IS IT THE READ OR THE INTERRUPTION? DECOMPOSING READ-GATE GAINS IN AGENTIC RAG

Anonymous authors
Paper under double-blind review

ABSTRACT

Agentic retrieval systems can finalize without issuing a read action on retrieved documents. Requiring a read can improve accuracy, but the gain may reflect retrieved evidence, interruption, or changes in answering behaviour. We investigate this distinction by branching each trajectory at its first attempt to finalize without reading, holding the preceding interaction fixed. For gpt-5-mini at minimal reasoning effort, we compare read demands with information-free tool-call demands in pre-registered factorial studies on 1,800 Wikipedia and 800 medical questions. The information-free control often returns acknowledgements or questions instead of answers. Adding the same explicit answer obligation to both demands reduces the read/noop accuracy gap on Wikipedia from 3.3 to 1.2 percentage points; both the reduction and residual survive Holm correction. On the medical corpus, no additional read benefit is detected. There, against no gate, the read demand without the obligation loses 3.0 points, whereas an obligated noop demand gains 3.5 points. A forecast frozen before the medical interventions overestimates the read demand's payoff by 5.3 points. On Qwen2.5-7B, non-answer responses are rare and no obligation benefit is detected. These findings show that accuracy gains from enforced reading alone do not isolate the contribution of retrieved evidence. Evaluations must also control how interventions instruct agents to answer and terminate.

1 INTRODUCTION

Tool-using retrieval agents answer questions by interleaving discrete actions — `search`, `read`, `final` — under a controller model, the large language model (LLM) that chooses the next action. This makes a behaviour observable that end-to-end retrieval-augmented generation (RAG) hides: the agent can commit to an answer *without* issuing a single explicit `read` action (search returns ranked snippets, `read` one chunk; a snippet can already contain that chunk, §2.1). We call the first such attempt in a trajectory a *trigger*. Triggers range from rare to common across controllers: on three multi-hop benchmarks gpt-5-mini at minimal reasoning effort emits one on 22–66% of questions (Appendix G), and every one of the seven controllers we measure does so at some rate (§6). A simple runtime rule — reject the premature final and demand a read — raises that controller's paired accuracy by about five points (§4). Zero reads need not mean a wrong answer: the snippets already expose evidence, and the prompt permits some snippet-only answers.

Why the rule helps is less clear than that it helps. Rejecting a final changes several things at once: the agent gets *another decision*, must *commit to an action*, receives *that action's content*, and is told *how it may end*; only the content is the evidence the gate's name promises. Gate-versus-no-gate comparisons confound these; the closest prior work — budget forcing (Muennighoff et al., 2025), self-correction (Shinn et al., 2023; Gou et al., 2024), search-reasoning loops (Li et al., 2025), test-time intervention (Yang et al., 2026) — optimizes an intervention, and replay-based attribution (§7) does not ask whether the control was obliged to answer. We hold the trajectory byte-identical up to the trigger and vary only the intervention injected there: a **matched ladder** of a structured delay, an information-free tool demand, a one-shot read demand and the persistent Read-Gate, with any-tool, decoy-read and restate side controls (§2.2).

The ladder answered part of the question and exposed a confound. The read demand beats the information-free action on two question-disjoint panels (+3.8 and +3.1pp), but its edge over a structured delay does not replicate, and the control often does not answer at all. Asked *"Castle*

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

Hill is a plantation near what primary residence of Thomas Jefferson?", the ungated agent answers *Monticello* from snippets and is right; demanding an information-free tool call first yields *"Understood — noop called."*, graded wrong; demanding a read yields *Monticello* again (Figure 1). The read/noop gap is partly the control breaking answers, not the read repairing them. We therefore add to every intervention a two-sentence *answer obligation* F — finalize with an answer, not an acknowledgement, plan or question — and run a pre-registered 2×2 factorial (read or noop demand, with or without F) plus obligated delay and decoy-read arms on the same 1,800 questions and on 800 new medical questions (§5).

Findings. *(M1) Matching the answer obligation shrinks the read/noop gap.* On the 1,800 Wikipedia questions the gap falls from $+3.28$ to $+1.22$pp; both pre-registered primaries (the residual gap and the interaction) pass Holm correction. On the medical corpus the residual is -0.38pp and the interaction $+2.12$pp, with an interval that includes zero. *(M2) No advantage of the unobligated read demand over an obligated information-free interruption is detected.* The obligated information-free action gains $+5.78$ and $+3.50$pp over no gate on the two corpora (exploratory contrasts); the unobligated read demand gains $+5.28$pp on Wikipedia but *loses* 3.00pp on medical questions, and an obligated structured delay has the highest medical accuracy. On a second controller that rarely hands back non-answers, no benefit from F was detected. *(M3) Transported recovery and disruption rates did not forecast the payoff.* A payoff forecast frozen before the medical intervention arms ran, transporting the read demand's Wikipedia recovery and disruption rates (the share of wrong skipped answers it fixes and of right ones it breaks) to the medical trigger and wrong-share rates, predicted $+2.33$pp; the observed effect was -3.00pp: disruption did not transport.

Contributions. (i) A comparison method: a prefix-matched factorial that crosses what an intervention demands with how the agent may end. (ii) The finding that an unmatched obligation inflates read-demand comparisons: matching it shrinks the read/noop gap to a small residual on Wikipedia, and on a medical corpus no residual is detected. (iii) Boundaries and resources: a frozen deployment forecast that failed on a new corpus, a second controller with no detected obligation effect, and an audited release of over 30,000 matched trajectories with manifests, pre-registration records and table-regenerating scripts.

2 Matched interventions at a premature final

2.1 The observable behaviour and the trigger

The observable behaviour. A trajectory is a sequence of discrete tool actions — `search`, `read`, `final` — together with the retrieved snippets and read passages behind each call. Our object of study is a *procedural* property of this sequence, not an inference about latent model intent: a *zero-read finalization* is an answer emitted as plain assistant content or as a `final` tool call with a successful full-chunk `read` count of zero (retrieved snippets may have been seen). Zero-read denotes the absence of a read *action*, not of exposure to chunk text: a search snippet can contain the whole chunk (§3, §5). This strict, action-count definition is read directly from the trace, needs no heuristic, and is the only prevalence signal the main text uses; heuristic subtypes and the entity-matcher validation are appendix diagnostics (Appendix J.4).

