niche theory, and bet-hedging theory, to ask one underlying question three ways: does spreading queries across several architectures pay off, and does any single architecture hold an exclusive advantage large enough to matter? Read the eleven architectures as a portfolio of eleven financial assets, with one return observation per asset for each of the 87 queries with complete three-judge coverage across all eleven. Under that reading, the single best architecture (P4: mean 0.642, risk 0.076, Sharpe ratio 8.44, a standard measure of return per unit of risk) does not minimise in-sample risk.

Concretely: a portfolio that blends architectures can beat the single best architecture on risk without giving up much return. Two long-only portfolios, both fully invested, fit and evaluated in-sample on the same 87 queries, illustrate this. The *minimum-variance* portfolio, whose sole objective is to minimise risk, puts 45% of its weight on P4 and splits the rest across P3, P7, and P8, for a portfolio risk of 0.067, below *every* single architecture's own risk including P4's, at a Sharpe ratio of 9.51. It *Pareto-dominates*, beats on both mean score and risk simultaneously, eight of the eleven single architectures, at the cost of 0.008 of P4's own mean. The *maximum-Sharpe (tangency)* portfolio applies a different weighting, one whose objective is the return-to-risk ratio.

Out of sample, though, most of that advantage shrinks: blending still lowers risk reliably, but no longer reliably beats the best single architecture on return. Both in-sample figures are, by construction, an upper bound on what blending can deliver, and a leave-one-source-out check, refit on four sources and evaluated on the held-out fifth, tempers each substantially. The maximum-Sharpe portfolio beats the best single archi-

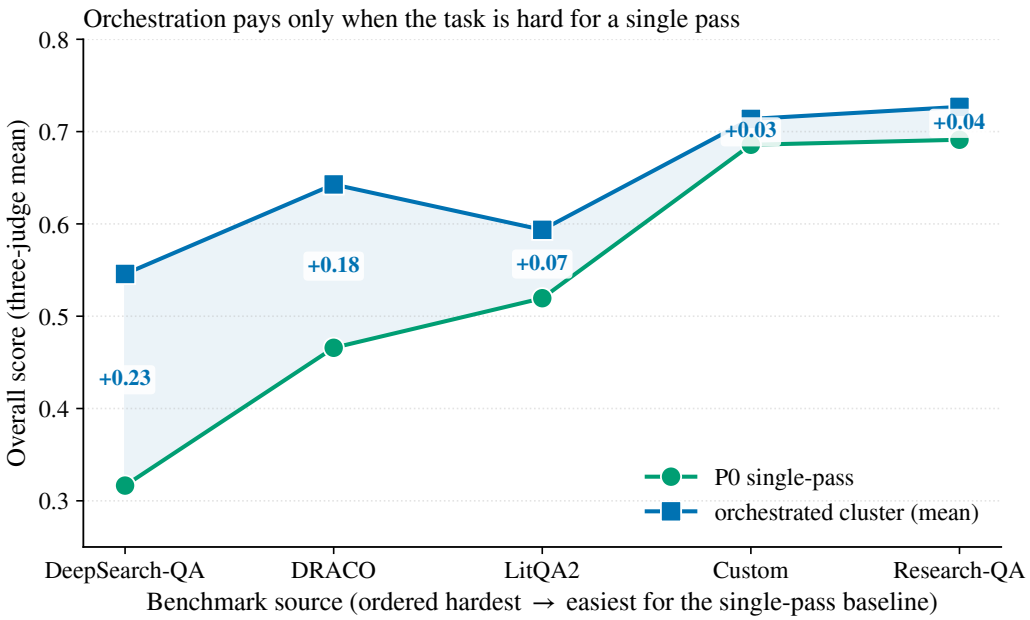

Figure 4: The orchestration premium is competence-conditional. Gap between the single-pass baseline P0 and the orchestrated-cluster mean (three-judge score) by benchmark source, ordered hardest to easiest for a single pass. The premium is 0.23 where P0 collapses (DEEPSEARCH-QA) and falls to 0.03–0.04 where P0 is already strong (CUSTOM, RESEARCH-QA); the pattern×source interaction is significant. Gaps are raw-scale (Section 5.9).

tecture in only 3 of 5 held-out folds (mean Sharpe improvement −4.25, dragged negative by the two smallest folds where a single architecture's held-out Sharpe is inflated by low sample variance; median improvement +0.08, the more robust summary). The minimum-variance portfolio's risk sits below the least-risky single architecture in 3 of 5 folds; the remaining two folds do not clear that bar. One structural property does survive out-of-sample cleanly: the blend's *diversification ratio*, how much less risky the blend is than the query-weighted average of its components' own individual risks, where a ratio above 1 means the blend wins, exceeds 1 in every held-out fold (1.23–1.65). So blending several architectures is reliably less risky than its individual components, even in folds where it does not beat the single best architecture outright. This is a weaker claim than "a deployer need not pick one." A separate, cross-validated test of essentially this question, a leave-one-query-out router choosing per query while the portfolios above hold a single in-sample weighting fixed, finds no significant improvement over picking the single best-fixed architecture (Section 12). So combining architectures is not established here as a practical, deployable win. It is established only as an in-sample property that is structurally robust, but modest.

Two further cross-field re-analyses of the same eleven-way spread corroborate the weaker, "no architecture dominates" reading; they do not corroborate the stronger blending claim. A Pianka [78] *niche-overlap* analysis, an ecology measure of how much two species compete for the same resources, applied here to per-query win-shares finds no architecture holds an exclusive niche large enough to matter: pairwise overlap is 0.48–0.90 across the top six, meaning even the most distinct pair of architectures still wins on substantially overlapping sets of queries. A second check applies Cohen [17]'s *geometric-mean bet-hedging* criterion to per-source quality. The criterion comes from evolutionary biology: it favours whichever option is safest across an uncertain mix of conditions, which need not be the option with the best average. This check complicates the niche-overlap reading above: the arithmetic-mean and geometric-mean rankings over sources are identical (Spearman 1.00, zero of eleven arms changing rank). P1's raw lead therefore holds up under an averaging rule that penalises volatility across sources: P1 is the choice the bet-hedging criterion favours under source-composition uncertainty even though the per-source table above shows it losing on individual

sources.

The two checks above examine the architectures; a third turns the same correlated-jury logic on the panel itself. A Ladha [46] analysis of judge votes against mechanically verifiable answers reaches the redundancy question through judge-error correlation, a different route from criterion-pairwise agreement. Because it draws on the same three-judge panel, its independence from the main analysis is partial at best. It lands close to the same judge-redundancy discount from a different statistical model (N_{eff} = 1.57, 95% CI [1.43, 1.77], against Section 5.4's 1.65), corroborating that figure. None of this overturns the judge-robust separations established above; it restates them as a portfolio problem.

5.4 A capability gap that dwarfs architecture, without a clean "scale" story

Holding architecture and tools fixed and swapping the model, the single-pass GPT-4o pipeline (P0, 0.49) outscores its Qwen2.5-7B twin (P9, 0.26) by $\Delta \approx 0.23$. The top-cluster pipeline P4 sits Δ(P4−P9) = 0.38 above the untrained 7B agent and Δ(P4−P10) = 0.30 above the stronger, RL-trained one (Figure 1; the top-cluster mean sits ≈0.38 above P9 and ≈0.30 above P10). On the raw scale, this gap is more than twice the entire baseline-to-top GPT-4o range (0.185) and about five times the intra-cluster range (0.073). On the length-adjusted scale of Section 5.9 the capability gap (≈0.11) still exceeds the 0–0.05 orchestration bound by at least twofold, so the ordering, though not the fivefold ratio, is scale-robust. Bayesian signed-rank contrasts place every cross-family comparison beyond the region of practical equivalence (ROPE) with posterior 1.00: the model's own Bayesian estimate of how likely the effect is real, a different question from a p-value. Because contamination is not symmetric across this contrast, we also recompute all three capability gaps – P0-vs-P9, the top-cluster pipeline P4, and the six-pattern cluster – on the 17-query contamination-clean residual: every one attenuates somewhat but still excludes zero, so the gap survives decontamination. Section 11 tabulates the full-set and clean-slice values side by side. Two cautions keep this from being a parameter-count law. *First, the contrast is not single-factor.* P9/P10 differ from GPT-4o in more than size alone: pre-training corpus, alignment stack, and 4-bit quantisation all differ too, so the label is a *capability* gap between the frontier and small open models; we avoid the word "scale" for it. P9's signature failure, flooring on 17 of 20 DeepSearchQA queries (Section 9), reads as an alignment/format collapse; nothing here resembles smooth size-driven degradation. *Second, capability at this task does not track size.* A contemporaneous 8B agent trained against a task-matched long-form rubric reward reaches near-frontier quality [83]. Orchestration's failure to close the gap therefore reflects a capability deficit that model scale *or* task-matched post-training can fill, neither of which our fixed scaffold supplies.

We de-confound the two sub-axes of "capability" directly with a four-arm local-model curve (Figure 5) run under the frozen P9 scaffold: P9 itself (Qwen2.5-7B), plus three further post-hoc probes on the same scaffold, P14 (DeepSeek-R1-Distill-Qwen-7B), P13 (Qwen3-8B), and P17 (Qwen2.5-14B-Instruct). All four arms decode on one identical backend (`llama.cpp` GGUF greedy) over the same hash-pinned 89-query frozen source set and are scored by the family-independent GPT-5.2 judge. Holding release vintage fixed at 2024-09 and doubling parameters from 7B to 14B lifts the overall score by +0.025 (95% CI [0.008, 0.043], paired bootstrap p=0.005). This is a clean capacity effect, with the decode backend held constant. Length control matters and reorders the raw ranking. The raw leader is a newer 8B model (0.356), but much of its lead is verbosity. Once each arm is placed at grand-mean output length (shared slope +0.175 per thousand words), the length-adjusted best arm is the 14B model (0.311 vs. the newer 8B's 0.299 and the 7B baseline's 0.278). Capacity therefore moves the score in the expected direction on the in-house curve, but the effect it isolates is small relative to the frontier gap – again consistent with capability being the binding axis rather than raw parameter count. A current-Claude crosscheck of this four-arm curve agrees on both the capacity direction and the length-adjusted reordering to the 14B arm, using the identical length-control and paired-bootstrap machinery as the banked GPT-5.2 result.

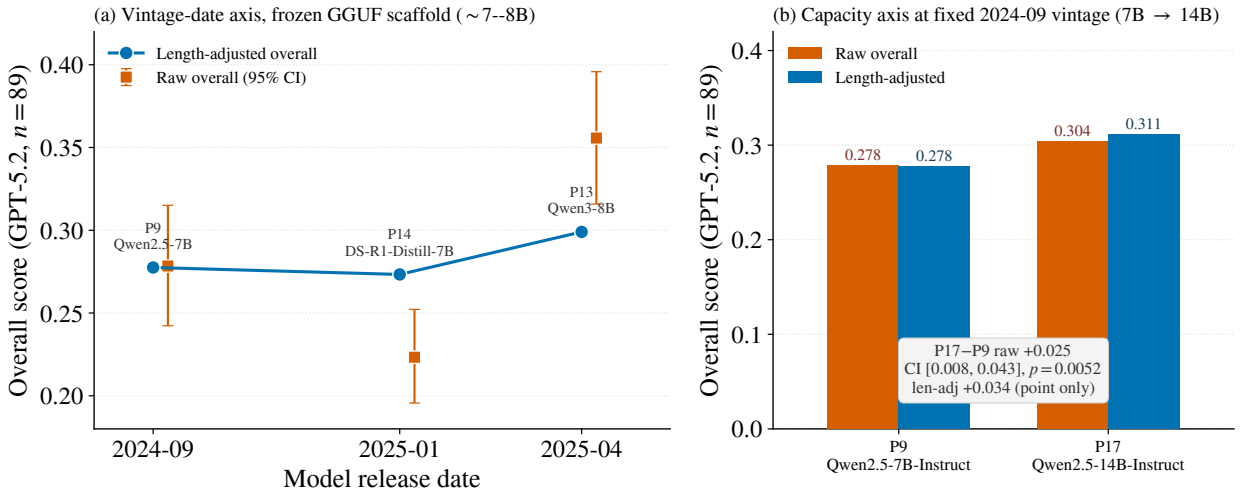

Figure 5: Capability is the binding axis on the local-model curve; raw parameter count does comparatively little work. Four arms run under the frozen P9 scaffold, all decoding on one `llama.cpp` GGUF greedy backend over the same hash-pinned 89-query frozen source set and scored by the family-independent GPT-5.2 judge; length is controlled by pooled OLS (shared slope +0.175 per thousand words, so each arm's length-adjusted score is its counterfactual at grand-mean output length). (a) The release-*vintage* axis at roughly constant 7–8B capacity: Qwen2.5-7B (P9, 2024-09, length-adjusted 0.278) → DeepSeek-R1-Distill-Qwen-7B (P14, 2025-01) → Qwen3-8B (P13, 2025-04); the raw leader is the newer 8B model, but once verbosity is removed its edge shrinks and the curve tracks release recency; size moves it little along this axis. (b) The *capacity* axis at fixed 2024-09 vintage, doubling parameters from 7B (P9) to 14B (P17): the raw overall score rises +0.025 (95% CI [0.008, 0.043], paired bootstrap p=0.005), a clean capacity effect with the decode backend held constant, and the 14B arm is the length-adjusted best of the four (0.311). The isolated capacity effect is small relative to the frontier gap.

Capability is the cheaper axis at the frontier, read with a coverage caveat. The local-model curve above and the architecture sweep of Section 5.3 are two margins of one surface: capability (7B→8B→14B→GPT-4o→GPT-4.1, all under GPT-5.2) against architecture (P0→P4). The architecture margin is a 17.5× token-spend step (Section 5.6). The full surface is under-identified, since no P4-on-local cell exists to pin down the interaction. The frontier capability step also carries its own coverage gap: the gpt-4o rung is the full 90-query manifest (0.3845), while the gpt-4.1 rung is a pre-selected, deterministic 45-query cost-control subset (the 30 variance-stratified queries used by the oracle contrast plus 15 more), run to completion with no dropped queries. Its composition was fixed in advance, which rules out survivorship bias, but this smaller sample, composed differently, still falls short of a query-matched comparison to the 90-query gpt-4o figure. We read the 0.487 gpt-4.1 mean as indicative – it does not rise to a certified like-for-like level; a query-matched re-scoring of P0 on the same 45-query subset under gpt-4o would be the clean fix. (The reduced-contrast replications of Section 5.10 do lose queries to a cost budget on the gpt-4.1 backend, but only for the P4 arms; this P0 rung is unaffected.) The capability step (P0, gpt-4o→gpt-4.1) reads as +0.103 against the architecture step (P0→P4 at gpt-4o) of +0.082 (95% CI [0.049, 0.116]). A model-generation upgrade is therefore at least comparable to, and on this uncorrected reading slightly larger than, the entire architecture stack, at roughly 1/17.5 the token spend (a cost comparison; compute across model generations remains incommensurable).

A second reading checks whether capability and architecture spend are reaching the same score by two different routes. The interaction slice, P4 on gpt-4o (0.467) against P0 on gpt-4.1 (0.487), sits within 0.020: we had read this as the same *isoquant* (an economics term for a curve of equally good trade-offs) reached by two different paths. Given the coverage gap just noted, that reading falls short of a clean no-extrapolation result and should be treated as suggestive. A related, cleaner check is whether the two frontier-tier architecture premiums agree with each other: they are close in point estimate (0.082 at gpt-4o versus 0.086 at gpt-4.1).

Here it matters which P0 baseline each premium uses. The gpt-4.1 premium is a genuine paired difference over the 32 queries P4 also completed there, measured against its own, narrower P0 baseline there (≈ 0.441), which is a different, smaller-sample number from the 0.487 full-45-query P0 mean used in the capability-step comparison above. The closeness of the two premiums is consistent with additive separability over this range, though without a direct test of that interaction term it falls short of a certified equivalence claim. Whether separability holds at the low-capability end remains untested, and the isoquant reading would need a P4-on-local cell to say.

A third, separate reading asks what exchange rate this implies down the local-model ladder: roughly three parameter-doublings per architecture step, as a point estimate. But the 95% CI (1.2–15 doublings) spans a multiplier range of roughly 2× to 30,000×, several orders of magnitude of spread. The data leave this exchange rate unbounded, so we read it as a curiosity. Our stance elsewhere, that we do *not* attribute the capability gap to parameter count alone (Section 5.4), stands regardless. One further caveat travels from Section 5.8: the P0-to-P4 premium used throughout this comparison is an upper bound on the pure-architecture effect, since the matched-budget probe attributes a borderline ≈40% of a same-size premium to compute. If a similar share applies here, the frontier exchange rate above shrinks by a comparable fraction, which only strengthens the qualitative reading that capability is the cheaper of the two axes, independent of the coverage caveat above.

Orchestration also fails to close the gap from below. We measure that directly. Running the two strongest GPT-4o architectures on the 7B backbone (P1 and P4 wiring; same tools and scaffolding; GPT-5.2 single-judge on a 12-query subset) yields premiums over the single-pass 7B baseline of +0.014 (P1 architecture, 95% CI $[-0.05, +0.08]$, p=0.68) and +0.067 (P4 architecture, CI $[-0.01, +0.14]$, p=0.23). Neither is significant at this sample size. The premiums are of the same small order as the (equally non-significant) premiums those architectures earn at GPT-4o scale on the same subset (+0.053 and +0.005). Even at the CI upper bounds they close a fraction of the 0.38 gap (Table 15). The probe is too small to rank backbones or architectures. What it bounds is the headroom: orchestration shows no sign of substituting for capability, with closure beyond +0.14 excluded at 95%. What is robust about the capability gap itself is its *direction*, and its survival under every judge individually. This includes the two cross-family Anthropic judges that cannot favour GPT-4o on family grounds (Opus: P0 0.47 vs. P9 0.24; Sonnet: P0 0.61 vs. P9 0.35). Because GPT-5.2 is GPT-family-adjacent and may inflate the GPT-4o side, we take the gap's *magnitude* from the full panel but its *existence* only from the cross-family judges. Those two Anthropic judges do agree with each other more than either does with GPT-5.2 (cell-level $r = 0.82$ vs. ≈ 0.73), making them the most redundant pair on the panel. We quantify this with an effective-judge count: judges who mostly agree supply less independent information than judges who disagree, so raw judge count overstates a panel's independent evidence.[3] The two same-lab Anthropic judges carry barely more than a single independent vote between them, and this is a property of that specific pair, not of panel size in general: a second, independent check (a spectral participation ratio, which counts effective dimensions from the eigenvalues of the judge-correlation matrix rather than from pairwise correlations directly) agrees on the Anthropic pair but does not corroborate it for a different pair.

Judge subset	N_{eff} (verdict formula)	N_{eff} (spectral check)
Two same-lab Anthropic judges	1.32	≈ 1.4
Full three-judge panel	1.65 (95% CI [1.64, 1.67])	—
Two OpenAI-lineage scorers[4]	1.41	—

[3]Computed over paired criterion verdicts on the fully crossed cells; the formula is $N_{\text{eff}} = N^2/(N + 2\sum_{i<j}\phi_{ij})$, with ϕ the pairwise verdict correlation.

[4]GPT-5.2 and a GPT-4.1 judge that re-scored the full corpus as a robustness scorer and enters no headline analysis; a separate pair from the Anthropic one above, so this does not corroborate that figure – it establishes that panel redundancy is a same-lab property, not an artefact of any one judge pair.

The full-panel CI is tight given the panel's 36,113 crossed cells. The redundancy is not concentrated in one rubric dimension. Recomputed per dimension (Table 2), the within-minus-cross-family excess is Holm-significant across all nine ($p \leq 0.005$ in every case), tightest on `organization` ($N_{\text{eff}} = 1.31$) and loosest on `attribution_quality` ($N_{\text{eff}} = 2.70$). Panel redundancy is itself established prior art: Kohli [41] find a nine-judge panel carries only about two effective votes. Our measurement is an independent, stricter instance of the same phenomenon on a cross-family three-judge panel; we claim no co-discovery. A genuinely independent fourth family (e.g., Gemini) is the clean fix (Section 12).

Table 2: Effective-judge count per rubric dimension. $\phi_{\text{Opus,Son}}$ is the same-lab (within-family) pairwise verdict correlation, $\phi_{\text{GPT,Claude}}$ the cross-family average; $\Delta_{\text{w-c}}$ is the within-minus-cross excess, Holm-corrected across the nine dimensions. All nine clear the correction.