The trigger. We call the first `final` attempt on a trajectory whose successful-read count is still zero the *trigger*, in either finalization channel during the ordinary agent loop; a final forced by the harness at the loop or token limit is not an interceptable trigger, and a trigger need not follow a search (55 of 685 Wikipedia and 40 of 458 medical triggers preceded any search call). The trigger is what the ladder of §2.2 intervenes on and what distinguishes trajectories that ever attempt a premature final from those that never do; every arm is byte-identical up to this moment. Triggering is read off the trace, independently of correctness.

2.2 Branch-point design and arms

Every arm observes the same agent, retrieval index and ordered question panel and differs from the no-intervention baseline only from the trigger (§2.1); trajectories that never trigger are byte-identical across arms (audited), so a paired difference between two arms is the effect of what happens *at* the

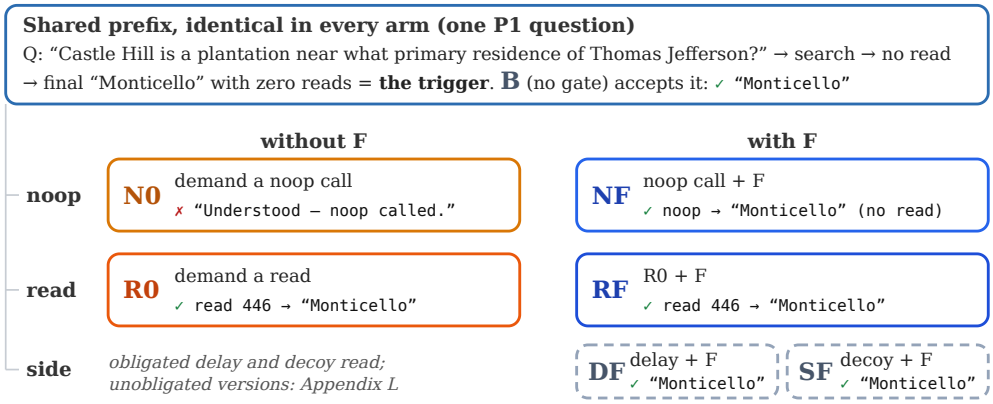

Figure 1: **Branch-point design on one real question.** Every arm shares the pre-trigger trajectory and continues from the first zero-read final; boxes show each arm's frozen answer and judge label for one P1 question (this question's labels, not rates). F is the two-sentence answer obligation: answer the original question, not with an acknowledgement, plan or question. The original five-rung ladder is drawn in Appendix Figure S3.

branch point (Figure 1). Every contrast is an intention-to-treat (ITT) difference: judged accuracy averaged over all questions of a panel, trigger or not, on identical question IDs.

Arms. The factorial of §5 uses B (no gate: accept the premature final), N0 (reject the final once and demand a call to an information-free noop tool) and R0 (reject once and demand one read); NF and RF add the same two-sentence answer obligation F, and DF and SF are an obligated structured delay and an obligated decoy read. Each arm is an *intervention bundle* — message, demanded tool, tool result — so a contrast attributes gain to what the bundle adds, not to one property, and no arm forbids further search or read after the trigger. The two primaries ask: with the obligation matched, does *requesting* a read rather than a noop still help (RF−NF: an ITT effect of the request, since compliance is not forced and both arms may read afterwards)? and how much does the read/noop gap change once F is added ($(\mathrm{R0}-\mathrm{N0})-(\mathrm{RF}-\mathrm{NF})$)?

The original ladder (§4). Before the factorial, the same branch point carried a five-rung ladder (Appendix Figure S3): NO-GATE (=B); a structured DELAY, a content-free message saying the next reply counts as the final answer; NOOP (=N0); READ-ONCE, the same read demand as R0, its chunk pinned by hash; and the persistent READ, which re-arms at every zero-read final and abstains at the loop or token limit (the earlier $n = 1{,}000$ grid's Read-Gate, Appendix J.1). Post-hoc controls probe its terms: ANY-TOOL demands any ordinary tool, SHAM-READ returns a hash-selected, not-yet-shown decoy chunk under the read demand, and RESTATE asks for a restated answer without requiring a tool call. Persistent enforcement added little over one shot (§4), so the factorial branches on the one-shot demands. Verbatim messages: Appendix A.1.

Caching and the interruption control. Every LLM call is served through a deterministic response cache keyed on the full request, so replaying the pre-trigger context unchanged returns the cached premature final and the structured delay is the smallest interruption that differs from NO-GATE (the noise study P3 of §5 bypasses the cache after the branch).

Pre-specification and audits. Each experiment is frozen in a hashed manifest before execution; the analysis audits matched triggers, identical prefixes, one-shot firing, post-trigger-only tool exposure and pinned message bytes (Appendix A.3).

3 EXPERIMENTAL SETUP

Agent and tools. The agent is a ReAct-style loop (Yao et al., 2023) — reasoning interleaved with tool calls — over native function calling with three tools: `search` (hybrid retrieval over the benchmark chunk corpus: BM25 (Robertson & Zaragoza, 2009) and Qwen3-Embedding-0.6B (Zhang et al., 2025) dense scores fused by reciprocal-rank fusion (Cormack et al., 2009) with k_{RRF}=60, returning the top-K chunk identifiers, each with a sentence snippet: the leading five sentence segments for BM25 candidates, similarity-selected sentences for dense-only candidates (Appendix K.3); K=5 by default), `read` (returning the full text of one chunk), and `final`. Every cell in the primary ladder and its controls uses `gpt-5-mini-2025-08-07` at minimal reasoning effort (OpenAI, 2026a) (deterministic response cache, temperature 0 where the API accepts it; at most 10 loops and a 128k-token budget); the prevalence map in §6 adds gpt-4o-mini, gpt-5.6-luna, Gemini 2.5 Flash, Qwen2.5 3B/7B and Qwen3.7-plus cells with their own settings; cell definitions and denominators are in Appendix Table S14 and full settings in the released manifests.

Data and panels. We use HotpotQA (Yang et al., 2018), 2WikiMultihop (Ho et al., 2020) and MuSiQue (Trivedi et al., 2022) with the per-benchmark chunk corpora and question pools released with LinearRAG (Zhuang et al., 2025a;b) (1,000 questions per benchmark); "Wikipedia" denotes their union. The *ladder panel* is 200 questions per benchmark drawn by a fixed seeded shuffle (600 total); the *replication panel* is the next 400 questions per benchmark in the same seeded order, disjoint from the ladder panel (1,200 total); all five arms ran on both panels. The replication panel was reserved for the pre-specified core replication; the post-hoc controls (any-tool, restate, sham-read), designed after the core results were known, ran on both panels' 1,800 IDs. All arms see identical ordered question IDs, frozen with hashes in the manifest. Both panels lie within an earlier $n = 1{,}000$ no-gate/persistent-gate grid on the same pools (Appendix J.1). *Prevalence cells* (§6) are no-intervention arms on the ladder panel's seeded first-200 IDs per benchmark under the same prompts, tools, loop cap and judge; inherited cells keep their original size (Appendix G.3).

Format studies. P1 replays the prefixes of all 1,800 panel questions and branches six new arms at the trigger (685 trigger rows); P2 runs all seven arms on 800 new GraphRAG-Bench medical questions (Xiang et al., 2025) (458 trigger rows); P3 redraws 60 trigger questions twice more under the six intervention arms (Appendix K).

Judging. Answers are judged against gold answers by `gpt-5-mini`. Cross-family agreement with a Gemini 2.5 Pro judge is Cohen's $\kappa = 0.831$ on the 678 answers of this paper that enter a discordant outcome of a headline contrast; substituting the stricter Gemini labels moves each pooled ladder contrast by at most 0.45pp and flips no sign, a targeted check rather than a bound (Appendix J.3). The format studies use the same judge with a frozen completion amendment — verdicts left malformed at the 300-token cap, scored wrong by the original parser, were re-judged under the same prompt at 1,500 tokens (Appendix K) — and an AI-rater audit (Appendix K.2).

Estimands and inference. Each contrast is a paired difference in judged accuracy over all questions of a panel (ITT); *fixes*/*breaks* are the wrong→right / right→wrong discordant pairs, and the main text reports point estimates with these counts. Ladder 95% confidence intervals (CIs) and original primary families are in Appendices B and B.1. Appendix Tables S22 and S25 report format/Qwen intervals and tests: exact paired McNemar for the residual and question-level Wald for the interaction, with Holm within each study. No-intervention accuracy is 59.7% on the panel and 57.1% on the replication under the original judge parsing (58.8% pooled under the completed judgments of §5).

Cost and provenance. Every cell runs under a frozen manifest with per-arm ceilings; the two panels comprise 15,000 ladder and side-control trajectories over matched IDs (about $33 of new API spend) and the three format studies 17,120 more ($64 at list price; Appendices F and K).

4 WHAT THE ORIGINAL LADDER FOUND

Table 1 reports the ladder's paired ITT differences on the 600-question panel and the question-disjoint 1,200-question replication (intervals, p-values, primary families and remaining rungs: Appendices B–B.1 and L).

Table 1: **The original ladder: paired ITT contrasts.** Differences in judged accuracy (pp; fixes/breaks in parentheses) on the 600-question panel, the disjoint 1,200-question replication and pooled. Bold: the pre-specified primary of both manifests. Remaining rungs, side controls and intervals: Appendix Table S1.

Contrast	Panel	Replication	Pooled
READ-ONCE − NOOP (read vs. information-free action; primary)	**+3.83** (28/5)	**+3.08** (60/23)	**+3.33**
READ − NO-GATE (persistent gate)	+4.83 (34/5)	+5.92 (86/15)	+5.56
DELAY − NO-GATE (structured interruption)	+1.17 (16/9)	+5.08 (74/13)	+3.78
READ-ONCE − DELAY (read demand over interruption)	+3.33 (29/9)	+0.25 (40/37)	+1.28
READ-ONCE − SHAM-READ (requested vs. decoy chunk; post hoc)	+2.50 (21/6)	+0.42 (27/22)	+1.11

The read demand beats an information-free action (confirmatory). The pre-specified contrast — a one-shot read demand against a demand to call an information-free noop tool — is +3.83pp on the panel (28 fixes vs. 5 breaks; 95% CI [+2.00, +5.67]) and +3.08pp on the replication (60 vs. 23; [+1.67, +4.58]), a primary of both manifests that passes each experiment's pre-specified test. All variation lives on the 197/600 and 488/1,200 trigger rows, where the contrast is +11.68 and +7.58pp. Persistent re-arming adds only +0.33 and +0.58pp; only the later pooled family supports a repetition benefit (adjusted $p = 0.023$, Table S2).

Its edge over a structured delay does not replicate (boundary). A content-free message that rejects the final once and says the next reply counts as the answer gains +1.17pp on the panel but +5.08pp on replication; READ-ONCE−DELAY falls from +3.33 to +0.25pp ([−1.17, +1.75]), failing the full-ladder replication's pre-specified test (per-benchmark terms: Appendix C). A restate control recovers +3.28 of the delay's +3.78pp pooled (rejection framing is a plausible carrier; Appendix D). The requested chunk's increment over a decoy chunk is +2.50pp on the panel but +0.42pp on replication ([−0.75, +1.58]), so the requested content is not established as the carrier either.