Dimension	$\phi_{\text{Opus,Son}}$	$\phi_{\text{GPT,Claude}}$	$\Delta_{\text{w-c}}$	N_{eff}
Instruction	0.489	0.217	+0.272	1.86
Coherence	0.495	0.234	+0.261	1.83
Citation	0.404	0.201	+0.203	1.95
Attribution	0.180	-0.006	+0.186	2.70
Info. recall	0.499	0.325	+0.174	1.70
Factual	0.492	0.352	+0.140	1.67
Coverage	0.515	0.396	+0.119	1.60
Org.	0.680	0.633	+0.047	1.31
Depth	0.5016	0.4563	+0.045	1.54
Overall	0.512	0.354	+0.158	1.65

$N_{\text{eff}}/k = 0.552$ (caution at < 0.5, [41]); panel sits just above the line.

5.5 RL post-training helps the 7B agent, though the gain does not reach the frontier

Holding the 7B backbone and tools fixed, the RL-trained DeepResearcher (P10, 0.34) beats the base Qwen2.5-7B (P9, 0.26) by $\Delta = 0.078$ with posterior 1.00. The contrast does not isolate RL cleanly: P10 is RL-trained, and it also runs a learned multi-search agentic loop (up to 8 iterations), where P9 is single-pass, so the 0.078 bundles RL post-training with that added scaffold. Either way the effect is real but small, in the range the RL-for-search literature reports for comparable backbones. It is about one-fifth of the frontier–7B gap (0.078 vs. 0.38) and nowhere near the GPT-4o cluster. The effect also holds out of distribution: on five external benchmarks (DeepSearchQA, DRACO, FreshWiki, LitQA2, ResearchQA; 75 paired queries, GPT-5.2), P10 beats P9 by +0.13 (95% CI $[0.092, 0.170]$). P9's mean is itself partly an aggregation artefact of its degenerate DeepSearchQA floor (Section 9).

5.6 Cost: a flat cluster is an economic choice, not an architectural one

Because the top cluster is statistically flat (no judge-robust separation at the power we have), the decision it leaves is settled on price. Figure 6 plots overall score against per-query compute cost: a quality-versus-budget trade-off in which several cluster members are dominated. P4 (STORM), the cluster's most prominent member, is also its *most expensive* ($5.66/query), yet it scores no higher than the cheaper P7 ($2.42), P6 ($2.65), or P1 ($3.91). It therefore sits off the efficiency frontier, as do P5 and P8. Among the GPT-4o patterns the Pareto frontier runs from the single-pass P0 ($0.324/query, 1.51 points per dollar) through P2 ($0.91, 0.64) to the cheaper cluster members P7 and P6, and finally P1, the highest-scoring pattern. (The local-7B agents sit below every GPT-4o pattern at negligible marginal cost. Their GPU-amortised cost is not commensurable with API spend, so we draw the frontier over the GPT-4o patterns only.) The reading is therefore conditional on budget. If cost-efficiency is the objective, P0 and P2 dominate on quality per dollar.

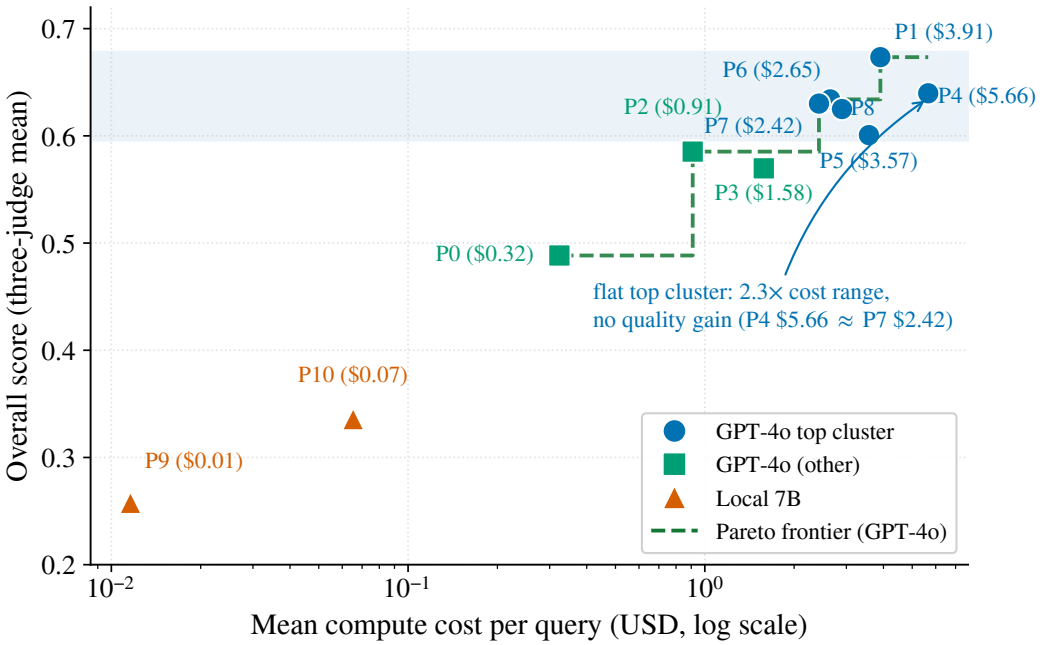

Figure 6: Overall score vs. per-query compute cost (log scale; three-judge mean, N=90 queries, 87–89 for three patterns). The dashed line is the Pareto-efficient frontier over the GPT-4o patterns (the 7B agents' GPU-amortised cost is not commensurable with API spend); the shaded band is the top cluster (five judge-robust patterns plus P5, which joins only under uncorrected equivalence testing). The expensive cluster members P4, P5, and P8 are dominated and sit off the frontier, which runs P0 → P2 → P7 → P6 → P1. P0 ($0.32) and P2 ($0.91) give the best quality-per-dollar, while P7 and P6 are the cheapest patterns that reach cluster-level quality. Scores are raw-scale point estimates; within-cluster orderings sit inside run noise.

If cluster-level quality is required, P7 or P6 are the cheapest patterns that reach it, and paying up to P4's price adds nothing P1 does not already give for less. The one architectural lift the raw scale prices as worth paying for is the step from P0 into competence on hard queries (Section 5.3). Even that step inherits the length-robust bound of Section 5.9, and past it, more orchestration compute returns little. (Costs are a token-priced proxy; local-7B cost is GPU-amortised and not directly comparable to API spend, and any within-cluster ordering is inside per-run noise $\sigma_{\text{run}} = 0.05$.)

5.7 Best-of-N over a single pass matches the cluster only at matched spend

This subsection and the two probes that follow are three robustness checks on the same quantity: the baseline-to-cluster gain. Each stress-tests it against a different alternative explanation, in order: resampling, then raw compute, then report length. The step from P0 into the cluster is the one architectural gain we keep pointing to. Is it a synthesis gain, or just an effect of running a stochastic pipeline more than once and keeping the better output? The variance experiment gives us twelve independent P0 generations per query, so we can test this. But the naive version of the test is biased, and we report it only to dismiss it. Selecting the best of k samples *by the same judge score that then evaluates the selection* inflates the curve through selection on judge noise (a winner's curse). The naive curve reaches 0.468 at four samples against the orchestrated cluster's 0.470. But a pure-noise calculation – selecting the maximum of k draws from a flat process with the measured within-query replicate SD (σ=0.050) – predicts 0.471 at k=4 and 0.478 at k=5. The naive crossing sits at the noise floor and is uninterpretable.

The unbiased version decouples selection from scoring. We split each report's criterion verdicts into two disjoint halves within every dimension, select the best sample by the first half, and score the selected sample on the held-out second half; the cluster reference is recomputed on the same held-out basis, so the scales

match. The split removes criterion-level selection noise. But judge error that is systematic at the *report* level – severity, length and style response – is shared across the two halves. That shared error is selected on and rewarded, so the decoupled curve remains upward-biased for best-of-N. Empirically the shared component is substantial: within-query deviations correlate at r=0.43 across the two halves, so decoupling removes only the criterion-noise share.

The decoupled curve tells a cleaner story. Best-of-k stays *below* the cluster through k=6 (0.448 at k=4, 0.455 at k=5, against the cluster's 0.468). It draws level at roughly seven samples (0.467) and stays at that level through k=12. The crossing point is split-direction-dependent. The reversed split (selecting on the held-out half, scoring on the first) plateaus $\approx$0.02 below its own cluster reference and never crosses through k=12. The direction-averaged curve approaches the cluster within noise (0.013 below at k=7, CI spanning zero) without a clean crossing. The paired decoupled-minus-cluster gap carries a query-bootstrap 95% CI that spans zero at every k, so "level at $k\approx$7–12" is a point-estimate reading. Given the shared-error bias above, it is also a lower bound on the true k: the 30-query subset cannot resolve gaps below $\approx$0.04. What the data certify is that the decoupled curve never significantly exceeds the cluster, and that the naive k=4–5 crossing disappears once selection noise is removed. Seven to twelve P0 samples cost \$2.3–\$3.9 per query against the cluster's mean \$3.5: the two routes land at matched spend, and resampling saves nothing on the bill. The gain still comes from selection; averaging alone would not produce it – the mean of the samples stays at P0's own level, 0.42 – and oracle selection on held-out criteria remains an upper bound on any deployable selector.

The control cuts both ways. Orchestration does *not* significantly beat the best-of-N envelope of its own baseline, consistent with concurrent evidence that structured coordination collapses to simple sampling [16]. But neither does resampling beat orchestration. At realistic selector quality – the oracle bound is unattainable, and the decoupled bound itself remains selector-friendly – the cluster is the cheaper route to its own level. The control is GPT-5.2 single-judge and runs on the variance subset, where the single-judge P0-to-cluster gap (0.04) is smaller than the same single-judge gap on the full 90-query manifest (0.066). Re-scoring the replicates with the cross-family panel is the clean confirmation, left to future work (Section 12). The full naive, pure-noise, and decoupled curves are in Table 16.

5.8 Matched-budget probe: part of the orchestration premium is compute

The tool *interface* is held fixed across all patterns, but tool *usage* is not: the token and tool-call budget rises 17.5× from the single pass to the heaviest pipeline (Section 5.6 gives the per-query dollar costs). So the P0-to-cluster gap confounds architecture with compute, and a sceptic can read "bounded returns to orchestration" as "bounded returns to spending more on the same scaffold." A disentanglement arm re-runs the P1 architecture with its token/tool-call budget clamped down towards the baseline's, holding the architecture fixed: from 12.1× P0 to 3.2× P0 in dollar terms. The clamp closes most but not all of the spend gap; "matched budget" below is shorthand for this clamped condition. The arm is scored by GPT-5.2 on the 29-query subset it covers.

The probe certifies one claim: at matched budget P1's advantage over P0 is no longer detectable. Pairing by query, the full-budget P1 beats P0 by +0.068 on the overall score (Wilcoxon p=0.013, significant). At matched budget the gap falls to +0.041 and is no longer distinguishable from zero (p=0.78; bootstrap 95% CI $[-0.016, +0.103]$). That interval still contains the unmatched effect, so the matched null is a power-limited statement, not evidence that the advantage is absent. Comparing a significant result to a non-significant one is not itself a valid test, so we also test the clamp directly: the effect of clamping (full-budget P1 vs. clamped P1, paired by query) is −0.027, short of significance under both tests we ran (Wilcoxon p=0.053; sign-flip permutation p=0.21; bootstrap 95% CI $[-0.064, +0.016]$). Three limits bound the probe. The direct clamp effect is uncertified. The compute-share point estimate ($\approx$40%) carries an uninterpretable bootstrap 95% CI of $[-29\%, 149\%]$; it collapses two noisy paired gaps into a single figure, and we do not certify the resulting split. And the coverage is one pattern (P1), one judge (GPT-5.2), and 29 queries.

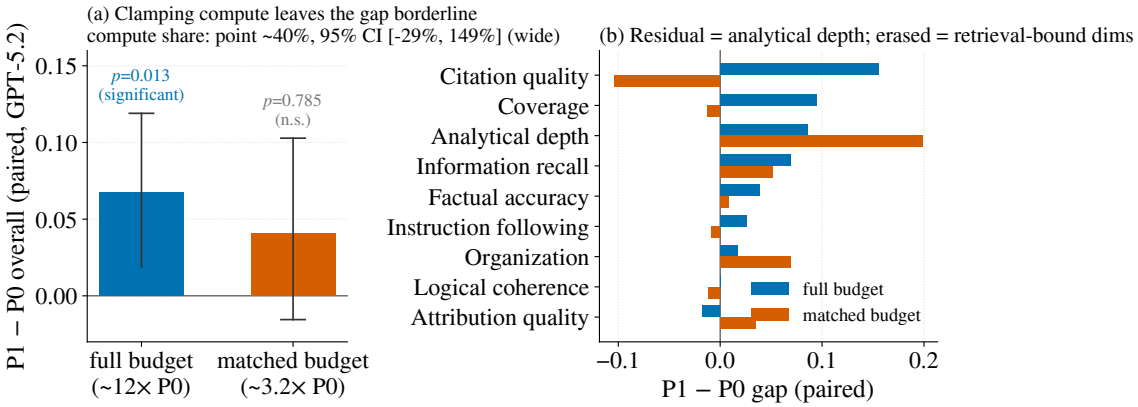

Figure 7: The matched-budget probe erases the retrieval-bound share of P1's advantage. Paired P1-vs-P0 gaps under full budget and under the 3.2× clamp (GPT-5.2, 29 queries). The overall gap falls from +0.068 (p=0.013) to +0.041 (p=0.78); per dimension, the clamp erases the citation-quality and coverage components while analytical depth strengthens. The compute share is a point estimate ($\approx 40\%$) with an uninterpretable 95% CI ([−29%, 149%]) and is not certified (Section 5.8).

The per-dimension breakdown supports the mechanism (Figure 7). The component the clamp erases is exactly the retrieval-bound pair that more search drives, and the attenuation is itself significant for both: `citation_quality` falls from +0.155 to −0.103 (change −0.259, 95% CI [−0.388, −0.121]), and `coverage` falls from +0.095 to −0.013 (change −0.108, [−0.196, −0.032]). `analytical_depth` moves the other way: it *strengthens* under the clamp (+0.086 → +0.198), and it is the one dimension whose clamped gap excludes zero ([0.078, 0.319], Wilcoxon p=0.004). That pattern is consistent with a synthesis property of the orchestration, and it fits a compute artefact poorly. These per-dimension intervals carry no multiplicity correction across the nine dimensions and share the same 29 queries. (The parallel P4 arm covers only 9 of the 30 variance queries and shows no base P4-over-P0 gap to decompose, so we do not lean on it.) The architecture effect does not invert, but it is smaller than the raw cluster gap implies. We accordingly frame the comparison as returns to architecture *and* its compute footprint jointly. It converges with the best-of-N result above: both controls bound the pure-architecture share of the orchestration premium from above, and both find that compute alone accounts for a meaningful slice of the premium.

An enterprise-scale harness study reaches a convergent structural reading from a cost angle. Across six foundation models and two orchestration layers, a model's harness-driven quality gain correlates almost perfectly with its own baseline capability ($r = 0.99$, n=6) [81]. That fits a picture in which orchestration multiplies whatever capability the underlying model already has.

5.9 Length-adjusted contrasts: much of the raw premium rides on report length

The judge audit finds positive length coefficients ($\beta(\log \text{wc}) \approx 0.04$–$0.11$, largest for the two Claude judges). The cluster writes systematically longer reports ($\approx$ 1.9k words versus P0's $\approx$ 1.2k and P9's $\approx$ 0.7k). And Section 5.4 shows that length adjustment reorders the local-model curve. We apply the identical pooled-OLS specification: score on pattern dummies plus output length, each pattern evaluated at grand-mean length.

The adjustment is decisive for the orchestration gap and not for the capability gaps. At grand-mean length the three-judge cluster-vs-P0 gap of 0.145 falls to −0.003 (query-clustered 95% CI [−0.038, 0.032]). With query fixed effects it is +0.025 ([−0.009, 0.058]). Under the milder log-words specification it is +0.049 ([0.022, 0.076]), smaller but still positive. The capability gaps survive every specification: P0-vs-P9 falls from 0.231 to 0.113 ([0.077, 0.151]), and cluster-vs-P9 from 0.376 to 0.110 ([0.050, 0.168]), both excluding zero.

Two readings bound the interpretation. The pooled within-pattern length slope partly reflects retrieval

success – a pipeline that finds more writes more – so judge verbosity alone cannot explain it fully; the full-strength adjustment over-corrects and the log-words one under-corrects. The length-robust orchestration gain is therefore bounded between roughly zero and 0.05, against 0.145 raw. That tightens the bounded-returns headline and sharpens the judge finding of Section 6: under the full-strength specification the family-clean GPT-5.2 judge alone puts the adjusted cluster gap at −0.029 ([−0.061, 0.000]). So on the one judge with no citation-density bonus, the length-adjusted orchestration premium is indistinguishable from zero. Verbosity fails to explain the capability gap under any specification. Grand-mean length (≈ 1,600 words) sits near P9's 95th percentile, so the adjusted P9 contrasts involve mild extrapolation. Wherever later sections quote premiums without a length qualifier – the per-source stratification, the cost frontier, the external battery – those are raw-scale quantities, and they inherit the 0–0.05 length-robust bound established here. One quantity, the baseline-to-cluster premium, recurs throughout the paper under enough different judge bases and adjustments that we gather every version of it here rather than leave it scattered:

Basis	Premium	95% CI	Note
Three-judge panel, raw (headline)	+0.145	[0.110, 0.183]	Section 5.3
GPT-5.2 alone, family-clean, raw	+0.066	[0.038, 0.094]	no citation-density bonus
Three-judge, length-adjusted (full-strength)	−0.003	[−0.038, 0.032]	spans zero
Three-judge, length-adjusted (query fixed effects)	+0.025	[−0.009, 0.058]	spans zero
Three-judge, length-adjusted (log-words)	+0.049	[0.022, 0.076]	excludes zero
GPT-5.2 alone, length-adjusted (full-strength)	−0.029	[−0.061, 0.000]	spans zero
Single-judge, live retrieval (oracle-arm subset)	+0.044	–	Section 7
Single-judge, ideal sources (same subset)	+0.010	–	Section 7
Per-source range, five benchmark families	+0.03 to +0.23	–	Figure 4

Every row above is the same premium; none is a slice of another, and none supersedes the raw headline. What varies row to row is the judge, the length adjustment, and the evidence condition (live retrieval vs. oracle sources).

5.10 External and cross-backbone replication

Both headline findings survive when we vary the base model and the query distribution. We regenerate the reduced contrast (the single-call baseline P0, the strongest pipeline P4, and its oracle arm) on a second frontier backbone (GPT-4.1, same vendor lineage as GPT-4o). The orchestration gain over the single-call baseline is bounded but positive (+0.086, 95% CI [0.027, 0.148], n=32 queries). The oracle lift over live retrieval is significant (+0.079, 95% CI [0.037, 0.119], n=17). The one backbone-dependent detail is that the two effects are comparable in magnitude here (0.079 versus 0.086). The oracle lift therefore does not *dominate* the orchestration gain as it does on GPT-4o: the retrieval-and-synthesis ceiling reproduces, but the size of its margin over orchestration is not itself backbone-invariant. This is a reduced-contrast replication on 100 of 120 target reports. The 20 heaviest queries exceeded the per-query generation budget on the costlier backbone and are dropped symmetrically from each paired contrast, and the replication does not re-run the full inner-cluster tie. Symmetric dropping protects the paired contrast itself but not external validity: we have not tested whether query weight correlates with difficulty, so if the dropped queries are systematically easier or harder than the retained 100, this replication samples a slice of the original manifest that is not fully representative. Unlike the P4-only probes above, a current-Claude crosscheck covers the broader six-pattern cluster on a separate, larger gpt-4.1 generation batch (n=58 for P0, 21–40 per cluster pattern), not the 100-of-120-report reduced contrast just described; the GPT-5.2 anchor below is recomputed on this same larger batch, so the two are directly comparable to each other but not to the +0.086/+0.079 figures above. The cluster-over-P0 gap excludes zero under all three judges, so the second-backbone replication is not a GPT-family-adjacent artefact of the judge that also generated it:

Judge	Cluster-over-P0 gap (95% CI)
GPT-5.2	+0.056 ([0.020, 0.093])
Opus	+0.051 ([0.016, 0.089])
Sonnet	+0.078 ([0.027, 0.132])

The six-pattern *ranking* does not travel with the gap: per-pattern ordering across the two backbones is weakly correlated (Spearman $\rho = 0.11$), and P4, the strongest pipeline on gpt-4o, is not the strongest on gpt-4.1. This strengthens finding (i): the aggregate cluster-over-baseline premium is backbone-robust, but *which* architecture leads inside the cluster is not. That split is consistent with capability dominating architecture as the source of any single pipeline's momentary lead.