4.1 THE CONTROL DOES NOT ALWAYS ANSWER

Audit lesson. An open-ended delay control measured format collapse: rejected with *"Use the next decision step however you judge appropriate"*, the agent *handed back* — replied with an acknowledgement, plan or question instead of an answer — on 70.6% of 197 trigger rows, graded wrong (Appendix I.2). The structured delay reduced that arm's hand-backs to 4.1% but not the others': the frozen detector flags 12.7%/11.3% of NOOP trigger answers (panel/replication) against 5.1%/5.3% under READ-ONCE, and 28 of the pooled net 60 read-over-noop improvements fall in pairs where either answer is flagged (Appendix Table S9 and Appendix I.3). This partition is descriptive, not a mediated share of the gate's gain; §5 estimates what adding the obligation changes.

5 MATCHING THE ANSWER OBLIGATION

Design. The answer obligation F is a two-sentence instruction appended once to the gate message at the first trigger: *"When finalizing after any required action, answer the original question. Do not return an acknowledgement, a procedural plan, or a question to the user as the final answer."* Nothing before the trigger changes, and the noop result string is unchanged. Seven arms share each question's byte-identical prefix (§2.2): B, N0/R0 without F, NF/RF with F, and the obligated delay and decoy read DF/SF. F obliges an answer, not a read: no arm forbids reading, and NF/DF sometimes read voluntarily after the trigger (rates below). P1 reruns the 1,800 panel questions; P2 draws 800 new GraphRAG-Bench medical questions (§3). Two primaries were frozen per study and Holm-corrected within it: the read residual under a matched obligation (RF−NF) and the change in the read/noop gap after adding F, (R0 − N0) − (RF − NF); seven secondaries are pointwise, and no row is excluded on hand-back, compliance or outcome.

M1: matching the answer obligation shrinks the read/noop gap (pre-registered). On P1 (Table 2) the read/noop gap without the obligation is +3.28pp (new continuations, completed judgments; the ladder's own run gave +3.33, Table 1); with the obligation it is +1.22pp ([+0.17, +2.28]; 58 fixes vs. 36 breaks, Holm $p = 0.030$), and the interaction is +2.06pp ([+0.56, +3.56], Holm $p = 0.015$).

Table 2: **Matching the answer obligation: pre-registered contrasts.** Paired ITT differences in judged accuracy (pp; fixes/breaks in parentheses) on 1,800 Wikipedia questions (P1) and 800 new medical questions (P2). The last block was added after an authorized interim look. Intervals for every row and the Holm-corrected p of the primaries: Appendix Table S22.

Contrast	interpretation	P1 Wikipedia (n = 1,800)	P2 Medical (n = 800)
Pre-registered primaries (Holm over two tests per study)			
RF−NF	read residual under F	+1.22 (58/36)	−0.38 (30/33)
(R0−N0)−(RF−NF)	gap removed by matching F	+2.06	+2.12
Pre-registered secondaries (no test)			
RF−DF	read vs. delay under F	+0.94 (53/36)	−1.50 (18/30)
RF−SF	requested vs. decoy under F	+0.67 (50/38)	+0.12 (25/24)
R0−N0	read/noop gap without F	+3.28 (90/31)	+1.75 (71/57)
R0−B	read demand alone	+5.28 (109/14)	−3.00 (36/60)
RF−B	read demand with F	+7.00 (133/7)	+3.12 (42/17)
NF−N0	F added to noop demand	+3.78 (87/19)	+8.25 (88/22)
RF−R0	F added to read demand	+1.72 (66/35)	+6.12 (73/24)
Post-peek exploratory (descriptive)			
NF−B	obligated noop vs. no gate	+5.78 (110/6)	+3.50 (50/22)
SF−NF	decoy read vs. noop under F	+0.56 (51/41)	−0.50 (24/28)

On the medical corpus the residual is −0.38pp ($[-2.25, +1.50]$) and the interaction +2.12pp with an interval that includes zero at $n = 800$. Neither the delay nor the decoy read is separated from the obligated read demand on either corpus (Table 2).

Table 3: **Every arm against no gate.** Judged accuracy over all questions (%), its exact paired difference from B (pp), and the share of trigger rows (685 P1, 458 P2) whose final answer the frozen first-200-character detector flags as a hand-back. Point estimates; intervals for the pre-registered contrasts in Appendix Table S22.

		P1 Wikipedia (n = 1,800)			P2 Medical (n = 800)		
Arm	intervention at the trigger	acc. (all)	−B	flagged (trig.)	acc. (all)	−B	flagged (trig.)
B	no gate	58.8	—	0.1%	70.4	—	0.0%
N0	noop demand	60.8	+2.00	13.3%	65.6	−4.75	24.5%
NF	noop demand + F	64.6	+5.78	0.4%	73.9	+3.50	0.0%
R0	read demand	64.1	+5.28	6.1%	67.4	−3.00	3.9%
RF	read demand + F	65.8	+7.00	0.1%	73.5	+3.12	0.0%
DF	structured delay + F	64.9	+6.06	0.3%	75.0	+4.62	0.0%
SF	decoy read + F	65.2	+6.33	0.3%	73.4	+3.00	0.0%

M2: no advantage of the unobligated read demand over an obligated interruption is detected. Adding F (Table 3, Figure 2) to the noop demand is worth +3.78 and +8.25pp, to the read demand +1.72 and +6.12pp. The obligated information-free action gains +5.78pp on Wikipedia and +3.50pp on medical questions over no gate — close to the read demand's +5.28pp on Wikipedia and above its −3.00pp on medical questions, where the obligated structured delay (DF, +4.62pp) has the highest observed accuracy and N0 itself loses 4.75pp. Directly, NF−R0 is +0.50pp $[-0.61, +1.56]$ and +6.50pp $[+4.12, +9.00]$ (post hoc). With F, the detector's flag rate collapses: it flags 13.3% and 24.5% of N0 trigger answers but 0.4% and 0.0% of NF answers, and 31 of the 87 Wikipedia fixes and 65 of the 88 medical fixes from N0 to NF are questions whose N0 answer was a flagged hand-back (Appendix Table S24); this is a post-treatment description, not a causal share. Are those converted answers right? Two blinded AI raters under author-approved rules rate the NF answers on those questions correct on 30 of 31 Wikipedia and 64 of 65 medical items, and agree with the judge on 93.9% of 912 audited rows, the judge being the stricter party on the medical F arms (Appendix K.2). The read arm also changes under F: of the 49 net medical improvements of RF over R0, 43 occur on questions B answered correctly; separately, rows flagged as hand-backs under R0 contribute six (Appendix Table S27). These overlapping descriptive partitions do not identify a mechanism.

Reading the residual. The residual +1.22pp on Wikipedia survives Holm correction but is small: about a third of the unmatched gap. Two caveats limit what it measures. First, the control is not read-free: after the trigger NF read voluntarily on 13.4%/8.5% of trigger rows and DF on 11.7%/34.7% (Wikipedia/medical), against 88.8%/79.0% under R0 and 90.4%/83.2% under RF, so the residual compares a demanded read with an occasional voluntary one. Second, on Wikipedia the demanded read often repeats text a pre-trigger search result already contained (238 of the 608 R0 rows with a read, 35% of triggers; Appendix Table S26), so the contrast is not purely one of new evidence. Three draws of a 60-question trigger sample give RF − NF of −5.0, +5.0, +5.0pp (Appendix K), no noise bound for the full ITT. We do not claim reading is unnecessary: the interval excludes zero and equivalence was never tested.

A second controller. On Qwen2.5-7B-Instruct, whose ladder core showed no gate gain (§6), we branched the four factorial arms at its frozen prefixes (562 eligible questions, 136 triggers; Appendix K.1). Hand-backs were rare ($\leq$ 1 flagged trigger answer per arm) and we detected no benefit from F: NF−N0 is −0.19pp $[-0.94, +0.56]$ (−0.18pp with errors scored wrong). The pre-specified complete-case analysis (533 questions, 107 triggers, after context-limit errors concentrated in the read arms) gives RF−NF +0.94pp $[-0.38, +2.25]$; scoring errors as incorrect on all 562 gives +0.18pp $[-1.42, +1.60]$; neither primary passes (Holm $p \geq 0.53$). It bounds the obligation effect's generality without identifying its cause.

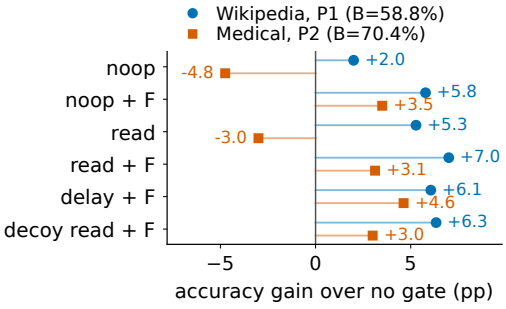

Figure 2: **Every intervention against no gate.** Paired ITT gains (pp) over B (Table 3; intervals: Appendix Table S22). Arms with F cluster.

Disclosure. Each study's analysis plan was frozen in its manifest before collection; the judgment-completion amendment (Appendix K) was frozen during P1 collection before any contrast was inspected, and one authorized interim look at 688 P1 questions followed, after which Table 2's last block was added. Scoring the malformed verdicts as wrong instead (the original parsing) moves P1 contrasts by at most 0.2pp and medical ones by up to 1.75pp. No primary decision changes, but the medical RF−NF point estimate changes sign and the medical R0−B interval excludes zero only under the completed parsing (Appendix Table S23). The medical forecast of §6 was frozen after the medical B arm and before any medical intervention arm ran.

6 Where a forced read pays

The medical loss raises a deployment question: can a gate's benefit be forecast from how often an agent skips reading and how often those skipped answers are wrong? We take prevalence (opportunities), wrong-share (what can be recovered) and pilot recovery/disruption rates in turn.

6.1 Prevalence across controllers and domains

Prevalence spans two orders of magnitude (boundary). Figure 3 plots the share of no-intervention trajectories that emit a final before any successful read over 22 controller×dataset cells of ten configurations of seven controllers, from Qwen2.5-3B at 80.7% and gpt-5-mini at minimal effort at 37.9% (55.5% on the medical scout) down to Qwen3.7-plus at 0.5–1.5%. Configurations differ in prompt, effort and tool policy, so this is not a size comparison; high prevalence alone establishes no benefit.

Prevalence depends on corpus and effort. For gpt-5-mini at minimal effort it ranges from 21.8% (HotpotQA) to 65.5% (MuSiQue), and from 8.5% on the novel scout to 55.5% on the medical one (Appendix Table S14); on the 600 panel questions, raising effort to medium cuts interceptable triggers from 197 to 20 (Appendix H.1). A pilot must measure prevalence on the target corpus and configuration.

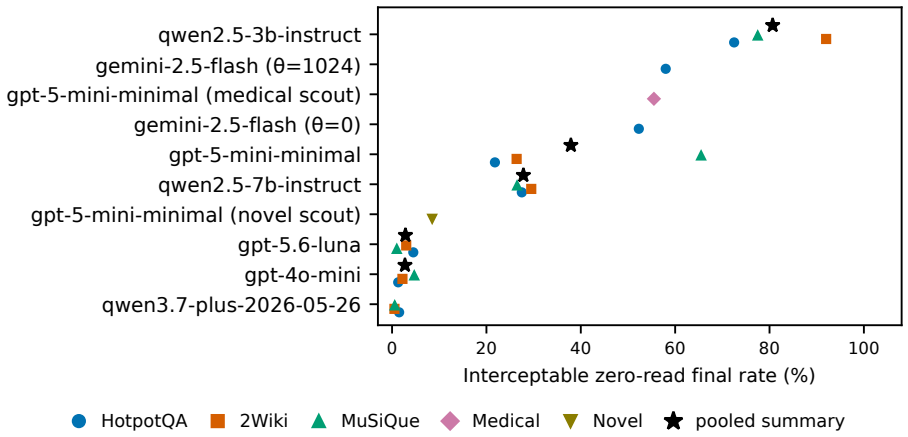

Figure 3: **Interceptable zero-read prevalence.** Share of no-intervention trajectories that finalize before any successful read (loop-cap forced finals excluded), by controller configuration; markers are the 22 controller×dataset cells, stars pooled summaries. *Novel*: a 200-question GraphRAG-Bench scout (Appendix G.3); θ: thinking budget. Counts and intervals: Appendix Table S14.