The query distribution itself replicates on a fresh external-validation battery held wholly out of the main run (Table 12). The battery is a 600-report set, five query families $\times$ {P0, P1, P4} $\times$ 40, with means recomputed directly from the 600 verdicts. The headline reproduces out of sample, and by a wide margin. The two orchestration-cluster pipelines both clear the single-pass baseline: P1 at 0.486 and P4 at 0.453 (n=200 reports each), against P0 at 0.183. The cluster-over-baseline gap travels intact; this is on the raw scale, since no length adjustment is applied to the battery. Absolute levels do not travel: P0's mean falls from 0.49 to 0.18 across manifests. Every absolute quantity in this paper is therefore indexed to the released manifest, and what generalises is the ordinal structure. The battery also sharpens the "single-pass baseline is weak" point: on this harder slice the GPT-4o single pass (P0) falls to $\approx$0.18 and sits at or below the local-7B agents on GPT-5.2's own scale. That places it level with P9 at 0.18 and below P10 at 0.21. Both 7B figures are GPT-5.2-single-judge scores on the main manifest; elsewhere in this paper P9 and P10 carry the panel means 0.26/0.34. On this slice, then, an unorchestrated frontier pass delivers little more than a 7B backbone. The battery's own caveats (partial replication with the RACE layer BLOCKED, and a flat-cluster check that formally returns FALSE, driven entirely by the within-cluster P1 $\leftrightarrow$ P4 swap) are set out in Table 12.

Both replications score with the independent GPT-5.2 judge alone, short of the full panel, so we do *not* claim strict judge independence for either: GPT-5.2 is GPT-family-adjacent and shares OpenAI lineage with the GPT-4.1 generator. These are replications *within the same judge lineage* of the two headline findings, not cross-family ones; the cross-family panel of the main study is where the independence claim lives. We read them as *internal* replications that remove the single-backbone and single-distribution caveats for the two claims the paper leans on, not as statements about any one model's standing.

5.11 An external framework bake-off: bounded returns replicates while leaderboards shift

The replications above still run our own eleven architectures. This bake-off asks whether the same bounded-returns pattern holds on external codebases, and it does, though the judges disagree sharply on the fine-grained ranking. It runs six independent, publicly released deep-research codebases (OPEN DEEP RESEARCH, GPT RESEARCHER, II-RESEARCHER, STORM, OWL, and DEERFLOW) on 30 frozen queries and the identical rubric, scored by the full three-judge panel. OPEN DEEP RESEARCH is additionally run on a Tavily-backed search backend as a seventh arm, so the seven-arm leaderboard below is best read as six independent systems, one of them run twice under two backends. Each framework runs its own default or documented configuration; we did not equalise a budget across them: iteration and tool-call caps range from two (STORM's conversation turns) to eight (II-RESEARCHER), and DEERFLOW runs in its shallowest single-agent mode. GPT RESEARCHER's planning and writing steps (`smart_llm`/`strategic_llm`) run on GPT-4.1, while every other framework, and GPT RESEARCHER's own fast-path calls, run GPT-4O-MINI. "Fixed backbone" therefore describes six of the seven arms exactly and the seventh only partially. This is an as-deployed comparison of external systems, not a controlled ablation; differences among the lower-ranked frameworks in particular do not isolate architecture from budget or backbone. Completion also differs (DEERFLOW 27/30, OWL 29/30, the rest 30/30), and, as elsewhere in this paper, means below are computed over completed queries only. The seven-arm leaderboard, all three judges:

Framework (backend)	GPT-5.2 (95% CI)	Opus	Sonnet	Completion
Open Deep Research (Azure)	0.435 ([0.392, 0.480])	0.716	0.594	30/30
Open Deep Research (Tavily)	0.433 ([0.391, 0.476])	0.841	0.734	30/30
ii-researcher	0.423 ([0.390, 0.457])	0.726	0.657	30/30
GPT Researcher	0.417 ([0.363, 0.469])	0.754	0.639	30/30
OWL	0.352 ([0.302, 0.404])	0.726	0.608	29/30
DeerFlow	0.334 ([0.279, 0.393])	0.753	0.594	27/30
STORM	0.330 ([0.274, 0.388])	0.747	0.610	30/30

On the GPT-5.2 anchor the top four arms land within 0.018 of one another (the table above gives each arm and judge). This is the same bounded-returns pattern as Section 5.3, reproduced on codebases this paper did not write. With only six independent systems, though, this is a smaller-n replication than the eleven-architecture main study. The panel is not unanimous about *which* four, or in what order, though at seven arms none of the pairwise rank agreements reaches significance (Spearman's two-sided critical value at $n = 7$ is $|\rho| \approx 0.79$); the next table gives the pairwise rank agreements and each judge's absolute leniency. The two Claude judges agree with each other more than either agrees with GPT-5.2, and this pattern points the same way as the within-family vs. cross-family split behind the panel's 1.65 effective judges (Section 5.4), though it supplies no independent statistical confirmation of that split.

Metric	Value	95% CI
Rank agreement (Spearman ρ): GPT-5.2 vs. Opus	−0.214	–
Rank agreement (Spearman ρ): GPT-5.2 vs. Sonnet	+0.143	–
Rank agreement (Spearman ρ): Opus vs. Sonnet	+0.536	–
Leniency (mean overall score, 0–1 scale): GPT-5.2	0.389	–
Leniency (mean overall score, 0–1 scale): Opus	0.752	–
Leniency (mean overall score, 0–1 scale): Sonnet	0.634	–
Backend Δ overall (Tavily − Azure): GPT-5.2	−0.002	[−0.037, 0.034]
Backend Δ overall (Tavily − Azure): Opus	+0.125	[0.073, 0.177]
Backend Δ overall (Tavily − Azure): Sonnet	+0.140	[0.074, 0.206]
Backend Δ domain-HHI (0–1 scale, Tavily − Azure; judge-independent, a property of the report text)	−0.168	[−0.253, −0.097]

The bake-off also isolates a backend effect the main study cannot, since every one of its own patterns uses one fixed search backend. Swapping Open Deep Research's default backend for Tavily moves overall score under GPT-5.2 by an amount spanning zero, but under Opus and Sonnet by amounts that exclude zero (n=30 paired queries; table above). The swap also lowers citation-domain concentration, measured by the Herfindahl-Hirschman Index (HHI): a lower HHI means citations are spread across more distinct domains, and the Tavily arm cites a more diverse set (table above, excluding zero). The two Claude judges reward that diversity on this one within-codebase, paired contrast; GPT-5.2 does not. This is an external, same-mechanism replication of the citation-density-bonus finding of Section 6. The bake-off's message is double-edged. Architecture choice is bounded across independent codebases as it is across our own, on a smaller and less controlled sample than the main study. Judge choice cannot be averaged away, since it changes both the leaderboard's order and its response to a backend swap.

5.12 Apparatus by-products: a distilled judge and two post-hoc probes

Two by-products of the apparatus qualify how far the panel and its cheaper approximations can be trusted. We fine-tuned a 7B judge (DR-Judge) on 30,465 panel-consensus verdicts to test whether the panel can be cheaply approximated. It *missed* its pre-specified $\kappa \geq 0.7$ target. On the held-out test split (Table 10) it reaches $\kappa = 0.45$ overall: 0.65 on verdicts where the panel was unanimous, but only 0.20 where the panel

split. The disputed cases are exactly the information-rich ones the panel needed three judges to resolve, and a small distilled model cannot reproduce that tie-breaking. We report this as a methods finding on the limits of small-judge distillation, and scope DR-Judge's intended release to a triage tool for unanimous verdicts only (staged for the archive, see "Reproducibility and Artefact Availability"). It is barred from scoring P9/P10 under the family-of-judges constraint (Section 4).

Two probes added after the main run are reported on a GPT-5.2 single-judge basis (Table 14). A partial Sonnet re-score in the underlying data (80 and 52 cells) agrees in direction and is excluded from the analysis for comparability. A verbatim ReAct controller (P11, *post-hoc, GPT-5.2 single-judge*) scores 0.33, below even the single-pass P0 (0.385 on the same single-judge basis) and roughly equidistant between the 7B tier and the orchestration cluster. Its 8-turn budget is largely consumed by sequential search/read actions, and the model rarely terminates voluntarily. The deficit does not shrink when the turn budget doubles (0.361 vs. 0.358 at 16 turns on the 29 paired queries, Wilcoxon p=0.71, itself a power-limited null). We read this narrowly: primitive ReAct underperforms orchestration on the GPT-5.2 axis, and we pass no general verdict on the ReAct family. Our own GRPO-trained 7B agent (P12, *post-hoc, GPT-5.2 single-judge*) scores 0.182, statistically tied with its untrained P9 base (0.180; paired Wilcoxon on the same 90 queries, p=0.84). We had attributed this negative result to using the noisy DR-Judge as the reward model, though two further, undisclosed factors plausibly contribute and are not separable with this run. First, the training recipe is a compute-driven simplification of textbook GRPO: a group size of 2, well below the 8–64 typical in the literature, and $\beta = 0$, dropping the KL term to the reference entirely. That term normally penalises the model for drifting too far from its pre-RL starting point; dropping it removes that guardrail. Both were deliberate cost choices for a single 7B run, and we ran neither as an ablation. Second, DR-Judge itself is queried at reward time in a format it was never fine-tuned on: a holistic $[0, 1]$ quality score across four rubric facets, whereas its 30,465-verdict training set and measured κ are both on a single-criterion binary satisfied/not-satisfied task. So the measured κ does not directly characterise the reliability of the signal P12 actually trained against. We report this as a bounded limit on this particular RL recipe supervised by a rubric judge, not a working system, and we do not claim to have isolated reward noise as the sole cause.

6 Citation Quality Is Largely a Measurement Artefact, Not a Research Difference

The main comparison and its replications leave the cluster indistinguishable on overall score. Within the top cluster, the dimension that varies most is `citation_quality` (Table 4): P1 and P7 earn ≈ 0.78, while P4 sits at 0.49 and P5 at 0.48. It is tempting to read this as a real difference in how well these pipelines source their claims. It is not. We show, by audit and regression, that the dimension largely counts how *many* citation-shaped strings a report contains and registers little about whether they are grounded. This behaviour originates in the rubric the judges apply, and it is repairable.

Provenance audit. We classified all 22,559 citations in the base reports as grounded (a resolvable URL or academic identifier) or placeholder ("Web Search Synthesis"-style strings emitted by the retrieval backend; Table 7). The GPT-4o pipelines split sharply: P4, P3, P8, and P5 emit 51–62% placeholder citations, whereas P1, P6, and P7 emit at most 0.33%. Yet the placeholder-heavy pipelines are *not* penalised in proportion: P4 emits 62% placeholders and still earns `citation_quality` 0.49. At the report level, `citation_quality` correlates about as strongly with raw citation count ($r = 0.49$) as with source provenance ($r = 0.47$).

The placeholder class is the "Web Search Synthesis" entries emitted by the Azure Responses `web_search_preview` tool, which wraps Bing-grounded synthesis. These are *not* necessarily fabricated, but they carry no resolvable URL that our extractor can verify. The finding therefore concerns the rubric: a citation-*presence* rubric credits non-resolvable citation-shaped text much like resolvable URLs. The credit falls at the boundary between the rubric and our classifier. In this sense the citation-quality gap belongs to measurement; it tells us little about substance.

Density earns a provenance-independent bonus. A report-level regression of `factual_accuracy` on citation features with pattern fixed effects isolates the effect (n=916; R^2_{adj}=0.43, the share of the variation in the score this regression explains). Each β coefficient below is that feature's own estimated effect on the score, holding the others fixed. Because the reports are clustered (every pattern answers the same queries – the statistical sense of "cluster," unrelated to the architecture cluster above), we report cluster-robust p-values in place of the anticonservative i.i.d. values. We rest the claim on the query grouping (G=90), which has enough clusters for the asymptotics to hold. The pattern grouping (G=11) has too few clusters for asymptotic cluster SEs. For that grouping we report a wild-cluster restricted bootstrap, a resampling method built specifically for cases with very few independent groups, where the usual formulas become unreliable; it is anticonservative-corrected and turns both coefficients non-significant. Source provenance is a genuine positive signal ($\beta_{\text{provenance}} = +0.108$, query-clustered $p = 3\times10^{-8}$; pattern wild-cluster $p = 0.071$, n.s.). Grounded citations do raise the substance score. Citation *density*, too, carries an *additional*, independent positive coefficient ($\beta_{\log\text{count}} = +0.040$, query-clustered $p = 6\times10^{-6}$; pattern wild-cluster $p = 0.066$, n.s.) after provenance, length, and pattern are controlled. The pattern-clustered contrast, with only eleven clusters, settles nothing either way. Adding query fixed effects, so that density is identified only from patterns answering the same query, shrinks the density coefficient to +0.022 but leaves it significant under the admissible query-clustered test (p=0.008); the pattern-grouping wild-cluster test, with its eleven clusters, is indeterminate at p=0.281. Provenance under the same specification is +0.103 (query-clustered p=1×10^{-8}, wild-cluster p=0.033), now clearing the pattern-grouping test even though the pooled provenance coefficient above does not. Adding more citation-shaped text raises the factual-accuracy score whether or not the citations are grounded. This is the confound: the judges reward the *presence* of brackets as an addition to the grounding they also reward, never as a replacement for it. (The dependent variable, `factual_accuracy`, has weak per-judge agreement, $\alpha \approx 0.13$. We therefore read these coefficients as panel-consensus signal. A per-judge refit locates both effects under the same inference rules as the pooled fit: query-clustered p-values as primary, and an exact wild-cluster bootstrap [7] at the pattern grouping. The density bonus is carried by the two Claude judges and is *absent* under GPT-5.2:

- Opus: $\beta_{\log\text{count}} = +0.072$ (query-clustered $p = 2\times10^{-7}$, wild-cluster $p = 0.008$)
- Sonnet: +0.049 ($p = 5\times10^{-6}$; wild-cluster $p = 0.093$, n.s.)
- GPT-5.2: $\beta = -0.001$, query-clustered $p = 0.92$, n.s.

The provenance reward shows the same Claude-sided pattern:

- Opus +0.140, Sonnet +0.161 (query-clustered $p < 10^{-6}$ for both)
- GPT-5.2 +0.024 (query-clustered $p = 0.32$, n.s.)

At the coarse pattern grouping, though, neither Claude judge's provenance coefficient individually clears the wild-cluster test (Sonnet narrowly, p=0.053; Opus more clearly, p=0.145), even though the pooled query-fixed-effects provenance coefficient above does clear it. GPT-5.2's factual-accuracy score responds to neither citation feature. The Claude judges respond to both. The density bonus is therefore a family-level property of the Claude judges, too broad to write off as a single-judge artefact. With only one non-Claude judge, though, the corroboration within the panel comes from one independent family only, and the pooled +0.040 is a conservative panel average.)

Strict entailment does not rescue the dimension; it reframes it. We ran a FActScore/SAFE-style [66, 106] pipeline that extracts atomic claims and asks whether each report's own cited sources entail them. On a stratified subset, $\approx$ 94% of extracted claims carried *no resolvable citation index* at all. Under a strict "directly stated" entailment rule, the verified-support rate collapsed towards zero across every pattern, including the patterns with concrete URLs throughout. This finding concerns the entailment method itself; it says nothing about which pattern is better. Most factual claims in a synthesised research report are *derived* (combined, contextualised, triangulated across sources) and only seldom lifted verbatim from a single source,

so the verbatim-entailment bar that works for short-form biographies does not fit the deep-research register. Concurrent work on citation evaluation reaches this limitation independently, arguing that natural-language-inference support is a poor proxy for genuine citation quality [112]. A partial-support rubric does recover material evidence flow (supports rate 16% → 40% on a top-cluster subset). That is what a deep-research grounding metric will require.

We measure the fabrication-vs-substance split on our own corpus. We audit fabrication rates through two channels. First, we resolve a stratified sample of cited URLs over HTTP: 96.0% of cited URLs resolve and only 3.5% are dead or fabricated (3.53% among determinate outcomes). This places our corpus at the *low* end of the 3–13% band the deployed-agent audits report [71, 79], so the corpus is not pathologically fabricated. Second, we measure substance by claim-level entailment of the cited subset. Only 22.2% of cited claims are entailed by their own source (95% bootstrap CI [15.9%, 28.4%]), rising to 37.1% among the cites that actually carry a resolvable URL. Both channels force the same reading: URLs mostly resolve, but the substance behind them is thin. This support rate is measured over n=176 cited claims (the sampled claims that carried an inline citation marker), well short of a corpus-wide census.

The orchestration gain does not reduce to differential search success. Because each architecture induces its own evidence set (Section 4), a sceptic can ask whether the cluster-over-baseline advantage is manufactured by the cluster pipelines retrieving successfully more often, with orchestration contributing nothing. On this account, dead links, empty fetches, and rate-limited calls would hurt the single-pass baseline disproportionately. We test this directly by conditioning the cluster-vs-P0 score gap on per-arm search success, here comparing P0 against every pattern that outscores it (P1–P8), a broader grouping than the six-pattern cluster used elsewhere. The gap survives. Regressing the arm-level score on cluster membership while controlling for harmonised retrieval yield (documents per external-fetch attempt) leaves the cluster coefficient at +0.266 (arm-level 95% CI $[0.135, 0.398]$, $t(8)$), statistically unchanged from the naive gap of +0.259. The same regression at the per-query level gives +0.181, but grouping the standard errors by arm (G=11 arm groups, the statistical sense of "cluster," unrelated to the architecture cluster above) gives a mixed result across the three standard corrections for having only a few groups: two of three standard small-sample corrections exclude zero, one does not.

Correction	Result	Excludes zero?
CR1 (small-sample SE adjustment)	CI $[0.034, 0.329]$	Yes
CR2/Bell-McCaffrey (more conservative)	CI $[-0.032, 0.395]$	No
Wild-cluster bootstrap[5]	test-inversion CI $[0.026, 0.326]$	Yes

Since these group-robust constructions do not all agree, we demote the per-query check to descriptive and lean on the arm-level regression, which is unambiguous, as the certified result. The treatment varies only across the eleven arms, so the naive per-row interval overstates precision. The per-query fit therefore corroborates the arm-level result; it does not add independent resolution. Coverage is partial: the per-query fit uses 533 of 990 pattern×query cells with usable trace data. On small-G wild-bootstrap evidence, here and earlier in this section, we apply one standard: at eleven or twelve arm groups the test is power-limited, so its nulls are uninformative while its rejections remain valid tests at the stated level. By that standard, the arm-level result and the per-query check's valid (wild-cluster, CR1) constructions all certify; only the intentionally conservative CR2 correction abstains. True search failure is roughly uniform and, where it varies, runs the *wrong* way to explain the gain. The dead-or-fabricated URL rate spans only 0.0–0.088 across all eleven arms;

[5] [62]; "cluster" here again means an arm group, not an architecture. Two random-weighting schemes agree: Rademacher p=0.011, Webb p=0.012.

the six-pattern cluster's mean dead rate is marginally *higher* than the baseline group's (+0.0085); and resolve rate correlates *negatively* with score across arms ($r = -0.36$). Differential search failure therefore cannot be the engine of the cluster advantage, which is what the search-time-contamination and run-reliability literature makes a live concern [104]. (Per-arm rate-limit and timeout counts are not attributable from the stored traces, so we lean on the dead-URL audit, which is comparable across arms, as the failure signal.)

The confound is rubric-conditioned, and corroborated externally. The effect is a property of the rubric these judges were given. The same cross-family, multi-judge design discriminates grounding well when the rubric *supplies the cited source to the judge*, as in FACTS Grounding [34]. Our collapse occurs precisely because a citation-*presence* rubric never shows the judge what a citation points to. Three concurrent audits of deployed deep-research systems reach the same place from the field side. Deep-research agents emit substantially more citations per query than search-augmented baselines yet report dead-or-fabricated URLs in the 3–13% band [79]. An audit of cited claims finds only 39–77% factual accuracy even as link validity stays above 94% [71]. An eight-dimension audit of commercial *deep-research-mode* agents (as distinct from these products' plain web-search modes) finds unsupported-statement rates from 53.6% (Gemini) to 97.5% (Perplexity), with a single calibrated exception at 12.5% (GPT-5's deep-research mode) [68]. All three are consistent with our own measured 3.5% dead rate and thin 22% support rate above: more citation volume does not yield fewer errors per citation, on our own pipelines or on commercial ones.