Table 4: **A frozen forecast of the read demand's payoff on a new corpus.** $\Delta = p\,[wr - (1-w)d]$ with (r, d) from the P1 R0/B pairs and (p, w) from the medical B arm, frozen before any medical intervention ran; p over all questions, w, r, d within triggers; intervals in the text.

	p (trigger)	w (wrong)	r (recovery)	d (disruption)	R0−B (pp)
P1 Wikipedia, observed	38.1%	67.0%	23.7%	6.2%	+5.28
Medical, forecast (P1 r, d; Medical p, w)	57.2%	34.3%	23.7%	6.2%	+2.33
Medical, observed	57.2%	34.3%	22.9%	19.9%	−3.00
Forecast error (observed − forecast)					−5.33

6.2 Recovery, disruption and observed payoff

On trigger rows the paired effect of a gate decomposes exactly as $\Delta_{\text{trig}} = w\,r - (1-w)\,d$ and $\Delta = p\,\Delta_{\text{trig}}$, where p is trigger prevalence, w the *wrong-share* (triggers whose skipped answer was wrong), r the recovery rate on those and d the disruption rate on triggers already right. Logs identify p, labels w, and only a paired intervention arm identifies (r, d).

A positive screen need not establish pilot benefit (boundary). For gpt-5-mini at minimal effort on the 600-question panel, $w = 67.0\%$, $r = 24.2\%$, $d = 7.7\%$ and $\Delta_{\text{trig}} = +13.7$pp $[8.1, 19.3]$ (ladder panel; P1's 1,800-question rerun gives 23.7%/6.2%, the values Table 4 transports). Our heuristic screen (prevalence $\geq 15\%$, wrong-share $\geq 40\%$) rejected the medical scout (55.5%, $w = 31.5\%$; Appendix G.3). A same-harness Qwen2.5-7B three-arm core (NO-GATE/NOOP/READ-ONCE on the panel IDs, 541 valid pairs, the population the factorial of §5 later revisits with 562 valid B runs; Appendix H.2) clears both thresholds ($w = 51.8\%$ on 114 triggers) and still shows no observed benefit: $r = 6.8\%$ against $d = 9.1\%$, $\Delta_{\text{trig}} = -0.9$pp $[-6.1, 4.4]$ (Appendix Table S15). The interval permits a positive effect; only the pilot-measured (r, d) separate the two controllers.

A frozen forecast on a new corpus. The medical study of §5 tested prospectively whether Wikipedia recovery and disruption rates transport, not the identity itself. Before any medical intervention arm ran we froze a forecast of the one-shot read demand's (R0) payoff there — (r, d) from the Wikipedia R0/B pairs, (p, w) from the 800 medical B trajectories — of +2.33pp ($[+0.73, +3.90]$ frozen; joint bootstrap $[+0.70, +3.90]$, Table 4). The observed R0−B was −3.00pp $[-5.38, -0.62]$, an error of −5.33pp $[-7.92, -2.75]$ (rates in Table 4). Prevalence and wrong-share were measured on the target; of the transported rates, recovery point estimates were similar (23.7% vs. 22.9%) and disruption was not: d rose from 6.2% to 19.9% because already-correct zero-read medical answers

were more often derailed by the unobligated read. Asked which tumour size is T1 in AJCC staging, the ungated agent answers *2 cm or less* and is right; the demanded read returns a bone-cancer chunk and the agent re-anchors to *8 cm or smaller for bone tumours*, graded wrong. With F the corpus gap in disruption narrows ($d = 3.1\%$ vs. 5.6%, post hoc; Appendix Table S27). The corpus had failed the screen ($w = 34.3\%$), yet the transported-rate forecast predicted a gain: the transport of (r, d) failed, not the screen.

7 Related work

Runtime enforcement and replay. AgentSpec specifies trigger–predicate–enforcement rules (Wang et al., 2026); Evidence-Carrying Termination checks typed evidence certificates before authorizing completion (Liu, 2026). Causal Agent Replay resamples or replaces steps to estimate outcome changes (Shah, 2026), and REFLECT uses a targeted replay's outcome to refine error attribution (Lin et al., 2026). We ask which read-gate contrasts survive interruption, action, content and answer-obligation controls; the ladder's telescoping sum is accounting, not attribution (Appendix L.1).

Agentic RAG and retrieval control. RAG conditions generation on retrieved evidence (Lewis et al., 2020). WebGPT, ReAct, Toolformer and ToolLLM interleave generation with tool calls (Nakano et al., 2021; Yao et al., 2023; Schick et al., 2023; Qin et al., 2024), while IRCoT, FLARE and A-RAG study interleaved retrieval for multi-step QA (Trivedi et al., 2023; Jiang et al., 2023; Du et al., 2026); none targets our procedural case, finalizing without an explicit read. Self-RAG, CRAG and RL-trained searchers learn when to retrieve or correct (Asai et al., 2024; Yan et al., 2024; Jin et al., 2025; Chen et al., 2025; Song et al., 2025; An et al., 2026); parametric answers can beat retrieval on popular facts (Mallen et al., 2023) and irrelevant passages degrade answers (Yoran et al., 2024), one source of the disruption a demanded read can cause. READ-GATE instead is a deterministic, training-free harness invariant before finalization, unlike feedback-driven self-correction (Shinn et al., 2023; Gou et al., 2024), whose limits without external feedback are known (Huang et al., 2024); our judging follows LLM-as-judge practice (Zheng et al., 2023; Liu et al., 2023). Budget forcing appends *Wait* at a stopping decision (Muennighoff et al., 2025); our structured delay is analogous, and we measure the answer obligation's effect factorially. The ladder compares gate and delay at an identical prefix, and read-over-delay superiority does not replicate; the cost frontier (Appendix H.1) compares a one-shot demand with raised test-time effort (Snell et al., 2025).

8 Limitations and conclusion

Scope. The factorial magnitudes are measured only for gpt-5-mini at minimal effort (at medium effort the read demand gains $+0.17$pp $[-0.33, +0.67]$, Appendix H.1); Wikipedia snippets often hold the whole chunk, medical ones never (Appendix Table S26); the Qwen2.5-7B factorial detects no obligation benefit (rare hand-backs), the cross-controller map is descriptive, and the medical study is one external corpus. *Identification.* Each rung is a bundle, so a contrast attributes gain to what one bundle adds, and an obligation without an interruption cannot be injected at the branch point, so no F-only cell exists; the answer obligation removes one format difference, not all; RF$-$NF compares a read request with a noop request, not evidence with its absence, since the control may read voluntarily; no measured moderator explains the panel gap (Appendix D). *Inference.* Replications are question-disjoint, not corpus-external, except P2; decoder noise is estimated only on a 60-question trigger sample; controller and judge share a family, and both judge audits (Appendices J.3 and K.2) are targeted checks, not independent human annotation. *Pre-registration.* Primary families changed across ladder experiments and P1 took one interim look; all are recorded (Appendix B.1, §5).

Conclusion. For gpt-5-mini at minimal effort, our title's question has a conditional answer: largely the interruption, once the control must also answer. An obligated interruption is not separated from the unobligated read demand on Wikipedia and beats it on the medical corpus, and with every intervention arm so obliged the read demand's edge over an information-free action shrinks from $+3.3$ to $+1.2$pp on Wikipedia and is not detected on a medical corpus (-0.4pp $[-2.3, +1.5]$). Neither a gate-versus-no-gate gain nor an unmatched read/noop gap is by itself evidence that retrieved content helps. Deployment needs a pilot: transported Wikipedia rates mispredicted the medical payoff by 5.3 points.

AI USE STATEMENT

Large language models, operated through a coding agent, were used to write, test and review the experiment, analysis and figure code, and to draft, edit and review the manuscript. They also served as the two LLM raters of the judge-validity audit, for the blinded annotation passes on the manual-audit sheets and for the entity-matcher validation labels; those labels are reported as AI ratings with author adjudication, not as human annotation (Appendices E.2, K.2 and J.2). The authors designed the study, approved every paid run, reviewed all AI-assisted code, analyses, numbers, citations and prose, and take responsibility for the final content and artifacts.

REPRODUCIBILITY STATEMENT

Appendices A–L follow the references. Every experiment is defined by a frozen manifest (arms, ordered question IDs with hashes, model snapshots, prompts and byte-exact gate messages, cost ceilings, pre-specified contrasts and decision rules, and hashes of the implementation files) and closed by an audited analysis whose checks are listed in Appendix A.3. The supplementary artifact contains the implementation, manifests, question-ID lists and the pinned revision of the benchmark corpora, all raw trajectories and judge outputs for the ladder, delay, valence-control, replication, medium-effort, second-family, scout and smoke experiments, the three answer-obligation studies with their frozen forecast and interim-look records, the cross-family judge re-audit and the AI-rater audit pack, the derived analyses behind every table and figure, and scripts that regenerate the headline analyses without paid API calls (verified in a clean environment; a few inherited summaries — read-quality proxies, prevalence, cost and scout tables — ship as frozen outputs with source hashes rather than as regenerations, and the post-review read-exposure check ships the compared strings rather than full transcripts).

REFERENCES

Alibaba Cloud. Model Studio: supported models. https://www.alibabacloud.com/help/en/model-studio/models, 2026. Accessed 2026-08-31; Qwen3.7 snapshots.

Kang An, Ziliang Wang, Xuhui Zheng, FaQiang Qian, WeiKun Zhang, Yuhang Wang, and Yichao Wu. Erase to improve: Erasable reinforcement learning for search-augmented LLMs. In *International Conference on Learning Representations (ICLR 2026)*, 2026.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024.

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 38 (NeurIPS 2025)*, 2025.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 758–759, 2009. doi: 10.1145/1571941.1572114.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Shaohan Wang, Pengyu Wang, Xiaorui Wang, and Zhendong Mao. A-RAG: Scaling agentic retrieval-augmented generation via hierarchical retrieval interfaces. *arXiv preprint arXiv:2602.03442*, 2026.

Google. Gemini 2.5 Flash model documentation. https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash, 2026. Accessed 2026-08-31.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations (ICLR 2024)*, 2024.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, 2023.

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. In *Conference on Language Modeling (COLM 2025)*, 2025.

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

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)*, pp. 5420–5438, 2025. doi: 10.18653/v1/2025.emnlp-main.276.

Xiaofeng Lin, Yingxu Wang, Tung Sum Thomas Kwok, Daniel Guo, Sahil Arun Nale, Charles Fleming, and Guang Cheng. REFLECT: Intervention-supported error attribution for silent failures in LLM agent traces. *arXiv preprint arXiv:2606.09071*, 2026. URL https://arxiv.org/abs/2606.09071.

Jason Liu. When may an agent stop? Evidence-Carrying Termination for tool-using LLMs. *arXiv preprint arXiv:2608.23623*, 2026. URL https://arxiv.org/abs/2608.23623.

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*, pp. 2511–2522, 2023.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 9802–9822, 2023.

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. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*, pp. 20275–20321. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1025. URL https://aclanthology.org/2025.emnlp-main.1025/.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

OpenAI. GPT-5 mini model documentation. https://developers.openai.com/api/docs/models/gpt-5-mini, 2026a. Accessed 2026-08-31.

OpenAI. GPT-5.6 Luna model documentation. https://developers.openai.com/api/docs/models/gpt-5.6-luna, 2026b. Accessed 2026-08-31.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, 2023.

Jaineet Shah. Causal Agent Replay: Counterfactual attribution for LLM-agent failures. *arXiv preprint arXiv:2606.08275*, 2026. URL https://arxiv.org/abs/2606.08275.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations (ICLR 2025)*, 2025.