A tool-layer intervention. If the score tracked grounding, swapping to a backend that returns cleaner URLs should raise it. It does the opposite. Replacing the Bing-grounded backend with Tavily, which returns far fewer but cleaner-URL results, *depresses* overall scores on all six tested patterns (mean $\Delta = -0.10$; Table 9; *GPT-5.2 single-judge*). Per-dimension, the drop is carried by `citation_quality` itself and by the coverage/organisation dimensions a thinner source pool leaves less to work with. `factual_accuracy` stays flat under GPT-5.2 here too, the same precise null as above: a citation-count effect confined to the citation dimension. The mechanism is citation volume: Tavily returns 16–88% fewer citations per pattern (matched query IDs, mean-of-means basis). Under a density-rewarding rubric, fewer citations means lower scores even when they are better grounded. A 24-report Claude Sonnet cross-check confirms the depression is not judge-specific. Pooled across all six patterns it replicates with the same sign and a larger magnitude ($\Delta = -0.29$, pattern-clustered bootstrap 95% CI $[-0.44, -0.16]$), though at n=4 queries per pattern the per-pattern breakdown is too underpowered to certify individual patterns. The robust conclusion is that *what moves the score is citation count*, and both judges' behaviour supports it.

Some pipelines forage past the point of diminishing evidence, and it gains them nothing. We borrow an idea from foraging theory: Charnov [9]'s marginal-value theorem says an optimal forager should leave a source once it stops returning more than the environment's average yield elsewhere. Applied to the within-run search traces, it asks: does a pipeline keep searching past the point where an optimal forager would already have moved on? The environment's own average marginal yield is 13.3 documents per search (raw returned counts, with within-run duplicates retained), pooled across eight traced pipelines. P1's iterative loop shows the textbook diminishing-returns curve: 8.3 documents on the first search, collapsing to 1.3 on the second and 1.0 by the fourth, an eightfold drop. Yet roughly 93% of its search calls are issued after its own marginal yield has already fallen below the environment average. It keeps foraging in an exhausted patch because the loop runs to a fixed iteration budget, and the stopping rule ignores yield entirely. P10, the RL-trained 7B agent, is the pathological extreme. Every one of its 17.4 search actions per run returns zero documents via the search verb itself (its citations arrive through a separate extraction step), so effort and harvest are completely decoupled: extreme search volume for near-zero direct yield. An arm-level correlation across the eleven arms points the same qualitative direction (raw search count Pearson $r = -0.34$; yield per attempt $r = +0.37$), but

at $n = 11$ neither excludes zero (Fisher-transformed 95% CIs $[-0.78, 0.33]$ and $[-0.30, 0.79]$). The two are not independent evidence of each other, since yield-per-attempt and search count share a denominator, and P10 is a single extreme outlier (highest search count, near-zero yield, low score) that is pulling much of the correlation's visual weight. The within-pipeline evidence above (P1's own eightfold yield collapse and its own statistic that 93% of its searches come late, and P10's zero-yield search verb) does not depend on this cross-arm correlation: the model is not being rewarded for searching past the optimal stop within a run; it is simply not stopping there. This is the same competence-over-effort distinction the rest of the paper draws for synthesis, but we stop short of the general claim that efficiency predicts quality *across* architectures better than volume does. "Search" is also a poor common unit across these architecturally different retrieval wrappers: per-attempt document counts span roughly a twentyfold range across arms in the underlying trace data. The pooled 13.3-document environment average is therefore a rough benchmark, with no standing as a well-defined ecological constant. This analysis covers roughly a third of the manifest's canonical queries per pipeline; traces are sparse for several patterns, and two, P11 and P12, have no usable traces at all. We therefore read it as a mechanism illustration on the two cases discussed; certifying a population estimate would need far more traces.

7 The Oracle-Retrieval Test

The results so far point past the orchestration layer towards retrieval and synthesis, but they do it by inference. Here we test the claim directly. We give every pattern the same sources and ask what changes.

Design. For each query we freeze one identical document set: the pooled union of the grounded sources that the patterns retrieved under live search. We serve that fixed corpus to all nine GPT-4o patterns through the same tool interface. Nothing else moves: same model, same scaffolds, same rubric. We use *oracle* in the fixed-evidence sense of prior deep-research benchmarks: a fixed, pattern-independent evidence pool, here the systems' own pooled retrieved sources. No gold-curated reference set enters the design. Because the corpus is pooled across patterns and not within, a heavy-retrieving pattern gets no more of it than a single-pass one. This is the fairness condition the intervention needs. We regenerate every report on the 30-query variance-stratified subset and re-score with GPT-5.2, then compare each pattern against its own live-retrieval reports on the same queries. To guard against a single-judge artefact we additionally re-score the six cluster patterns' oracle and baseline reports with a current-generation Claude Opus and Sonnet crosscheck,[6] holding each judge's version fixed across the two conditions so the oracle-minus-baseline delta carries no version artefact. The corpus is the pooled-existing tier, so it sets a floor: a gold-curated pool could only do better.

Better sources fix citations but leave facts unchanged. Figure 9(a) is the result. Pooled across the six cluster patterns and paired by query, ideal sources raise `citation_quality` by +0.162. Because the six patterns share the same queries and the same pooled corpus, the pattern is the resampling unit, and the individual report pairs within a pattern are correlated draws. A two-stage block bootstrap over patterns (then queries within) gives a 95% CI of $[0.06, 0.26]$, 2.5× wider than the naive interval that treats the 179 pairs as independent, but it still excludes zero, and it excludes zero under a leave-P5-out check too, $[0.04, 0.27]$.

Factual accuracy does not move: `factual_accuracy` stays at −0.007, and every check below agrees it is indistinguishable from zero at this resolution. A two one-sided test on the six per-pattern deltas certifies it statistically equivalent to zero within a ±0.05 margin (TOST p=0.015), though not within the tighter ±0.02 margin (p=0.21). The 80%-power minimum detectable effect at this six-pattern resolution is 0.050, so the margin here sits exactly at the MDE, whereas Section 5.3 sets its standard at twice the MDE.

[6] This uses the current Opus and Sonnet releases, not the banked judges of the main panel (Section 4), so the Claude deltas are estimated on a different judge vintage than the banked panel's. The version difference is uncontrolled but delta-internal: it cannot manufacture a within-judge delta.

A verdict-level Beta-Binomial cross-check points the same way from far more data: it integrates the ≈2,900 underlying criterion verdicts pooled across both conditions (≈1,440 per condition), a far larger base than the n=6 interval. This model treats each individual pass/fail verdict as a coin flip and pools them statistically. It puts $P(|\delta| \leq 0.05) = 0.99$ and $P(|\delta| \leq 0.02) = 0.70$. But this cross-check pools verdicts within report, so it inherits exactly the within-cell pseudo-replication problem that Section 5 explicitly disqualifies Gate 1's own likelihood-ratio p-value for. Applying that same standard here, we report the Beta-Binomial number only for completeness. The claim we actually stand on is the TOST on the six per-pattern deltas at ±0.05. We read it as descriptive equivalence at this resolution, not as a null certified at twice the MDE. `factual_accuracy` also has weak inter-rater reliability ($\alpha \approx 0.13$, Section 11). A TOST null on this dimension only bounds our sampling-level power to detect a mean shift; a low-reliability construct may be too noisy to register a true non-zero effect, however large. The independent claim-level entailment decomposition below is a mechanical source-to-claim check, so the construct-reliability concern has no purchase there, and it carries the stronger version of this finding.

The split shows the paper's two arguments on one set of reports. Citation quality was retrieval-bound all along: give the pipelines clean sources, and the dimension that looked like a research difference in Section 6 mostly resolves. The per-pattern detail makes the mechanism concrete: the citation gain is largest exactly for the patterns that emitted the most placeholder citations under live search. The Pearson correlation between baseline placeholder rate and oracle citation gain is $r = 0.58$ across all 11 arms (the oracle regeneration also covered the two 7B agents for this correlation), and $r = 0.93$ on the six cluster patterns alone. With the base widened from the cluster to the full eleven-arm set, the gains run from +0.41 for the 25%-placeholder P2, through +0.21 to +0.35 for the 51–62%-placeholder P3, P4, P5, and P8, down to +0.07 and +0.00 for the already-clean P1 and P7 (P6, also already-clean, gains +0.10). Oracle sources do not teach these pipelines to source better; they replace the unresolvable placeholder strings of Section 6 with citable URLs, which is what the citation-presence rubric rewards.

Factual accuracy is bound somewhere else: in aggregate it shows no detectable change within the ±0.05 region of practical equivalence, and the per-pattern deltas are small and two-signed (P5 alone drops significantly, −0.046, CI $[-0.075, -0.013]$), so the synthesis ceiling holds at this resolution. The constraint on substance lives in how the model fuses evidence, and simply supplying the evidence does not relieve it. `information_recall` rises modestly (+0.071, CI $[0.033, 0.110]$), consistent with better sources surfacing more of the answer while that same synthesis bound caps how much of it the report states correctly.

The synthesis bound is a utilisation ceiling: more than a bare null. Of the factual shortfall, ideal retrieval closes only ≈ 0.09. The rest is what the model fails to *use*. Concretely, a claim-level entailment decomposition of the same oracle arm puts the gap between oracle and live grounding at 0.092 (95% CI $[0.030, 0.154]$), a within-metric difference that does not depend on any absolute level. Because the strict rule compresses all levels towards its floor, we bound what any partial-support rubric could change with a bracketing sensitivity: promoting every "neutral" verdict to supported, the loosest possible rule, moves the retrieval component from +0.092 to −0.011 (95% CI $[-0.019, -0.003]$). Loosening the support rule therefore shrinks the retrieval share, which makes the utilisation reading bracket-robust; the absolute level itself remains metric-dependent (footnote above). The measured decomposition therefore assigns almost all of the shortfall to a utilisation limit that no amount of retrieval quality reaches. The entailment verdicts are produced by a GPT-4o checker, the same base family as the generators. The family-of-judges constraint (Section 4) governs open-ended rubric judging, where family style preference has purchase that a binary source-to-claim entailment check largely lacks. A non-OpenAI re-verification of this sub-corpus remains the clean upgrade. This is the positive version of the argument a sceptic could otherwise dismiss as an under-

powered null. It is the paper's wedge past the well-known retrieval bottleneck.[7] The decomposition is over the five *entailment-covered* cluster patterns (P1, P4, P5, P7, P8; roughly 10–11 queries each) – a different five-pattern subset from *the inner five* of Section 5.3, which excludes P5, not P6; here P6 is the one missing, because it contributes oracle-arm judge scores above but has no live-retrieval claims in the claim-level entailment sub-corpus, so it is excluded from this decomposition. Per-pattern retrieval components span 0.005–0.182. The binding constraint lies in the model's failure to exploit ideal evidence it already holds: better retrieval could fix missing evidence, but not this.

The utilisation limit varies by claim type. Stratifying the oracle lift by what kind of claim it supports refines the null without contradicting it. This stratification runs on a wider matched population than the entailment-covered-pattern decomposition above: the eight patterns with both oracle and live claim data (the five entailment-covered cluster patterns plus P0, P9, and P10), pooled over patterns, and its own aggregate lift is therefore a related but distinct quantity from the five-pattern +0.092 retrieval component above; neither is a slice of the other (table below). On the 11 queries with matched oracle and live claims across these eight patterns, verified factual accuracy rises for single-hop and numeric claims (we call this class *grounding* claims) but stays flat for comparative and multi-hop claims (we call this class *synthesis* claims):

Claim population	Oracle lift (95% CI)	
All (aggregate, 8 patterns)	+0.079 ($[-0.002, 0.179]$)	spans zero
Grounding (single-hop, numeric)	+0.115 ($[0.036, 0.221]$)	excludes zero
Synthesis (comparative, multi-hop)	−0.004 ($[-0.103, 0.149]$)	spans zero

Ideal sources fix looking a fact up. They do not fix comparing or aggregating across facts, which is exactly the operation Section 6 and the dose ladder below locate the ceiling in. The flat aggregate therefore hides two different claim populations behind one number: a real grounding gain masked by a flat synthesis-claim channel. The split is directional at this sample size (claim-type tagging is heuristic and multi-hop claims are under-detected), so we read it as sharpening the mechanism and decline to certify it as a sub-result.

The clearest single view of the synthesis ceiling is an injected-gold dose ladder: the pipelines retrieve the gold, cite the gold, and paraphrase it into the prose, and still convert none of it to verified accuracy. It also closes the one loophole in the arm above, that the reports might never have *read* the injected sources, in which case a flat factual response would say nothing about synthesis. We run a dose ladder that raises the fraction of injected gold sources from 0 to 100%, as a CPU-only citation/lexical manipulation check on one instrumented query across P0, P1, and P4 (pooled gold-citation counts $0 \rightarrow 9 \rightarrow 18 \rightarrow 27 \rightarrow 36$ as the dose rises across the three architectures' reports on that single query, a 12-document gold pool). The reports cite essentially all of it: gold citations track gold available almost one-to-one, at a 100% resolution rate. The gold *content* also enters the prose (lexical recall ≈ 0.4 of the injected material). This one-query manipulation check establishes only that the gold material is retrieved, cited, and read. The factual-accuracy claim rests on the full ladder below, judged separately, where the slope is flat. A pre-registered mixed-effects fit sweeping the injected-gold fraction from 0 to 100% (GPT-5.2, n=450 dose points over 30 queries) returns a factual-accuracy slope of −0.001 (two-sided 95% CI $[-0.035, 0.033]$, p=0.95), against a citation-quality slope of +0.038 whose CI spans zero ($[-0.032, 0.108]$). The certified half of the split is the flat factual slope, set against the near one-to-one citation-count tracking above. Both halves of the ladder point the same way: the synthesis ceiling holds at the level of individual sources, so its support extends beyond the aggregate null (Figure 8).

[7] The *absolute* height at which grounding tops out, a *utilisation ceiling* of 0.320 (95% CI $[0.262, 0.378]$), should be read with more care. It is computed with the same strict verbatim-entailment metric that Section 6 argues is *unfit* for the deep-research register, where most claims are derived and multi-source. The strict bar reads the absolute level as artificially low, and a partial-support rubric lifts it.

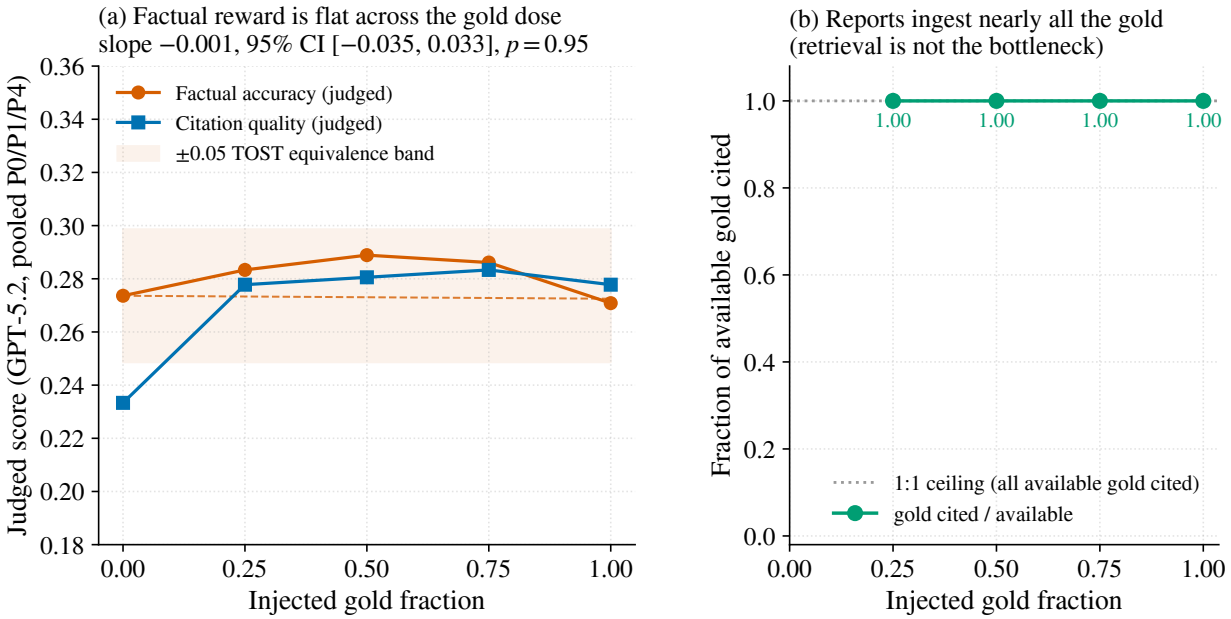

Figure 8: The injected-gold dose ladder makes the synthesis ceiling visible. (a) Judged factual accuracy is flat across the gold dose (mixed-effects slope −0.001, 95% CI [−0.035, 0.033], p=0.95; pooled P0/P1/P4, GPT-5.2, n=450 dose points over 30 queries). (b) A separate, CPU-only citation/lexical check on one instrumented query (P0/P1/P4, a 12-document gold pool) shows the pipelines cite essentially all the gold they are given: the ratio of gold citations to gold available sits on the 1:1 ceiling at every dose with gold to cite (the 0% dose has no available gold to ratio against and is omitted from the panel), at a 100% resolution rate. More gold brings more citations; verified facts stay flat.

Ideal sources shrink the orchestration gap. Every one of the nine GPT-4o patterns this design targets improves under oracle retrieval, and the cheap one improves most. Scores rise by +0.073 for the single-pass P0 and by as much as +0.090 for the supervisor pattern P2. (A separate correlation below also uses the two 7B agents' oracle re-scoring, where P10's point estimate falls instead, with a CI spanning zero.) The consequence is in Figure 9(b): the gap between P0 and the cluster mean narrows from 0.044 under live search to 0.010 under ideal sources, a compression of 0.034. Its paired query-bootstrap 95% CI, $[-0.02, 0.09]$, spans zero, so the compression reads as directional at this single-judge, 30-query resolution, short of a certified result. The high-baseline cluster patterns were already close to citation saturation, so they have less room to gain, while P0 starts lower and gains more. Part of this compression is by construction, since the pooled corpus grants the single pass the retrieval breadth it lacked. But that is the point of the intervention: P0's deficit was substantially a sourcing deficit, and orchestration's measured edge is partly a proxy for the better sourcing that the more elaborate pipelines happen to collect. When sourcing is equalised, the edge mostly closes. The cluster does not re-order into a clean architecture ranking under oracle retrieval, which is the outcome that would have refuted Section 5.3. It compresses towards the baseline instead.

What the test settles. The experiment we had named as the decisive missing one has now been run. The result matches what the earlier sections inferred: a retrieval ceiling on citation quality, and a separate synthesis ceiling on factual accuracy that ideal sources do not detectably lift (equivalent within ±0.05). Neither ceiling sits in the orchestration layer. A concurrent step-level faithfulness benchmark reaches the same conclusion observationally, from annotated reasoning traces: the bottleneck sits in synthesis, not retrieval [76]. The split also survives a change of judge. Re-scoring the cluster's oracle and baseline reports with a current-generation Opus/Sonnet crosscheck, each held within its own version so the delta carries no version artefact, reproduces the split on both: citation quality rises (Opus +0.045, Sonnet +0.095, both excluding zero), while factual

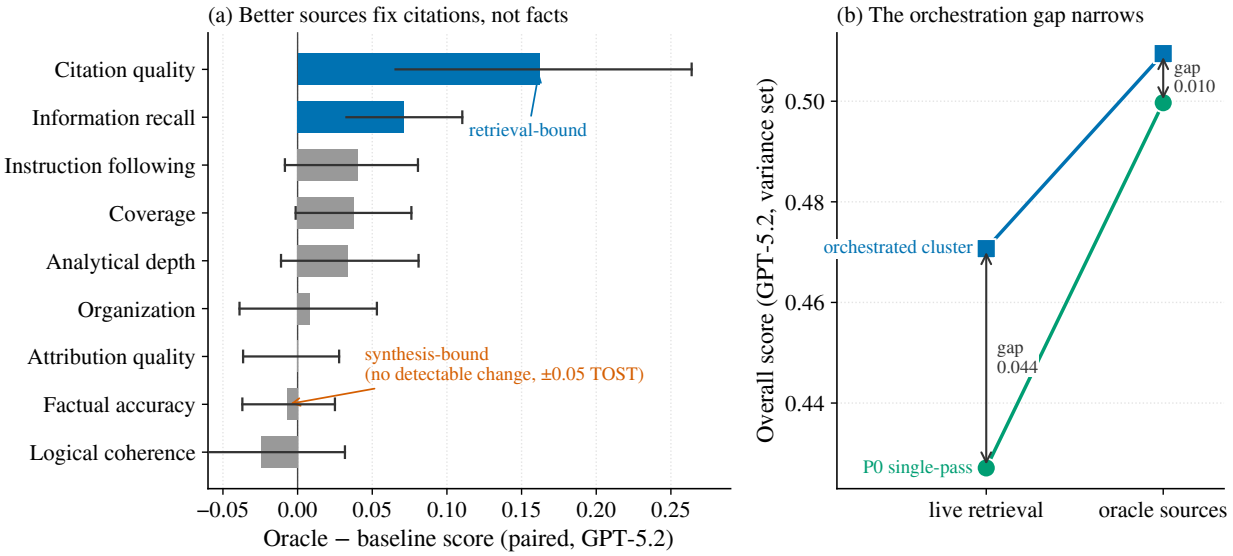

Figure 9: The oracle-retrieval intervention separates a retrieval ceiling from a synthesis ceiling. (a) Change in each dimension, pooled across the six cluster patterns, when every pattern is fed an identical ideal source set (paired, GPT-5.2, N=179 report pairs over 6 patterns; bars are mean oracle-minus-baseline deltas). The plotted whiskers are the 95% bootstrap CIs (citation quality $[0.066, 0.264]$), an interval run separately, resampled by pattern block, that closely agrees with the one quoted in the text ($[0.06, 0.26]$, which resamples the six patterns); a naive i.i.d. interval that treated the 179 pairs as independent would be 2.5× narrower, at $[0.123, 0.201]$. Citation quality jumps and both intervals clear zero; factual accuracy shows no detectable change and its interval spans zero (equivalent within ±0.05 by TOST). The same split holds when the cluster is re-scored by the current-generation Opus and Sonnet crosscheck, judge by judge (see text). (b) The single-pass baseline P0 and the orchestrated cluster, under live retrieval and under oracle sources: ideal sources lift the cheap pipeline more, so the architecture gap shrinks from 0.044 to 0.010 (point estimates; the change's 95% CI spans zero).

accuracy stays flat (Opus +0.004, Sonnet +0.007, both spanning zero).

Judge	Citation quality Δ (95% CI)	Factual accuracy Δ (95% CI)
Opus	+0.045 ($[0.020, 0.070]$)	+0.004 ($[-0.025, 0.032]$)
Sonnet	+0.095 ($[0.061, 0.127]$)	+0.007 ($[-0.019, 0.035]$)

GPT-5.2 and the Opus/Sonnet crosscheck therefore agree that ideal sources lift citation quality and not factual accuracy. The Opus and Sonnet citation gains are smaller than GPT-5.2's because the two Claude judges already rate the live-retrieval citations highly and so have less headroom. But the direction is unanimous. Two further points harden the reading. GPT-5.2 is the conservative judge for the citation claim – the one of the three that shows no citation-density bonus (Section 6). So its +0.16 registers a clean provenance effect. And the flat factual response cannot be a *global* strict-judge floor: a judge held at a floor could not have moved citation quality, and all three moved it. That leaves dimension-specific insensitivity of the judged factual dimension open. This is why the synthesis reading rests on the claim-level entailment channel above, with the judged dimension alone too weak to carry it. The one thing this pooled arm does not vary is the source-quality *dose*. The first rung of that ladder, the fitted injected-gold dose-response above, reproduces the same split (flat factual slope against one-to-one citation-count tracking), with the fuller ladder from high-recall sources up to gold-curated ones left as the remaining extension (Section 12).

7.1 A frozen-evidence defence: synthesis machinery, prior-override, and noise dosing

The oracle test above holds evidence *quality* fixed across architectures. A complementary triple experiment holds one identical evidence set fixed across *synthesis choices* instead, on a smaller and cheaper backbone (GPT-4O-MINI, where the rest of Section 5 uses GPT-4O). This bears on the mechanism question directly. It does not reproduce Section 5.3's exact contrast.[8] First, does the synthesis *scaffold* matter at all once evidence is frozen, or does the P1/P4-over-P0 factual gap of Section 5.3 require better retrieval to exist? We write the same frozen evidence through five scaffolds (a single pass, plus draft-then-revise, map-reduce, beam search over drafts, and a verifier-select loop) and compare each against the single-pass baseline on factual accuracy. One of four scaffolds separates, under both judges, and it is not the one with the most model calls:

Scaffold	Calls	GPT-5.2 effect	Sonnet-5 effect
Draft-then-revise	3	+0.079 ([0.038, 0.125], $p < 10^{-4}$)	+0.125 ([0.069, 0.181], $p < 10^{-4}$)
Map-reduce	7	n.s.	—
Beam search	7	n.s.	—
Verifier-select	45	n.s.	—

The three null GPT-5.2 point estimates run −0.013 to +0.029. We do not apply a multiplicity correction across this four-scaffold family, though the significant result would survive Holm correction at this p-value. Two of the three null CIs, map-reduce and verifier-select, are tight enough to exclude effects even half the size of draft-then-revise's; only beam's CI is wide enough not to rule that out. An independent Sonnet-5 judge, re-scoring the identical reports blind to the GPT-5.2 verdicts, replicates the same qualitative pattern above (a 58% larger point estimate than GPT-5.2's for draft-then-revise; the other three again do not separate). Factual accuracy does not track call count across the five arms, so the effect is architectural – one specific draft-then-revise loop does the work, which rules out a generic artefact of more compute buying more quality. This sharpens the oracle finding above without overturning it, and it qualifies Section 8's generalisation that only evidence-widening components help. Holding evidence exactly fixed does not make all synthesis machinery irrelevant; it isolates which piece of it is doing the work, and that piece is a narrow one.

Second, does the model follow frozen evidence that contradicts what it likely believes, or override that evidence towards its own prior belief? We construct evidence containing an attribute that counters the model's expected prior and classify each report's stance. This probe's construct validity is weaker than the scaffold probe above and the noise-dosing probe below. The counter-prior attributes are niche technical parameters (e.g., a compound's degradation temperature) that the model may hold no confident prior about at all, and the reporting prompt explicitly instructs the model to report sources' figures over its own knowledge. So a low override rate is not fully distinguishable from ordinary instruction-following. Overrides are rare under the primary, narrower classification: only 1 of 25 decided cases reverts to the prior (4%, Wilson 95% CI [0.7%, 19.5%], a wide interval at this n). A secondary, noisier whole-report classification on the same 30 probes gives a substantially higher rate (5 of 13 decided, 38.5%) – a coarser substring match over the full report and more prone to spurious hits, so not our primary reading, but the two differ by an order of magnitude. GPT-5.2 separately scores all 30 reports as reasonably faithful to the frozen evidence on factual accuracy (0.263, 95% CI [0.213, 0.317]). Here the two judge families genuinely disagree, and the disagreement goes well beyond strictness. An independent Sonnet-5 re-score of 29 of the same reports (one quarantined for Sonnet-family scoring, Section 11) rates factual accuracy substantially higher (0.591, 95% CI [0.483, 0.703]). More starkly, it rates attribution quality at 0.638 (95% CI [0.500, 0.759]) against GPT-5.2's 0.0, a literal zero on every one of the 30 reports. We inspected the underlying per-criterion verdicts directly: GPT-5.2 gives specific, substantive per-criterion reasoning on every report; none of the scores are

[8] This design speaks to the test-time-scaling readings of Zhang et al. [123] and Han et al. [27]. It can rule out "more retrieval passes" as a mechanism, since evidence is frozen identically for every scaffold, but it cannot adjudicate the retrieval-inclusive form of their claims. The result below is mixed: one of four scaffolds does separate, which falls short of a clean rebuttal.

null or malformed. The split is a genuine judge disagreement, not a scoring-pipeline artefact. It differs from the noise-dosing result below, which is noise scattered around a null: this split is large and systematic, and it lands on exactly the two dimensions, factual accuracy and attribution, that Section 6 and Section 5.1 already flag as the panel's least-agreed construct. We read the rule-based override rate, which is judge-independent by construction, as the load-bearing part of this probe. We read the faithfulness-score disagreement as a further illustration of the attribution-dimension construct problem. It does not show the judges substantively disagreeing about whether the writer followed the evidence.

Third, does orchestration's advantage over the single-pass baseline grow as the evidence degrades? The intuition is that synthesizing across noisier or more distractor-laden evidence is precisely where multi-step architectures should earn their keep. We dope the frozen corpus with distractor passages at four doses (0, 20, 40, 70%) and fit the slope of the cluster-minus-baseline gap against dose for P1 and P4 separately. Under GPT-5.2 the two patterns do not even agree in sign (P1 gap-slope +0.020, 95% CI $[-0.020, 0.059]$; P4 -0.015, CI $[-0.057, 0.024]$). Neither excludes zero. An independent Sonnet-5 crosscheck agrees on the null: P1 +0.033 (95% CI $[-0.050, 0.120]$), P4 +0.008 (95% CI $[-0.070, 0.087]$), again both spanning zero. This null is not offered as a defence: there is no statistically supported evidence, under either judge family, that orchestration's edge over the baseline widens as the evidence quality falls. An intuition this plausible deserves to be tested even though it does not survive the test.

8 Component Ablations

The oracle sets a ceiling on what any whole architecture can extract from better sources. If whole-architecture swaps barely move scores, do the individual components inside a pipeline? We ran a registry of seven component ablations across P3, P4, and P5, each removing one architectural piece and reusing the rest, scored on a symmetric two-judge basis (Table 13). The answer sharpens the cluster result without overturning it.

A few components matter and most do not. P4's cross-perspective triangulation is the single largest lever we measure ($\Delta = -0.060$ when removed), followed by its multi-perspective conversations (-0.036) and adaptive perspective count (-0.030), and P5's meta-evaluation loop (-0.047). Since report length itself moves scores (Section 5.9), we apply the same length-adjustment to this largest lever. Removing triangulation also shortens P4's reports (110 words on average, 60% of queries shorter). At grand-mean length the -0.060 effect attenuates to -0.041 (95% CI $[-0.057, -0.024]$, still excluding zero), so roughly a third of the abstract's headline "largest isolable component" figure is a length artefact, and the remainder is a significant content effect of the triangulation step. Even the *largest* significant component effect (-0.060 raw, -0.041 length-adjusted, both on the two-judge GPT-5.2+Sonnet basis this ablation check uses) is about one-sixth of the frontier–7B capability gap on the raw scale (≈ 0.38, three-judge). This ratio is not sensitive to the judge-basis difference, since the matched two-judge raw capability gap is close to the three-judge figure. We do not have a length-adjusted capability gap on the same two-judge basis as the ablation check. The only length-adjusted figure on record (0.110 for the cluster-vs-P9 gap, Section 5.9) is three-judge, and Opus alone pulls that figure up relative to what a two-judge version would show, so we do not report a length-adjusted like-for-like fraction here. The most valuable single piece of orchestration we can identify yields, at minimum, a small fraction of what swapping the backbone does. At the other end, P3's quality-evaluator self-loop is statistically *equivalent* to its removal under TOST at a ± 0.02 margin: a well-powered positive null. P3's topic-mining and P5's fixed-vs-adaptive width are indeterminate.

The pattern is interpretable, with one exception detailed in Section 7.1. Most components that help do so by changing *what evidence is gathered*: triangulating across perspectives, holding more parallel conversations. Most components that merely *re-process* already-retrieved evidence, such as P3's quality-evaluator pass over a fixed draft, do not move scores. The one counter-example is a draft-then-revise re-processing step, which Section 7.1 shows moves factual accuracy even when the evidence is held completely fixed. So "re-processing never helps" is too strong; the frozen-defence result is the cleaner test of which re-processing steps are the exceptions. Component surgery and whole-architecture swaps tell the same story from two

directions. Orchestration helps mostly insofar as it widens the evidence base or specifically revises against it, and its returns are small relative to the capability gap a model swap opens.

9 Failure Analysis

Components tell us what helps on average; failures tell us what breaks. The two families break in qualitatively different ways. Tagging the unsatisfied criterion verdicts by failure mode, the GPT-4o pipelines fail predominantly through *missing evidence*, *citation fabrication*, and *factual contradiction*, whereas the local 7B agents fail through *empty or sparse output*. Citation fabrication is an order of magnitude more common in GPT-4o reports, and empty output is roughly six times more common in 7B reports (χ^2 over the family×mode table, Cramér's $V = 0.44$, a 0–1 measure of how strongly two categories are associated, here a medium-to-large effect). The 7B "advantage" on fabrication is illusory: these systems fail earlier in the pipeline and are never assessed on citation quality. The starkest instance is P9 on DeepSearchQA, where it floors (≤ 0.05) on 17 of 20 queries. That is a degenerate "no gradeable output" regime: P9 returns nothing to grade on most DeepSearchQA queries, so the floors index missing output rather than uniformly mediocre quality. Its overall 0.26 mean, across all 90 queries, is therefore an aggregation artefact.

Five case studies, staged for release (see "Reproducibility and Artefact Availability"), make the cluster's internal texture concrete. The largest score gap we observe is P4 over P0 on a multi-criteria relocation query (0.72 vs. 0.01). Inspecting the transcript shows P0's near-zero score is a total generation failure on this query: every criterion verdict records no report content, and the returned text is the echoed query itself. This instance illustrates a pipeline-failure mode and carries no evidence about perspective-taking's architectural mechanism. The largest *reversal* is P1 over P4 on a narrow factual query (0.76 vs. 0.40), where iterative retrieval grounds a specific record that perspective expansion dilutes. Individual queries show large, opposite-signed effects *within* the cluster that is indistinguishable on average. That dispersion is why we report a cluster and stop short of ranking its members. The widest cross-judge disagreement not already used above (per-report SD = 0.38) arises on a report where Sonnet credits an unverified but plausibly written central claim as accurate, while GPT-5.2 withholds credit until it is independently corroborated. This is the same verification-strictness gap Section 5.1 documents at the panel level (Figure 2), visible here in a single report.

10 Discussion

10.1 A demonstrated ceiling beneath the orchestration layer

Three results locate the constraint on report quality below the orchestration layer: two observed across the main comparison and one now tested by intervention. The substance dimensions, `factual_accuracy` and `information_recall`, are bounded across the whole architectural axis. The architecture-fixed model swap moves them by 0.23 and 0.15 respectively, and moves the overall score by 0.23 (the widest cross-family contrast moves it by 0.38; Section 5.4). The two part company only under ideal sources: factual accuracy stays flat while information recall recovers a little (Section 7), so the synthesis ceiling is specific to factual accuracy. The one architectural signal that varies, `citation_quality`, is largely a judging artefact, and the underlying substance stays the same across patterns (Section 6).

The oracle-retrieval test of Section 7 feeds every pattern an identical ideal source set and finds exactly the structure the first two results imply. Citation quality is retrieval-bound and rises sharply under ideal sources. Factual accuracy is synthesis-bound and shows no detectable change within ±0.05. The orchestration gap compresses towards the baseline; no architecture ranking emerges. The binding constraint therefore sits below the wiring, in what evidence the tools surface and how the model fuses it.

This locates the same failure mode a concurrent study finds directly: deep-research agents adopt misleading source content as fact during long-horizon synthesis even when the identical content is correctly flagged by a verifier run in isolation [129]. Evidence *detection* already succeeds there; the fault sits in evidence *integration*. This is exactly the synthesis locus that our oracle intervention isolates.

It is a dual ceiling: retrieval bounds citation quality, and a separate synthesis limit, one that ideal sources

do not detectably lift, bounds factual accuracy. Both ceilings sit outside the orchestration layer this literature optimises.

The claim-level decomposition of Section 7 casts this synthesis limit as a positive finding, and its decisive half is metric-robust. Of the factual shortfall, the part that pooled near-ideal sources actually close is just 0.092 (Section 7); retrieval failure therefore accounts for very little of it, and almost all of the shortfall is what the model fails to use. The corresponding *absolute* ceiling near 0.32 is entailment-metric-dependent, and we do not lean on its level (Section 7). The residual is therefore a *utilisation* failure: the model cannot use ideal evidence it already holds, a bound no retrieval improvement reaches.

The reading agrees with the independent evidence and sharpens it. DeepScholar-Bench's oracle ablation [74] saturates retrieval-quality scores while nugget coverage stays near one-half. A contemporaneous controlled study finds that on multi-hop questions staged retrieval can even beat a one-shot ideal-evidence oracle: supplying all gold passages at once overloads the context window, and most of the residual errors are synthesis failures, occurring even when the evidence is present [4]. A study that varies the retriever and the generator independently over a frozen corpus finds that improving retrieval alone does not improve end-to-end answers, and can even raise confident hallucination when the answer is not in the corpus [64]. The same fragility shows up in retrieval-augmented generation, where adding more documents degrades accuracy by up to a fifth even when the context length is held fixed [48]. What our intervention adds is the architecture-controlled form of the result: the dual ceiling holds across nine pipelines on a single fixed model and tool layer, which lets us attribute it to the layer itself, since a particular agent's design cannot explain a result common to all nine.

10.2 Repositioning architectural claims

Our nulls neither prove nor refute the thesis that a single agent can beat a multi-agent one, but they narrow it. The cheap single-pass P0 underperforms the mean of the cluster (the six top patterns that are judge-robustly inseparable, Section 5.3) by ≈ 0.15 on the full panel, and the best single pipeline by ≈ 0.18, so *some* orchestration pays on the raw scale. The length-adjusted share of that step, though, is bounded at roughly 0–0.05 (Section 5.9). That step also does not exceed what resampling achieves: on the variance subset under GPT-5.2, a noise-decoupled best-of-N over independent P0 samples approaches the cluster within noise by seven to twelve samples, a split-direction-dependent point estimate (Section 5.7), at a spend comparable to the cluster's own. Whether that account extends to the larger full-panel gap is open until the replicates are re-scored with the cross-family panel.

Among the most elaborate, multi-agent-like GPT-4o patterns, P2's supervisor fan-out (outside the cluster) and P5's hierarchical scheduling (inside it) alike do not beat the cheaper iterative-retrieval P1. This is consistent with the equal-budget finding of Tran and Kiela [97] and Wang et al. [103]'s finding that a single prompt can match a full debate. It is also consistent with the empirically sceptical literature reviewed in Section 2 (the MAST failure catalogue, the capability-saturation law of Kim et al. [40], and the pre-registered human-judged replication and entropy-dynamics account discussed there), and with the ensembling literature *up to* the cluster ceiling.

The STORM triangulation claim [84] gets our strongest component-level support (-0.060 when ablated). MindSearch-style graph decomposition [14] keeps P7 competitive but adds nothing separable. On the single-agent-ReAct claim of Hu et al. [31], our verbatim ReAct probe (P11, *GPT-5.2 single-judge*) underperforms even P0, and the deficit survives a doubled turn budget (16 turns, 0.361 vs. 0.358, p=0.71). Primitive ReAct therefore does not beat orchestration here, though the wider debate stays open.

10.3 RL helps a 7B agent only when its objective matches the task

RL post-training lifts the 7B agent over its backbone (P10 > P9, $\Delta = 0.078$, posterior 1.00), as Section 5 reports: real, but a fraction of the capability gap. The sharper lesson comes from *which* RL works. DeepResearcher (P10) is trained with verifiable rewards on short-form QA. Concurrent work such as DR Tulu [83]

reports that SFT-and-RL against an evolving *long-form* rubric reward lifts an 8B agent to near-frontier long-form quality. Our own GRPO agent (P12), whose reward signal and training recipe both carry disclosed, undisambiguated limits (Section 5.12), does not improve over its base at all. The pattern across all three is that RL for deep research pays when the training objective matches the long-form synthesis task and the reward is reliable. So "RL lifts 7B agents to frontier" is true only with those qualifications, and under our setup the frontier capability gap dominates the RL-training effect we can measure.

A wave of concurrent work sharpens both halves of this picture. QUEST trains open models from 2B to 35B parameters via mid-training and RL against auto-verifiable "rubric-tree" synthetic rewards, reporting quality approaching or surpassing frontier closed agents on eight benchmarks [111]. Its success at even 2B scale is consistent with our reward-reliability mechanism, since the reward there is verifiable, a property learned rewards lack. But it complicates a naive reading of our own capability-gap finding, since it suggests well-constructed training can partly substitute for scale on its own benchmark suite, a question our fixed-scaffold, fixed-training design cannot adjudicate. Closer to our own setup, RubricEM trains an 8B agent with stage-structured rubric judgments and reports it approaching proprietary deep-research systems [49]. Its rubric-judge design keeps the scoring model *frozen* and evolves the text rubric itself, deliberately avoiding a trained critic. Its reward design is therefore structurally different from DR-Judge's. RubricEM offers no evidence that a fine-tuned judge model like ours can work when built well; if anything it is a data point for the opposite lesson, that avoiding a trained judge altogether may be the more robust design. We therefore attribute P12's failure to DR-Judge's own disclosed, specific limits (Section 5.12: non-textbook GRPO hyperparameters and a train/test format mismatch), without claiming RubricEM shows learned-judge rewards in general are salvageable. A frozen-judge, evolving-rubric design is one concrete alternative worth testing before training another judge model.

10.4 What a three-judge panel can certify

Our panel certifies relative orderings and within-cluster non-separation; it does not certify absolute levels on substantive dimensions. Per-dimension Krippendorff α ranges from 0.84 (`organization`) to −0.10 (`attribution_quality`), and judge stringency contributes real variance (ICC(judge) = 0.18). Following Guerdan et al. [24], we read the weak substantive agreement as rating indeterminacy, distinct from rater error. That is why we lead every headline on the judge-robust criterion and the cross-family panel average; no single judge's absolute scale carries the claim. Auditing the instrument this openly (bias regressions, self-preference tests, within-judge re-tests, the citation-density audit) is a requirement for null-heavy evaluation papers.

10.5 Generalisation boundary

The study is English-only and biased towards Western sources. It fixes a single tool layer, a single base model, and a single query distribution, and leaves out three axes that recent work shows can matter: tool-extended agents (code execution, private corpora), inference-time verification, and reasoning-native agents, where tool-calling is baked into the model's own reasoning trace – unlike every scaffold in this taxonomy, which externally orchestrates it around a non-reasoning chat model. Whether bounded returns to hand-coded control flow also bounds returns to a model's internal planning is an open question this design cannot answer.

The base model and the query distribution we test directly, and both headline findings survive that test (Section 5.10). The remaining axis, inference-time verification, is a lever that *works*. It lies beyond the width, depth, scale, and training axes our taxonomy varies, and it moves scores where orchestration does not. This is a distinct mechanism from Section 7.1's draft-then-revise scaffold: draft-then-revise re-runs the same synthesis step on frozen evidence, while here a rubric-guided verification pass checks a live-retrieval draft against the scoring rubric before revising it. It is also distinct from Section 7.1's null verifier-select scaffold, which chooses among several already-generated candidate drafts rather than revising a single one against the rubric; the two land on opposite sides of significance. Rubric-guided verification raises the overall score

by +0.045 on the single-call baseline P0 (95% CI $[0.016, 0.077]$, Wilcoxon p=0.006) and by +0.038 on the strongest pipeline P4 ($[0.018, 0.060]$, p=0.002). Each gain is measured against its own budget-matched unguided-revision control, so it isolates the verification *signal* from the compute a second pass supplies (the control equalises compute but does not guarantee equal output length). Both gains are significant on their own, and each survives a Holm correction across this two-test family (P4's smaller raw p takes the larger multiplier: $p_{\text{holm}} = 0.004$ for P4, 0.006 for P0). The probe is 30-query, single-judge GPT-5.2, on P0 and P4 only, and is consistent with concurrent test-time-verification work that lifts hard-subset accuracy by 8–11% [99].

In absolute size the +0.04–0.05 gain is modest, comparable to the largest single component lever we measure in the ablation registry (Section 8). We do not compare it against the ±0.05 equivalence margin used for the within-cluster non-separation claim, since that margin states what our power can certify as indistinguishable and does not bound the substantive size of an already-significant, control-matched effect. This is exactly what the thesis predicts, and is the clearest positive lever the paper identifies. Retrieval and synthesis are binding constraints; coordination is not. An inference-time step that re-examines the drafted answer against the rubric yields real quality across both a weak and a strong base pipeline – a return that re-arranging the control flow alone never delivers.

The oracle test of Section 7 probes the retrieval edge directly, but a different retrieval substrate, a different base model, or a verification-augmented pattern could still move it.

11 Limitations

We list the gaps in descending order of how much they bear on the headline claims.

No *full* human-calibration pilot. The panel is anchored to human and verifiable ground truth, but only partially. On the human side, we score the panel against expert labels on 5,765 human-labelled items across seven scored judge-family × public-set cells drawn from DRB-RACE, ExpertQA [63], DeepFact-Bench [32], and HealthBench (Table 11). On the verifiable side, we score its factual verdicts against 322 mechanically checkable verifiable-answer reports (LitQA2 + DeepSearchQA). GPT-5.2 emerges as the most human- and gold-aligned judge overall. It is the only panel member whose verifiable-gold factual AUC excludes chance (0.66; correct-minus-incorrect delta 95% CI $[0.024, 0.107]$), and it scores 0.707 AUC on HealthBench, the only judge evaluated on that set. The human-label cells are split: Claude Sonnet leads where it is scored (DRB-RACE Spearman 0.295 vs. GPT-5.2's 0.222, ExpertQA AUC 0.74, DeepFactBench AUC 0.749). The anchor we rest on is therefore the verifiable-gold asymmetry, not any single human-set cell.

Several judge × human-set cells await scoring (e.g., the panel on ExpertQA and DeepFactBench, HealthBench for the Claude judges), and none of the human agreements is strong, so the anchoring bounds *ordinal* validity; it does not certify absolute levels. We reinforce it with a within-judge re-test (GPT-5.2 $\kappa = 0.91$) and a separate version-drift probe. The probe re-judges a stratified 40-report sample with the now-current Claude models (Opus 4.8, Sonnet 5) against the banked Opus 4.1/Sonnet 4.5 verdicts. It finds only moderate agreement ($\kappa = 0.43$ Opus, $\kappa = 0.41$ Sonnet) and a significant, direction-inconsistent level shift (Opus 4.8 more lenient by +0.134, Sonnet 5 stricter by −0.072, both 95% CIs excluding zero). This is why the version-bumped judge is excluded from the banked panel and every certified headline separation (Section 4). Where current-generation Claude models appear elsewhere in the paper, they appear only as an explicitly labeled, standalone crosscheck. Large cross-task studies find LLM-judge agreement with humans of at best $\rho \approx 0.5$ (the strongest model's graded-task ceiling in that benchmark) [5]. That is the uncertainty one should attach to our absolute levels, though not to the cross-family ordinal claims. Completing the pending cells and running a stratified expert pilot is the first calibration we would add.

Each architecture induces its own retrieval pool; none draws on a shared one, and one channel ran degraded. Holding the tool code constant fixes the retrieval *interface* but not the evidence pool. Each architecture authors its own sub-queries, so the cited-source sets are predominantly disjoint across patterns, with no query reaching a half-shared cited-URL set. We read this divergence as part of the architecture's

treatment, not a confound to strip out. Where a claim must isolate architecture from retrieval breadth, we use the oracle backend (Section 7) as the fixed-evidence control. Separately, during corpus generation (Mar 16–Apr 3) the academic sub-channel had Semantic Scholar down (HTTP 429 throughout; the academic cache is 97.4% arXiv / 2.6% Semantic Scholar). As a result, academic evidence was effectively arXiv-only and rate-limited. Degraded calls returned empty and were never cached, so no report is corrupted. The condition was uniform across all eleven patterns and all 90 queries, leaving every cross-pattern contrast unaffected; the dominant web channel ($\approx 3.5\times$ larger) was 100% healthy throughout. The practical consequence is that the academic slice of the citation-quality narrative is understated, making the oracle citation ceiling of Section 7 a conservative floor: properly provisioned academic sources would lift quality *more*. A related worry is that differential *search flakiness* drives the cluster-over-baseline gain, with architecture playing no causal role. This is the third robustness check on that gain, alongside the best-of-N and matched-budget probes of Section 5.7 and Section 5.8, and it too leaves the gain standing. Conditioning the cluster-vs-P0 gap on per-arm search success leaves the architecture cluster coefficient statistically unchanged under arm-grouped inference (the statistical sense of "cluster," not the architecture sense), and the dead-or-fabricated URL rate (0.0–0.088 across arms) runs the wrong way to manufacture the gain (Section 6).

The tool interface is fixed, but the compute budget is not equalised. Token and tool-call budget is left free, rising 17.5× from the single pass to the heaviest pipeline. Unlike the budget-equalised priors we position alongside [40, 97], our architecture contrast therefore bundles control flow with the compute it spends. The matched-budget probe of Section 5.8 bounds the confound directly: under the clamp, the pipeline's (P1's) single-pass advantage is no longer detectable. The point estimate attributes $\approx 40\%$ of the effect to compute, but its CI is too wide to certify that share. The direct clamp effect is itself short of significance under both tests, and the probe covers one pattern, one judge, 29 queries, with the clamp landing at 3.2× baseline spend, short of the intended 1× (Section 5.8). The length-adjustment control of Section 5.9 converges: the length-robust share of the raw cluster-over-baseline gain is bounded at roughly 0–0.05. This is why we frame the result as bounded returns to architecture *and* its compute footprint. A fully budget-matched sweep across all patterns is the clean version we did not run.

The panel spans two families; a genuinely independent family is missing. The independence constraint holds against the *patterns*, but two of the three judges (Opus, Sonnet) share the Anthropic pre-training family. So for family-correlated dimensions the panel carries fewer than three independent votes (an effective-judge count near 1.4 by the participation ratio of the judge-correlation spectrum). This matters most for the citation-density bonus of Section 6, which is a Claude-family effect (present for Opus and Sonnet, a precise null under GPT-5.2). GPT-5.2 is the sole non-Claude check on it. A genuinely independent fourth family (e.g. Gemini) is the triangulation left to Section 12.

The ceiling is demonstrated on the cluster and a single source tier, short of the gold tier. (Recall from Section 5.3: *the cluster* is the six top patterns that are judge-robustly inseparable, and *the inner five* is the cluster minus P5. A third set of similar size appears below: the five *entailment-covered* cluster patterns of Section 7 exclude P6, not P5, so they are not the inner five under another name.) The oracle-retrieval arm of Section 7 shows by intervention that citation quality is retrieval-bound and factual accuracy is synthesis-bound. The split reproduces under the main-panel judge (GPT-5.2) and, independently, under a current-generation Opus/Sonnet crosscheck (Section 7) – a distinct check from the banked three-judge panel, which did not re-score the oracle arm. The pooled-existing corpus is a floor on ideal sources; it stops short of gold curation. We run a single source-quality tier, with no dose-response curve at this tier. The Claude crosscheck's re-scoring reaches the six cluster patterns, fewer than all eleven (the primary GPT-5.2 regeneration does cover all eleven, Section 7). (A separate, smaller dose-ladder check does inject curated gold sources, Figure 8, but only tests the factual-accuracy slope against dose on three patterns. It does not re-run this section's full citation/factual dual-ceiling decomposition at that source tier.) The Bing-vs-Tavily probe that motivated it remains single-judge.

Not every analysis spans all eleven architectures. The full three-judge run covers eleven architectures

on the shared manifest, but the downstream analyses narrow by design. The oracle-retrieval intervention re-scores all eleven patterns (the nine GPT-4o patterns under the primary design, with the two 7B agents added via an identical protocol, Section 7). The top-cluster inseparability analysis concerns six. The aggregate utilisation-ceiling decomposition rests on the five entailment-covered cluster patterns with paired oracle and live data (P6 has none – a different five-pattern subset from *the inner five*, which excludes P5 instead), while its claim-type stratification runs on a wider eight-pattern population (those five plus P0, P9, and P10). Because the populations differ, reading the stratification as a further cut of the same figure would conflate them. The second-backbone replication regenerates one pipeline (P4) against one baseline (P0) under one judge, and the 7B-orchestration probe is twelve queries scored single-judge. The headline nulls are read strictly at the coverage each analysis actually has.

Substantive-dimension reliability is weak. Krippendorff α is 0.11 (`information_recall`), 0.13 (`factual_accuracy`), and −0.10 (`attribution_quality`). We therefore certify rankings but stop short of absolute substantive levels, and treat `attribution_quality` as the least trustworthy dimension throughout.

The family-of-judges constraint rests on empirical evidence; no structural guarantee backs it. GPT-5.2 and the two Claude judges may share pre-training distribution. Our self-preference test controls only the rank gap within a judge and cannot detect Anthropic-family preference (no Anthropic-family generator exists among the patterns). Qwen and DeepSeek are excluded as judges because they are the P9/P10 backbones. A genuinely independent fourth family (e.g., Gemini) is the natural triangulation.

Run noise approaches the equivalence margin. A replicate experiment puts per-cell run noise at $\sigma_{\text{run}} \approx 0.05$ (P0 eleven-replicate anchor 0.049; pooled estimates 0.047–0.059), of the same order as the ±0.05 ROPE. The single-run headline means should be read at that precision, and the equivalence claim is power-justified – a weaker claim than genuine substantive equivalence (0 of 15 pairs equivalent at the tighter ±0.02 margin). The method-stable judge-robust result (0 of 10 inner-cluster pairs separated) does not depend on this margin.

Post-hoc, single-judge probes. P11 (ReAct), P12 (our RL agent), the 7B orchestration arm, and the Bing-vs-Tavily intervention are reported on a GPT-5.2 single-judge basis and were added after the main run. Their numbers carry a single-judge tag wherever they appear and should not be read as panel-certified (Section 5.12).

Negative by-products. DR-Judge-7B missed its $\kappa \geq 0.7$ target, and its intended release is scoped to a triage tool only. P12's flat training curve is a bounded result on RL supervised by a rubric judge; it makes no general claim about RL.

Scope and contamination asymmetry. All 90 queries are English with a Western-source slant, and web search is non-deterministic across time despite stored logs. Because all patterns search the live web, they are also exposed to search-time contamination, the retrieval of benchmark-derived pages during evaluation [104], though symmetrically so, since all eleven share one tool layer. Benchmark contamination is plausibly symmetric across patterns that share a base model, so the orchestration contrasts are unaffected, but *not* symmetric across the frontier-vs-7B contrast: GPT-4o and Qwen2.5 have different pre-training corpora, so differential contamination could inflate the very capability gap that is our second finding. We therefore test that gap on the clean residual as well (below), and it attenuates but survives. Where a contrast must be insulated from search-time contamination entirely, we use the frozen-source arms: the pooled oracle corpus (Section 7) and the hash-pinned frozen-vintage source set (Section 5.4). These fix one identical, timestamped evidence set per query, removing live-search drift and benchmark-page retrieval from those specific comparisons; the live-search runs remain exposed.

We also test both headline contrasts against contamination directly. A regex-and-classifier detector over each report's citation and search snippets (metadata-host and question-context buckets) flags 73 of the 90 queries as potentially contaminated. Dropping them and recomputing on the 17-query residual leaves both intact. The top orchestration cluster stays flat (no within-cluster pair separates). The cluster-over-P0 lift is if anything larger: +0.176 versus +0.152 on the full set (both on the five-pattern inner-cluster basis this detector analysis uses). The one thing that moves is *which* cluster member ranks first (P1→P5), but that is the same

small-sample non-identifiability we report everywhere. On 17 queries the rank-1 shuffle sits inside the flat top band and is expected sampling noise. All three capability gaps shrink on the clean slice, but only one shift is large enough that the full-set value falls outside the clean-slice interval: P0-vs-P9 falls from 0.231 to 0.142, and cluster-vs-P9 (the six-pattern cluster of Section 5.3; 0.376 on Section 5.9's pooled-OLS pipeline, computed independently and closely agreeing, not a second population) falls from 0.375 to 0.318; the same 17-query recompute covers the P4-specific gap quoted earlier (Section 5.4) too, all three gathered here:

Gap	Full set	Clean 17-query slice	Full-set value vs. clean CI
P0-vs-P9	0.231	0.142 ([0.067, 0.224])	outside CI
P4-vs-P9	0.382	0.337 ([0.250, 0.423])	inside CI
Cluster-vs-P9 (six-pattern)	0.375	0.318 ([0.232, 0.404])	inside CI

The P0-vs-P9 shift is consistent with contamination flattering the frontier backbone on flagged queries. Finding (ii) therefore stands directionally on the decontaminated slice. We quote the 0.23 backbone gap with its decontaminated companion 0.14 in mind. The survival criterion, a flat top cluster, a positive P0 gap, and a positive capability gap, holds. The leader's identity, as throughout the paper, does not. The detector's precision is unaudited (no manual review of the 73 flags), so the residual is conservative by construction; certifying it clean is a separate, unmet bar.

Multiplicity beyond the headline. The four-gate hierarchy (Section 4) governs the headline separations, but the paper reports many secondary contrasts, and we control their family-wise error within each family. The nine-dimension oracle deltas and the per-dimension gates are Holm-corrected across the nine dimensions. The pairwise separations are Holm-corrected within judge, and this is invariant to a single joint Holm family over all 165 judge-level tests. The six-pattern equivalence-testing battery of Section 5.3 is Holm-corrected across its fifteen-pair family under both the paired-Wilcoxon and t-based TOST variants. The component-ablation registry is Holm-corrected across its arms, the per-dimension effective-judge excess tests are Holm-corrected across the nine dimensions, and the rubric-guided-verification result of Section 10.5 is Holm-corrected across its two-pipeline family (P0, P4). Contrasts we report as descriptive, the per-source leader board, the individual per-query case studies, and the citation-density/provenance regression battery of Section 6 (fit separately per judge, with and without query fixed effects, pooled and wild-cluster variants), are labelled as such and carry no family-wise guarantee.

We did not attempt a bias-corrected judge level. A Rogan–Gladen or prediction-powered correction is a standard technique for adjusting a fallible measurement back towards ground truth, once you know how often that measurement is right and wrong. Applying either one to our raw judge levels needs per-criterion judge sensitivity and specificity against gold. The one gold source fine-grained enough to try, the verifiable-answer slice underlying Table 11, is unsuitable for this specific purpose. Its own construction note records that the answer-present signal is confounded with general report completeness; it is not dimension-specific. Agreement with it is in the Cohen's κ "poor" band for all three judges, and two of three judges discriminate at within noise of chance (AUC 0.50–0.54). Building a formal correction on that foundation would turn a weak, construct-compromised proxy into an apparently precise number. We therefore report raw judge levels throughout. The defensible use of the panel is the rank-order and contrast robustness established across the paper.

Further robustness material exists in the released store beyond what we narrate here. Several completed, $0-marginal-cost analyses corroborate results already established in the main text. These are a per-adapter curve of reward-hacking onset and its current-Claude crosscheck (RL post-training, extending Section 5); a factor-analytic check of the rubric's own construct validity (an exploratory factor analysis of the nine dimensions on panel-marginalised scores finds one dominant general factor, consistent with the weak per-dimension substantive agreement of Section 10.4); a per-source complexity gradient; and a distilled-judge Youden-J operating-point analysis extending Section 5.12. None of these change a headline finding.

They are released in `canonical_numbers.json` for readers who want the fuller robustness picture.

12 Future Work

Our results name their own next experiments, in descending order of how much they would change the paper's claims.

The oracle-retrieval arm, extended to a dose-response. *Partly run, partly open.* We ran the pooled-source form of this experiment and re-scored the cluster under GPT-5.2 and, independently, a current-generation Opus/Sonnet crosscheck (Section 7). All three judge instances reproduce the dual ceiling. We also fitted the first rung of the dose-response (Figure 8, Section 7). The flat factual slope against near one-to-one citation-count tracking confirms that the synthesis ceiling holds *across the dose*. The remaining extension is the fuller ladder: climbing from pooled-existing to fresh high-recall to gold-curated corpora at higher resolution and under the full panel. There we expect the flat factual slope to persist at every rung, while the retrieval gain on citations continues to grow.

A compute-matched orchestration control, strengthened. We ran the single-judge form of this control, decoupled via a split-half design (Section 5.7). With selection decoupled from scoring, best-of-N over independent P0 samples draws level with the cluster at roughly seven samples on the point estimate, and stays level at matched spend: neither resampling nor orchestration beats the other on the variance subset. Two extensions would complete it. The first is to replace oracle selection based on held-out criteria with a realisable selector (self-consistency or a held-out scorer model), since even decoupled selection is an upper bound. The second is to re-score the P0 replicates with the full cross-family panel, so the control speaks to the larger panel-level gap; the smaller single-judge gap on the subset answers a narrower question.

A full human-validation pilot. Our panel is reliability-audited and *partially* validity-anchored (Table 11). Seven judge-family × public-set cells and a 322-report verifiable-gold slice are already scored, and they establish GPT-5.2 as the most human- and gold-aligned judge. What remains is to fill the pending cells and add a stratified pilot of 20–40 reports rated by domain experts. This would calibrate the panel against ground truth on every dimension, bound the absolute scores (which cross-task work puts at best $\rho \approx 0.5$ agreement with humans), and let us report which dimensions the panel can certify at the level, beyond just the ordering.

A grounding-aware citation rubric and a fourth judge family. The Section 6 confound is rubric-conditioned: a citation-*presence* rubric that never shows the judge the cited source rewards density. The fix is a source-supplied grounding rubric (FACTS-Grounding-style [34]) and a partial-support entailment metric for the deep-research register. Adding an independent judge family (e.g., Gemini, by now a standard third-lab addition to LLM-judge panels) would also let us separate GPT-family-adjacent inflation from a true capability gap. We lacked deployment access to a Gemini endpoint under this study's infrastructure.

Task-matched RL for the open models. Our RL result is bounded by the training objective more than by RL per se: P10 is trained on short-form QA, and our own P12 used a noisy rubric-judge reward. Re-training the 7B agents against a task-matched, long-form, evolving-rubric reward [83] would test whether the capability gap is closable by post-training. This is the axis DR Tulu suggests is decisive.

From "how much" to "when": a router closes the oracle gap but leaves the realised one open. The per-source leader rotation (Table 8) and the large sign-flipping per-query effects (Section 9) indicate that architecture matters *task-conditionally*. So we tested whether a per-query router can turn our null result on the grand mean into a positive, actionable one. On the 85 queries with complete three-judge coverage across all eleven patterns, a per-query oracle – the best architecture on each query with hindsight – beats the single best-fixed architecture (P1) by 0.056 (95% CI $[0.039, 0.075]$, excluding zero). The headroom is real. A deployable router does not reach it. A leave-one-query-out ridge-argmax selector, trained only on features available before running (query length, expected topic count, difficulty, source, causal-question count, entity density) with outcome scores withheld, realises a mean of 0.669 against the best-fixed baseline's 0.675. The delta is -0.007 (95% CI $[-0.027, 0.013]$, an interval that spans zero; probability of beating the

fixed baseline 0.26). At n=85 queries and this feature set, routing does not detectably beat always choosing the best-fixed architecture, let alone reach the oracle. The gap between oracle headroom and realised routing is itself the finding: a quantity that is actionable in principle but unrealised in practice, consistent with the paper's bounded-returns theme. Whether a larger manifest or richer query-level features would close that gap, or whether it is a hard ceiling at this sample size, remains open. Extending the manifest to multilingual and tool-extended (code, private-corpus) settings would map the generalisation boundary.

13 Conclusion

Across eleven deep-research architectures on a single shared tool layer, orchestration moves report quality far less than the system-by-system literature implies. The single-pass baseline plus eight orchestration pipelines span 0.185. Five of them are judge-robustly inseparable. The length-robust share of even the baseline-to-cluster step is bounded at roughly 0–0.05, and the frontier–7B capability gap is roughly twice the entire span. The one architectural signal that does vary is citation quality. This is largely an artefact of judges that add a density bonus on top of grounding – a confound two concurrent studies independently corroborate. An oracle-retrieval experiment that serves every pattern an identical ideal source set then shows the rest directly. Citation quality rises by 0.16. Factual accuracy shows no detectable change (equivalent within ±0.05). A claim-level entailment decomposition positively locates the shortfall: ideal retrieval closes only ≈0.09 of it; the rest is a utilisation limit, covering evidence the model already held and failed to use. And the baseline-to-cluster gap narrows while the ordering survives, which places a retrieval-and-synthesis ceiling beneath the orchestration layer. Three implications for the research agenda follow. First, orchestration cleverness has bounded, task-conditional returns on the GPT-4o-class backbones we test. Second, model capability and inference-time verification are the levers worth pulling: verification lifts scores by +0.04–0.05 where rearranging control flow does not, while retrieval quality is a narrower lever that raises citation quality without touching the synthesis bottleneck. Third, grounding-aware citation judging remains an unsolved methods problem. The field treats it as a rubric afterthought; it deserves first-class objective status.

Reproducibility and Artefact Availability

Every pattern mean, dimension score, regression coefficient, interval, and verdict count in this paper is regenerated from the underlying criterion-level data by a single build script, which also regenerates every figure and table the paper inputs. A reconciliation script then checks every numeric cell in the generated tables against the canonical number store, so that no table can drift from the data. (All 430 numeric cells of the tables this paper inputs trace to the store; over the wider release the count is 916 of 920, the four misses being arXiv identifiers in tables from companion analyses.) The headline inline numbers in the prose are drawn from that same canonical store, and a companion prose-reconciliation pass pins the load-bearing inline statistics of the abstract, contributions, and results (86 anchored statistics) to their store keys. (The verdicts are clustered within 990 (pattern, query) cells (985 realised reports) over 90 queries. The unit of independent evidence for between-pattern inference is the query, not the verdict.) The archive currently live at the DOI below is a prior revision (v2). It already comprises the full analysis apparatus: every build and reconciliation script, the 90-query manifest, the nine-dimension rubric and per-query coverage criteria, and the ablation registry specification, together with a data dictionary documenting every released column. But its canonical number store predates several results, including the oracle-retrieval synthesis-ceiling test this paper is titled around. A refreshed archive version (v3), rebuilt from this revision's canonical store immediately before submission, is staged for upload. It includes the underlying 248,536 criterion-level verdicts (broken down by population in Table 6) with judge rationales and evidence quotes, the oracle and Bing-vs-Tavily intervention sets, the P0/P1 replicate runs, the five case studies, and the DR-Judge-7B LoRA adapter (Section 5.12). The judge prompt templates are not bundled in this archive and, like the rest of the v3 material, are available from the authors on request. None of this v3 material is yet part of the public record. The code and data ship under the Apache License 2.0. Queries drawn from third-party benchmarks are redistributed as identifiers plus prompts

where the source licence permits, and as identifiers alone otherwise. Judge outputs will be released under the respective providers' output-usage terms. The archive accompanies this paper as supplementary material and is publicly archived at Zenodo under the concept DOI 10.5281/zenodo.21118281, which resolves to the latest version.

A Complete Results Tables

Table 3: Per-pattern overall scores (three-judge average; the SONNET column uses its recomputed overall, since its stored overall is corrupted upstream). *N* is judge×query cells.

Pattern	*N*	Mean	SD	GPT-5.2	Opus	Sonnet
P1 Iterative RAG	268	0.673	0.215	0.467	0.763	0.793
P4 STORM	270	0.640	0.172	0.467	0.640	0.812
P6 Reactive	261	0.634	0.200	0.461	0.639	0.802
P7 Graph	270	0.630	0.190	0.447	0.671	0.773
P8 Beam	270	0.625	0.193	0.432	0.674	0.768
P5 Hier. W&D	267	0.601	0.198	0.428	0.601	0.773
P2 Supervisor	270	0.585	0.197	0.398	0.653	0.705
P3 MERIDIAN	265	0.570	0.177	0.406	0.592	0.713
P0 Single-pass	270	0.488	0.247	0.385	0.475	0.606
P10 DeepResearcher	270	0.336	0.195	0.213	0.318	0.476
P9 Qwen2.5-7B	270	0.258	0.259	0.180	0.243	0.350

Table 4: Per-pattern × per-dimension mean scores (three-judge average).

Pattern	Info. recall	Factual	Coverage	Depth	Citation	Coherence	Org.	Instr.	Attrib.
P1 Iterative RAG	0.60	0.61	0.70	0.66	0.78	0.75	0.96	0.74	0.43
P4 STORM	0.55	0.55	0.74	0.83	0.49	0.75	0.99	0.72	0.21
P6 Reactive	0.58	0.56	0.64	0.62	0.72	0.76	0.98	0.72	0.30
P7 Graph	0.45	0.61	0.66	0.71	0.79	0.70	0.94	0.60	0.29
P8 Beam	0.46	0.59	0.70	0.80	0.59	0.73	0.97	0.64	0.23
P5 Hier. W&D	0.56	0.53	0.67	0.70	0.48	0.70	0.97	0.68	0.21
P2 Supervisor	0.50	0.54	0.65	0.66	0.52	0.73	0.98	0.61	0.22
P3 MERIDIAN	0.52	0.50	0.64	0.67	0.39	0.75	0.94	0.67	0.19
P0 Single-pass	0.36	0.46	0.47	0.46	0.55	0.65	0.88	0.58	0.30
P10 DeepResearcher	0.31	0.32	0.34	0.21	0.29	0.49	0.81	0.46	0.13
P9 Qwen2.5-7B	0.21	0.24	0.26	0.19	0.25	0.40	0.52	0.34	0.15

Table 5: Inter-rater reliability of the three-judge panel.

Statistic	Value
Krippendorff α (overall)	0.423
ICC(A,1)	0.489
ICC(A,k=3)	0.742
Per-dimension α:	
Info. recall	0.105
Factual	0.135
Coverage	0.542
Depth	0.572
Citation	0.542
Coherence	0.097
Org.	0.841
Instr.	0.276
Attrib.	-0.102

Table 6: Criterion-level verdict counts, regenerated from the underlying parquets (staged for a forthcoming archive version; see "Reproducibility and Artefact Availability").

Population	Verdicts
Base patterns (P0–P10)	111,984
Post-hoc single-judge probes (P11, P12, P11 at 16 turns, 7B replicates)	19,913
Ablations (7 interventions)	46,844
Variance experiment (reruns)	29,034
Bing-vs-Tavily intervention	13,139
Oracle-retrieval intervention	26,135
Tool-layer disentanglement probe	1,487
Total released verdicts	**248,536**
≥2-judge consensus triples	79,972
3-judge complete triples	50,298

Table 7: Citation-provenance audit: placeholder vs. grounded citations per pattern.

Pattern	Cites/rep	Placeholder %	Academic %	Real-URL %
P1 Iterative RAG	42.8	0.2	23.3	76.3
P4 STORM	21.4	61.6	9.4	29.0
P6 Reactive	24.6	0.3	23.4	76.1
P7 Graph	30.4	0.2	24.3	75.3
P8 Beam	34.8	56.8	14.1	29.1
P5 Hier. W&D	57.5	50.8	9.4	39.8
P2 Supervisor	13.8	25.0	24.6	50.4
P3 MERIDIAN	13.0	59.1	8.3	32.6
P0 Single-pass	6.0	0.0	28.6	71.4
P10 DeepResearcher	9.1	37.0	19.5	43.4
P9 Qwen2.5-7B	4.1	0.0	41.4	58.6

Table 8: Per-source leader board (three-judge mean). The *identity* of the best pattern is source-conditional, not pattern-intrinsic.

Source	Best	2nd	3rd
Custom (n=5)	P6 (0.769)	P5 (0.736)	P7 (0.726)
DeepSearchQA (n=20)	P7 (0.575)	P8 (0.573)	P4 (0.569)
DRACO (n=40)	P1 (0.679)	P4 (0.660)	P6 (0.638)
LitQA2 (n=10)	P1 (0.681)	P6 (0.603)	P8 (0.589)
ResearchQA (n=15)	P1 (0.811)	P6 (0.725)	P8 (0.722)

Table 9: Tool-layer intervention: Bing vs. Tavily backend (*GPT-5.2 single-judge*; paired bootstrap 95% CI over 28–29 stratified queries). A cleaner-URL backend that returns fewer citations *depresses* every pattern.

Pattern	Bing	Tavily	Δ	95% CI
P0	0.408	0.239	-0.169	[-0.260, -0.076]
P1	0.504	0.391	-0.113	[-0.164, -0.064]
P3	0.445	0.398	-0.047	[-0.087, -0.010]
P4	0.481	0.399	-0.082	[-0.118, -0.045]
P5	0.488	0.366	-0.122	[-0.180, -0.065]
P8	0.459	0.370	-0.089	[-0.128, -0.053]

Table 10: DR-Judge-7B agreement with the panel consensus (held-out test split).

Subset	κ	n
All test verdicts	0.453	3,824
Undisputed (panel unanimous)	0.652	2,143
Disputed (panel split)	0.199	1,681

Table 11: What the panel *is* anchored to (Section 11). Each cell reports a scored judge-family × ground-truth agreement; the human-set columns give the agreement metric for that public set (Spearman for DRB-RACE's graded tasks, AUC for the binary factual sets), and the verifiable-gold column gives factual-accuracy AUC against the LitQA2+DeepSearchQA slice (GPT-5.2 n=322; Claude Sonnet n=290; Claude Opus n=259, the judges' report coverage differs). Cells marked — await scoring. GPT-5.2 is the only judge whose verifiable-gold AUC excludes chance; it is also the only judge scored on HealthBench (0.707 AUC), while Claude Sonnet leads on the human-set cells it is scored on. DRACO-full is excluded for label leakage.

Judge family	DRB-RACE[a] (Spearman)	ExpertQA (AUC)	DeepFactBench[b] (AUC)	HealthBench (AUC)
GPT-5.2 (OpenAI)	0.222	—	—	0.707
Claude Sonnet (Anthropic)	0.295	0.74	0.749	—
Local 7B	—	0.545	0.556	—

Verifiable-answer gold (factual AUC; LitQA2+DeepSearchQA):
GPT-5.2 **0.66**, n=322 (excludes chance) Claude Opus 0.54, n=259 Claude Sonnet 0.50, n=290 (both near-chance)

[a] DRB-RACE agreement is Spearman over per-dimension cells nested in 50 annotated tasks (n=1000 cells for GPT-5.2, 965 for Sonnet); the small task count is the clustering unit for its confidence interval.

[b] DeepFactBench AUC rests on 621 claim-level judgements nested in only 15 report clusters (n=600 scored claims per judge); the 15-cluster basis is the reason its interval is the widest in the table and its cell is read as indicative rather than certified.

Table 12: External-validation battery (Section 5.10). A fresh 600-report set, five query families × {P0, P1, P4} × 40, generated by GPT-4o and scored by the independent GPT-5.2 judge; means recomputed directly from the 600 verdicts (n=200 reports per pattern). Both orchestration-cluster pipelines reproduce their lead over the single-pass baseline P0 out of sample. Three caveats attach. First, this is a *partial* replication: three of four planned layers are computed, and the fourth, a correlation of our verdicts against the DeepResearch-Bench RACE expert annotations, is carried as BLOCKED because those external annotations are not on disk and this harness does not fetch external data by design. Second, the pre-specified three-pattern flat-cluster check formally returns FALSE on this slice, driven entirely by the within-cluster P1 ↔ P4 order (a 0.034 gap, not a tier flip: both stay well above P0), the same inner-order non-identifiability we report throughout, not a new separation. Third, rank concordance against the main leaderboard is Spearman $\rho = 0.5$ (n=3 patterns, ordinal only), reflecting that same swap.

Pattern	Battery mean	n (reports)
P1 (Iterative RAG)	0.4862	200
P4 (STORM)	0.4526	200
P0 (single-pass)	0.183	200

Rank concordance vs. main leaderboard: Spearman $\rho = 0.5$ (n=3, ordinal; P1 ↔ P4 swap)
Tiering: P0 ≈ 0.18 on this harder battery sits at or below the local-7B agents' scores on the main 90-query manifold (P9 0.18, P10 0.21, same single-judge basis); on that same basis P0 itself scores 0.38, and its three-judge panel mean is 0.49 (individual judges span 0.38–0.61), so the apparent tie is specific to this harder battery, not a general P0-vs-7B result

Table 13: Component-ablation effects (two-judge GPT-5.2 + Sonnet basis). Each row removes one architectural component and re-uses the rest; Δ is the paired oracle-minus-baseline overall change. The Holm p column applies a Holm–Bonferroni step-down correction across the seven-contrast family; † marks the four contrasts that survive at α=0.05. Four of seven survive: P5's fixed-vs-adaptive width is raw-significant (Wilcoxon p=2.3×10^{-2}) but not significant after correction (p_{Holm}=6.9×10^{-2}), consistent with the indeterminate reading in the text.

Ablation	N	Δ	95% CI	Wilcoxon p	Holm p
P3 − quality-eval	87	+0.002	[-0.013, 0.019]	1.0	1.0
P3 − topic-mining	86	+0.008	[-0.010, 0.026]	5.3×10^{-1}	1.0
P4 − adaptive persp.	90	-0.030	[-0.044, -0.017]	3.1×10^{-5}	1.2×10^{-4}†
P4 − conversations	90	-0.036	[-0.052, -0.020]	4.2×10^{-6}	$< 10^{-4}$†
P4 − triangulation	90	-0.060	[-0.080, -0.043]	9.6×10^{-10}	$< 10^{-4}$†
P5 − adaptive width	82	-0.023	[-0.045, -0.004]	2.3×10^{-2}	6.9×10^{-2}
P5 − meta-eval	86	-0.047	[-0.065, -0.028]	6.0×10^{-6}	$< 10^{-4}$†

Table 14: Single-judge (GPT-5.2) per-pattern means, including the post-hoc P11 (ReAct) and P12 (own RL) probes.

Pattern	GPT-5.2 mean	SD	N
P4 STORM	0.467	0.098	90
P1 Iterative RAG	0.467	0.136	90
P6 Reactive	0.461	0.148	87
P7 Graph	0.447	0.107	90
P8 Beam	0.432	0.106	90
P5 Hier. W&D	0.428	0.129	89
P3 MERIDIAN	0.406	0.126	89
P2 Supervisor	0.398	0.127	90
P0 Single-pass	0.385	0.183	90
P11 ReAct	0.331	0.147	89
P10 DeepResearcher	0.213	0.133	90
P12 RL (ours)	0.182	0.160	90
P9 Qwen2.5-7B	0.180	0.155	90

Table 15: Orchestration premium at 7B vs. at GPT-4o scale (*GPT-5.2 single-judge*, 12-query subset). Running the P1 and P4 architectures on the Qwen2.5-7B backbone yields premiums of the same small order as at GPT-4o scale; neither closes more than a fraction of the 0.38 capability gap.

Architecture	Qwen2.5-7B backbone	GPT-4o backbone
Single pass (P0 architecture)	0.3039	0.472
P1 architecture (iterative RAG)	0.3174	0.525
P4 architecture (STORM)	0.3704	0.477
P1 premium over single pass	+0.014 [-0.05, +0.08], p=0.68	+0.053 [-0.03, +0.14], p=0.45
P4 premium over single pass	+0.067 [-0.01, +0.14], p=0.23	+0.005 [-0.07, +0.09], p=0.93

Table 16: Best-of-k over independent P0 replicates (*GPT-5.2 single-judge*, 30 variance queries). The naive curve, which selects and scores with the same judge score, is indistinguishable from the pure-noise max-order-statistic prediction at the measured within-query SD, so its crossing of the cluster is a winner's curse. The decoupled curve (selection on one criterion half, scoring on the held-out half) draws level with the cluster's held-out-basis reference at $k \approx 7$ (a split-direction-dependent point estimate: the reversed split never crosses through k=12, and the paired gap's query-bootstrap 95% CI spans zero at every k) and does not significantly exceed it at any k.

k	1	2	3	4	5	7	9	12
Naive best-of-k (same-judge selection)	0.427	0.446	0.462	0.468	0.475	0.488	0.493	0.495
Pure-noise prediction (σ=0.050)	0.420	0.448	0.462	0.471	0.478	0.487	0.494	0.501
Decoupled best-of-k (held-out scoring)	0.432	0.438	0.443	0.448	0.455	0.467	0.470	0.468
gap to cluster (95% CI)	$[-0.09, +0.01]$	$[-0.08, +0.01]$	$[-0.07, +0.02]$	$[-0.07, +0.02]$	$[-0.06, +0.03]$	$[-0.04, +0.04]$	$[-0.04, +0.04]$	$[-0.05, +0.04]$
Cluster reference (full / held-out basis)	0.470 / 0.468							

References

[1] Miguel Angel Alvarado Gonzalez, Michelle Bruno Hernandez, Miguel Angel Peñaloza Perez, Bruno Lopez Orozco, Jesus Tadeo Cruz Soto, and Sandra Malagon. Do repetitions matter? strengthening reliability in LLM evaluations, 2025. arXiv:2509.24086.

[2] Anthropic. How we built our multi-agent research system. Engineering blog post, Anthropic, 2025. URL https://www.anthropic.com/engineering/built-multi-agent-research-system. Orchestrator–worker system.

[3] Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. HealthBench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

[4] Mahdi Astaraki, Mohammad Arshi Saloot, Ali Shiraee Kasmaee, Hamidreza Mahyar, and Soheila Samiee. When iterative RAG beats ideal evidence: A diagnostic study in scientific multi-hop question answering, 2026. URL https://arxiv.org/abs/2601.19827.

[5] Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K. Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. arXiv:2406.18403; short paper.

[6] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. arXiv:2308.09687.

[7] A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008. doi: 10.1162/rest.90.3.414.

[8] Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? In *Advances in Neural Information Processing Systems (NeurIPS) – Datasets and Benchmarks Track*, 2025. arXiv:2503.13657; introduces the MAST taxonomy.

[9] Eric L. Charnov. Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, 9 (2):129–136, 1976.

[10] Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? A study on judgement biases, 2024. URL https://arxiv.org/abs/2402.10669. EMNLP 2024.

[11] Junjie Chen, Yuxi Dong, Haitao Li, Weihang Su, Yujia Zhou, Min Zhang, Yiqun Liu, and Qingyao Ai. Benchmarking LLM-as-a-judge for long-form output evaluation, 2026. URL https://arxiv.org/abs/2606.01629.

[12] Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more LLM calls all you need? Towards the scaling properties of compound AI systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2403.02419.

[13] Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. ReSearch: Learning to reason with search for LLMs via reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2503.19470.

[14] Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. MindSearch: Mimicking human minds elicits deep AI searcher. *arXiv preprint arXiv:2407.20183*, 2024.

[15] Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Sahel Sharifymoghaddam, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy Lin. BrowseComp-Plus: A more fair and transparent evaluation benchmark of deep-research agent, 2025. URL https://arxiv.org/abs/2508.06600.

[16] Hyeong Kyu Choi, Xiaojin Zhu, and Yixuan Li. Debate or vote: Which yields better decisions in multi-agent large language models?, 2025. URL https://arxiv.org/abs/2508.17536.

[17] Dan Cohen. Optimizing reproduction in a randomly varying environment. *Journal of Theoretical Biology*, 12(1):119–129, 1966.

[18] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. arXiv:2501.12948; DOI: https://doi.org/10.1038/s41586-025-09422-z.

[19] Simon Dennis, Michael Diamond, Rivaan Patil, Kevin Shabahang, and Hao Guo. In-context prompting obsoletes agent orchestration for procedural tasks, 2026. arXiv:2604.27891.

[20] Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*, 2025. Accepted at ICLR 2026. DOI: 10.48550/arXiv.2506.11763.

[21] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *Proceedings of the Conference on Language Modeling (COLM)*, 2024. arXiv:2404.04475.

[22] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6465–6488, 2023. doi: 10.18653/v1/2023.emnlp-main.398. ALCE benchmark; arXiv:2305.14627.

[23] Google. Try Deep Research and our new experimental model in Gemini, your AI assistant. Blog post, Google Gemini, December 2024. URL https://blog.google/products/gemini/google-gemini-deep-research/. Launched 2024-12-11 as part of Gemini 2.0.

[24] Luke Guerdan, Solon Barocas, Kenneth Holstein, Hanna Wallach, Zhiwei Steven Wu, and Alexandra Chouldechova. Validating LLM-as-a-judge systems under rating indeterminacy. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2503.05965.

[25] Nikita Gupta, Riju Chatterjee, Lukas Haas, Connie Tao, Andrew Wang, Chang Liu, Hidekazu Oiwa, Elena Gribovskaya, Jan Ackermann, John Blitzer, Sasha Goldshtein, and Dipanjan Das. DeepSearchQA: Bridging the comprehensiveness gap for deep research agents. *arXiv preprint arXiv:2601.20975*, 2026. Google DeepMind.

[26] Rajarshi Haldar and Julia Hockenmaier. Rating roulette: Self-inconsistency in LLM-as-a-judge frameworks, 2025. URL https://arxiv.org/abs/2510.27106.

[27] Rujun Han, Yanfei Chen, Zoey CuiZhu, Lesly Miculicich, Guan Sun, Yuanjun Bi, Weiming Wen, Hui Wan, Chunfeng Wen, Solène Maître, George Lee, Vishy Tirumalashetty, Emily Xue, Zizhao Zhang, Salem Haykal, Burak Gokturk, Tomas Pfister, and Chen-Yu Lee. Deep researcher with test-time diffusion. *arXiv preprint arXiv:2507.16075*, 2025.

[28] Tomas Havranek and Zuzana Irsova. Does multi-agent debate improve AI feedback on research papers?, 2026. arXiv:2607.14713.

[29] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[30] Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. arXiv:2308.00352.

[31] Chen Hu, Haikuo Du, Heng Wang, Lin Lin, Mingrui Chen, Peng Liu, Ruihang Miao, Tianchi Yue, et al. Step-DeepResearch technical report. Technical report, Step AI, 2025. arXiv:2512.20491.

[32] Yukun Huang, Leonardo F. R. Ribeiro, Momchil Hardalov, Bhuwan Dhingra, Markus Dreyer, and Venkatesh Saligrama. DeepFact: Co-evolving benchmarks and agents for deep research factuality, 2026. URL https://arxiv.org/abs/2603.05912. Introduces DeepFact-Bench and DeepFact-Eval.

[33] Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Huichi Zhou, Meng Fang, et al. Deep research agents: A systematic examination and roadmap. *arXiv preprint arXiv:2506.18096*, 2025. URL https://arxiv.org/abs/2506.18096.

[34] Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, Carl Saroufim, Corey Fry, Dror Marcus, Doron Kukliansky, Gaurav Singh Tomar, James Swirhun, Jinwei Xing, Lily Wang, Madhu Gurumurthy, Michael Aaron, Moran Ambar, Rachana Fellinger, Rui Wang, Zizhao Zhang, Sasha Goldshtein, and Dipanjan Das. The FACTS grounding leaderboard: Benchmarking LLMs' ability to ground responses to long-form input. *arXiv preprint arXiv:2501.03200*, 2025.

[35] 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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. URL https://arxiv.org/abs/2408.15232. Co-STORM; arXiv:2408.15232.

[36] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025. URL https://arxiv.org/abs/2503.09516.

[37] Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents that matter. *Transactions on Machine Learning Research*, 2025. arXiv:2407.01502; openreview.net/forum?id=Zy4uFzMviZ.

[38] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. arXiv:2405.01535.

[39] Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, Sue Hyun Park, Hyeonbin Hwang, Jinkyung Jo, Hyowon Cho, Haebin Shin, Seongyun Lee, Hanseok Oh, Noah Lee, Namgyu Ho, Se June Joo, Miyoung Ko, Yoonjoo Lee, Hyungjoo Chae, Jamin Shin, Joel Jang, Seonghyeon Ye, Bill Yuchen Lin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. The BiGGen Bench: A principled benchmark for fine-grained evaluation of language models with language models. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2025. arXiv:2406.05761.

[40] Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Yun Liu, Mark Malhotra, Paul Pu Liang, Hae Won Park, Yuzhe Yang, Xuhai Xu, Yilun Du, Shwetak Patel, Tim Althoff, Daniel McDuff, and Xin Liu. Capable language models can outgrow the benefits of collaboration. *Nature Machine Intelligence*, 2026. Originally circulated as "Towards a Science of Scaling Agent Systems", arXiv:2512.08296 (2025).

[41] Guneet Kohli. Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels, 2026. arXiv:2605.29800.

[42] Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. CoBBLEr; arXiv:2309.17012.

[43] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage, Thousand Oaks, CA, 2nd edition, 2004.

[44] Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. *arXiv preprint arXiv:2409.12941*, 2024. Introduces FRAMES; Google Research.

[45] John K. Kruschke. Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280, 2018. doi: 10.1177/2515245918771304.

[46] Krishna K. Ladha. The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, 36(3):617–634, 1992.

[47] Daniël Lakens. Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177.

[48] Shahar Levy, Nir Mazor, Lihi Shalmon, Michael Hassid, and Gabriel Stanovsky. More documents, same length: Isolating the challenge of multiple documents in RAG, 2025. URL https://arxiv.org/abs/2503.04388.

[49] Gaotang Li, Bhavana Dalvi Mishra, Zifeng Wang, Chen-Yu Lee, Tomas Pfister, et al. RubricEM: Meta-RL with rubric-guided policy decomposition beyond verifiable rewards, 2026. arXiv:2605.10899.

[50] Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv:2402.05120.

[51] Minghao Li, Ying Zeng, Zhihao Cheng, Cong Ma, and Kai Jia. ReportBench: Evaluating deep research agents via academic survey tasks, 2025. URL https://arxiv.org/abs/2508.15804.

[52] S. Yifei Li, Allen Chang, Chaitanya Malaviya, and Mark Yatskar. ResearchQA: Evaluating scholarly question answering at scale across 75 fields with survey-mined questions and rubrics. *arXiv preprint arXiv:2509.00496*, 2025.

[53] Wenjun Li, Zhi Chen, Jingru Lin, Hannan Cao, Wei Han, Sheng Liang, Zhi Zhang, Kuicai Dong, Dexun Li, Chen Zhang, and Yong Liu. Reinforcement learning foundations for deep research systems: A survey, 2025. URL https://arxiv.org/abs/2509.06733.

[54] Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. URL https://arxiv.org/abs/2501.05366. arXiv:2501.05366.

[55] Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. WebThinker: Empowering large reasoning models with deep research capability. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2504.21776.

[56] Xuefeng Li, Haoyang Zou, and Pengfei Liu. LIMR: Less is more for RL scaling. *arXiv preprint arXiv:2502.11886*, 2025.

[57] Zijian Li et al. WebWeaver: Structuring web-scale evidence with dynamic outlines for open-ended deep research. *arXiv preprint arXiv:2509.13312*, 2025. URL https://arxiv.org/abs/2509.13312.

[58] Nelson F. Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, 2023. doi: 10.18653/v1/2023.findings-emnlp.467. arXiv:2304.09848.

[59] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. arXiv:2303.16634.

[60] 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. URL https://arxiv.org/abs/2408.06292. Sakana AI.

[61] Mingyu Lu, Yushan Huang, Chris Lin, and Su-In Lee. Agents that matter: Optimizing multi-agent LLMs via removal-based attribution, 2026. arXiv:2605.27621.

[62] James G. MacKinnon and Matthew D. Webb. The wild bootstrap for few (treated) clusters. *The Econometrics Journal*, 21(2):114–135, 2018.

[63] Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2024. aclanthology.org/2024.naacl-long.167; arXiv:2309.07852.

[64] Saahil Mathur, Ryan David Rittner, Vedant Ajit Thakur, Daniel Stuart Schiff, and Tunazzina Islam. Retrieval improvements do not guarantee better answers: A study of RAG for AI policy QA, 2026. URL https://arxiv.org/abs/2603.24580.

[65] Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. arXiv:2311.12983.

[66] 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 the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. arXiv:2305.14251.

[67] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

[68] Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Kung-Hsiang Huang, Yixin Mao, and Chien-Sheng Wu. DeepTRACE: Auditing deep research AI systems for tracking reliability across citations and evidence, 2025. arXiv:2509.04499.

[69] Maksym Nechepurenko and Pavel Shuvalov. Coordination as an architectural layer for LLM-based multi-agent systems, 2026. arXiv:2605.03310.

[70] Justin D. Norman, Michael U. Rivera, and D. Alex Hughes. Reliability without validity: A systematic, large-scale evaluation of LLM-as-a-judge models across agreement, consistency, and bias, 2026. arXiv:2606.19544.

[71] Hailey Onweller, Elias Lumer, Austin Huber, Pia Ramchandani, Vamse Kumar Subbiah, and Corey Feld. Cited but not verified: Parsing and evaluating source attribution in LLM deep research agents. arXiv preprint arXiv:2605.06635, 2026.

[72] OpenAI. Introducing deep research. Blog post, OpenAI, February 2025. URL https://openai.com/index/introducing-deep-research/. Launched 2025-02-02.

[73] Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2404.13076.

[74] Liana Patel, Negar Arabzadeh, Harshit Gupta, Ankita Sundar, Ion Stoica, Matei Zaharia, and Carlos Guestrin. DeepScholar-Bench: A live benchmark and automated evaluation for generative research synthesis. *arXiv preprint arXiv:2508.20033*, 2025.

[75] Perplexity AI. Introducing Perplexity Deep Research. Blog post, Perplexity, February 2025. URL https://perplexity.ai/hub/blog/introducing-perplexity-deep-research. Launched 2025-02-14.

[76] Hoang Pham, Dong Le, and Anh Tuan Luu. GRACE: Step-level benchmark for faithful reasoning over context. *arXiv preprint arXiv:2606.16151*, 2026.

[77] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025. Center for AI Safety and Scale AI; 2,500 expert-crafted questions.

[78] Eric R. Pianka. The structure of lizard communities. *Annual Review of Ecology and Systematics*, 4: 53–74, 1973.

[79] Delip Rao, Eric Wong, and Chris Callison-Burch. Detecting and correcting reference hallucinations in commercial LLMs and deep research agents. arXiv preprint arXiv:2604.03173, 2026.

[80] Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*, 2023.

[81] Muayad Sayed Ali et al. The harness effect: How orchestration design sets the token economics of enterprise agentic AI, 2026. arXiv:2607.06906; 32 authors.

[82] Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987. doi: 10.1007/BF01068419.

[83] Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G. Finlayson, David Sontag, Tyler Murray, Sewon Min, Pradeep Dasigi, Luca Soldaini, Faeze Brahman, Wen-tau Yih, Tongshuang Wu, Luke Zettlemoyer, Yoon Kim, Hannaneh Hajishirzi, and Pang Wei Koh. DR Tulu: Reinforcement learning with evolving rubrics for deep research. *arXiv preprint arXiv:2511.19399*, 2025. Reinforcement learning with evolving rubrics (RLER) for long-form deep research.

[84] 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. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2024. arXiv:2402.14207.

[85] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. Introduces GRPO.

[86] Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, Aishwarya Balwani, Denis Peskoff, Marcos Ayestaran, Sean M. Hendryx, Brad Kenstler, and Bing Liu. ResearchRubrics: A benchmark of prompts and rubrics for evaluating deep research agents. *arXiv preprint arXiv:2511.07685*, 2025. Accepted at ICLR 2026; Scale AI.

[87] Wenxuan Shi, Haochen Tan, Chuqiao Kuang, Xiaoguang Li, Xiaozhe Ren, Chen Zhang, Hanting Chen, Yasheng Wang, Lu Hou, and Lifeng Shang. DeepDiver: Adaptive search intensity scaling via open-web reinforcement learning. *arXiv preprint arXiv:2505.24332*, 2025. URL https://arxiv.org/abs/2505.24332. NeurIPS 2025 Spotlight; Huawei Language Model Lab.

[88] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2303.11366.

[89] Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979. doi: 10.1037/0033-2909.86.2.420.

[90] 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. *arXiv preprint arXiv:2409.13740*, 2024. Introduces LitQA2; FutureHouse.

[91] Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-Searcher: Incentivizing the search capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2503.05592*, 2025. URL https://arxiv.org/abs/2503.05592.

[92] Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*, 2024.

[93] Hao Sun, Zile Qiao, Jiayan Guo, Xuanbo Fan, Yingyan Hou, Yong Jiang, Pengjun Xie, Yan Zhang, Fei Huang, and Jingren Zhou. ZeroSearch: Incentivize the search capability of LLMs without searching. *arXiv preprint arXiv:2505.04588*, 2025. URL https://arxiv.org/abs/2505.04588.

[94] Shuang Sun, Huatong Song, Yuhao Wang, Ruiyang Ren, Jinhao Jiang, Junjie Zhang, Fei Bai, Jia Deng, Wayne Xin Zhao, Zheng Liu, Lei Fang, Zhongyuan Wang, and Ji-Rong Wen. SimpleDeepSearcher: Deep information seeking via web-powered reasoning trajectory synthesis. *arXiv preprint arXiv:2505.16834*, 2025.

[95] Zhengwei Tao, Jialong Wu, Wenbiao Yin, Junkai Zhang, Baixuan Li, Haiyang Shen, Kuan Li, Liwen Zhang, Xinyu Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. WebShaper: Agentically data synthesizing via information-seeking formalization. *arXiv preprint arXiv:2507.15061*, 2025. URL https://arxiv.org/abs/2507.15061.

[96] Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, et al. Tongyi DeepResearch technical report. *arXiv preprint arXiv:2510.24701*, 2025. URL https://arxiv.org/abs/2510.24701. Alibaba Tongyi Lab.

[97] Dat Tran and Douwe Kiela. Single-agent LLMs outperform multi-agent systems on multi-hop reasoning under equal thinking token budgets. *arXiv preprint arXiv:2604.02460*, 2026.

[98] Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

[99] Yuxuan Wan, Tianqing Fang, Zaitang Li, Yintong Huo, Wenxuan Wang, Haitao Mi, Dong Yu, and Michael R. Lyu. Inference-time scaling of verification: Self-evolving deep research agents via test-time rubric-guided verification, 2026. URL https://arxiv.org/abs/2601.15808.

[100] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

[101] Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. arXiv:2305.04091.

[102] Leyao Wang, Yanan He, Peng Chen, Asaf Yehudai, Yixin Liu, Rex Ying, Michal Shmueli-Scheuer, and Arman Cohan. Time to REFLECT: Can we trust LLM judges for evidence-based research agents?, 2026. URL https://arxiv.org/abs/2605.19196.

[103] Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2402.18272.

[104] Yongjie Wang, Xinyue Zhang, Kunhong Yao, Zhiwei Zeng, Kaisong Song, Jun Lin, and Zhiqi Shen. Search-time contamination in deep research agents: Measuring performance inflation in public benchmark evaluation, 2026. URL https://arxiv.org/abs/2606.05241.

[105] Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. BrowseComp: A simple yet challenging benchmark for browsing agents. Technical report, OpenAI, 2025. arXiv:2504.12516; Blog + paper: https://openai.com/index/browsecomp/; code: openai/simple-evals.

[106] Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. Introduces LongFact and SAFE; arXiv:2403.18802.

[107] Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Gang Fu, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. WebDancer: Towards autonomous information seeking agency. *arXiv preprint arXiv:2505.22648*, 2025. URL https://arxiv.org/abs/2505.22648.

[108] Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. WebWalker: Benchmarking LLMs in web traversal. *arXiv preprint arXiv:2501.07572*, 2025.

[109] Minghao Wu and Alham Fikri Aji. Style over substance: Evaluation biases for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, 2025. arXiv:2307.03025.

[110] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

[111] Jian Xie, Huan Sun, et al. QUEST: Training frontier deep research agents with fully synthetic tasks, 2026. arXiv:2605.24218; OSU NLP Group, 19 authors.

[112] Yumo Xu, Peng Qi, Jifan Chen, Kunlun Liu, Rujun Han, Lan Liu, Bonan Min, Vittorio Castelli, Arshit Gupta, and Zhiguo Wang. CiteEval: Principle-driven citation evaluation for source attribution, 2025. URL https://arxiv.org/abs/2506.01829.

[113] 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. URL https://arxiv.org/abs/2504.08066. Sakana AI.

[114] Zongyou Yang, Yinghan Hou, and Xiaokun Yang. When the judge changes, so does the measurement: Auditing LLM-as-judge reliability, 2026. arXiv:2607.08535.

[115] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.10601.

[116] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.03629.

[117] Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice? Quantifying biases in LLM-as-a-judge, 2024. URL https://arxiv.org/abs/2410.02736.

[118] Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024. arXiv:2307.10928; Spotlight.

[119] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. LIMO: Less is more for reasoning. In *Conference on Language Modeling (COLM)*, 2025. arXiv:2502.03387.

[120] You.com. Introducing ARI: The first professional-grade research agent for business. Blog post, You.com, February 2025. URL https://you.com/resources/introducing-ari-the-first-professional-grade-research-agent-for-business. Launched 2025-02-27.

[121] Ming Zhang, Jiabao Zhuang, Wenqing Jing, Kexin Tan, Ziyu Kong, Jingyi Deng, Yujiong Shen, Yuhui Wang, Zhenghao Xiang, Qiyuan Peng, Yuhang Zhao, Ning Luo, Renzhe Zheng, Jiahui Lin, Mingqi Wu, Long Ma, Zhangyue Yin, Shihan Dou, Maxm Pan, Tao Gui, Qi Zhang, and Xuanjing Huang. Can deep research agents retrieve and organize? evaluating the synthesis gap with expert taxonomies, 2026. arXiv:2601.12369.

[122] Wenlin Zhang, Xiaopeng Li, Yingyi Zhang, Pengyue Jia, Yichao Wang, Huifeng Guo, Yong Liu, and Xiangyu Zhao. Deep Research: A survey of autonomous research agents. *arXiv preprint arXiv:2508.12752*, 2025. URL https://arxiv.org/abs/2508.12752.

[123] Zhen Zhang, Liangcai Su, Zhuo Chen, Xiang Lin, Haotian Xu, Simon Shaolei Du, Kaiyu Yang, Bo An, Lidong Bing, and Xinyu Wang. Argus: Evidence assembly for scalable deep research agents. *arXiv preprint arXiv:2605.16217*, 2026.

[124] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.05685.

[125] Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. DeepResearcher: Scaling deep research via reinforcement learning in real-world environments. *arXiv preprint arXiv:2504.03160*, 2025.

[126] Joey Zhong, Hao Zhang, Clare Southern, Jeremy Yang, Thomas Wang, Kate Jung, Shu Zhang, Denis Yarats, Johnny Ho, and Jerry Ma. DRACO: a cross-domain benchmark for deep research accuracy, completeness, and objectivity. *arXiv preprint arXiv:2602.11685*, 2026.

[127] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. arXiv:2205.10625.

[128] Junze Zhu, Weihao Chen, Xuanwang Zhang, Zhen Wu, and Xinyu Dai. Recognize your orchestrator: An entropy dynamics perspective for LLM multi-agent systems, 2026. arXiv:2606.01351.

[129] Pengyu Zhu, Lijun Li, Longju Yang, and Sen Su. Is deep research reliable? misleading knowledge induces false conclusions, 2026. arXiv:2607.20891.