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.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 10014–10037, 2023.

Haoyu Wang, Christopher M. Poskitt, and Jun Sun. AgentSpec: Customizable runtime enforcement for safe and reliable LLM agents. In *Proceedings of the 48th International Conference on Software Engineering*, 2026. URL https://arxiv.org/abs/2503.18666.

Zhishang Xiang, Chuanjie Wu, Qinggang Zhang, Shengyuan Chen, Zijin Hong, Xiao Huang, and Jinsong Su. When to use graphs in RAG: A comprehensive analysis for graph retrieval-augmented generation. *arXiv preprint arXiv:2506.05690*, 2025.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024.

Chenxu Yang, Qingyi Si, Mz Dai, Dingyu Yao, Mingyu Zheng, Minghui Chen, Zheng Lin, and Weiping Wang. Test-time prompt intervention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(40):34223–34231, 2026. doi: 10.1609/aaai.v40i40.40718.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, 2024.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

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, Datasets and Benchmarks Track*, 2023.

Luyao Zhuang, Shengyuan Chen, Yilin Xiao, Huachi Zhou, Yujing Zhang, Hao Chen, Qinggang Zhang, and Xiao Huang. LinearRAG data release: chunk corpora and question pools for HotpotQA, 2WikiMultihop and MuSiQue. Hugging Face dataset https://huggingface.co/datasets/Zly0523/linear-rag, distributed with https://github.com/DEEP-PolyU/LinearRAG, 2025a. Revision f7971bb84fe205d0c080df5f48b1393d9f580c68; accessed 2026-08-31.

Luyao Zhuang, Shengyuan Chen, Yilin Xiao, Huachi Zhou, Yujing Zhang, Hao Chen, Qinggang Zhang, and Xiao Huang. LinearRAG: Linear graph retrieval augmented generation on large-scale

corpora. *arXiv preprint arXiv:2510.10114*, 2025b.

A LADDER PROTOCOL DETAILS

A.1 ARM DEFINITIONS AND VERBATIM GATE MESSAGES

Every arm shares the trajectory up to the first zero-read `final` attempt. At that point the harness rejects the attempt (except NO-GATE) by appending one user-role message and continuing the loop; every one-shot arm then disarms permanently, and a later premature final is accepted and recorded as such. Every injected message is a module-level constant in `agents/base_agent.py`, pinned by sha256 in the test suite (`tests/test_mechanism_ladder_gates.py`; `tests/test_delay_v2_gate.py` for the delay variants); the strings below are copied byte-for-byte from those constants, not retyped from a description.

DELAY (Rung 1, `STRUCTURED_DELAY_MESSAGE`). "Finalization is deferred once. Your next message will be treated as your final answer unless it performs another step. Do not reply with an acknowledgement or a question."

RESTATE (valence control, Appendix D). "Finalization is deferred once. Please restate your final answer. Your next message will be treated as your final answer unless it performs another step. Do not reply with an acknowledgement or a question."

NOOP (Rung 2, `NOOP_GATE_MESSAGE`). "Finalization is deferred once. You MUST call the noop tool before submitting a final answer. The noop tool performs no operation and returns no information." The `noop` tool (no parameters) is exposed only after this message and returns `NOOP_TOOL_RESULT`: "noop acknowledged. No operation was performed."

ANY-TOOL. "Finalization is deferred once. You MUST call one of your available tools before submitting a final answer. Any tool is acceptable." No new tool is exposed; compliance is any ordinary tool call after the message.

READ-ONCE and READ (Rungs 3–4, `READ_GATE_MESSAGE_PLAIN`). "You attempted to answer before reading any evidence. You MUST call read(evidence_id) on a promising chunk before giving the final answer." This plain-channel variant fires when the premature final arrives as assistant-message content; when it arrives as a call to a `final`-style tool, the tool-channel variant `READ_GATE_MESSAGE_TOOL` fires instead: "You attempted to finish before reading any evidence. Call read(evidence_id) on a promising chunk before finishing." Both rungs inject the identical string on their first firing and differ only in whether the obligation re-arms: READ repeats it at every subsequent zero-read final and returns an explicit abstention if the loop or token limit is reached with the obligation open; READ-ONCE sends it once.

SHAM-READ. Sends the READ-ONCE message once; every subsequent `read` then returns, under the requested id, the text of the corpus chunk selected by the sha256 of the requested id (skipping the requested chunk and every chunk seen in any search result), and the read is recorded as a read.

Superseded control (`NEUTRAL_DELAY_MESSAGE`, v1 delay). This open-ended message produced the hand-back artifact reported in §4.1 (Appendix I.2); the structured DELAY above is its format-preserving replacement and is what every reported DELAY number uses. "Finalization is deferred once. Use the next decision step however you judge appropriate."

A.2 ONE-SHOT / DISARM SEMANTICS AND POST-TRIGGER TOOL EXPOSURE

Each rung's obligation is tracked by one boolean flag (`delay_gate_satisfied`, `rg_once_satisfied`, `noop_gate_satisfied`) that starts `False` and is set `True` the first time the gate fires on a trajectory. The firing condition for every one-shot rung is identical: the model's current final attempt has `ctx.read_chunk_ids` empty (zero successful reads so far) *and* the rung's flag is still `False`. Once the flag flips, that rung never fires again on the same trajectory, in either the plain or the tool-call channel — a later zero-read final is accepted and recorded as such (this is what READ-ONCE and NOOP mean by "one-shot"). DELAY, NOOP and READ-ONCE are mutually exclusive with each other and with the persistent READ policy at the agent's construction time (the class raises `ValueError` if two are enabled together), so a single run object always implements exactly one rung of the ladder. READ (Rung 4) uses a separate, persistent flag (`require_read_before_final`) that is never set once satisfied — the gate re-fires on every zero-read final until a read occurs or the loop / token budget

is exhausted, at which point the trajectory returns an explicit abstention (`abstained: True`, empty answer) rather than a forced answer; this is the only rung with an abstention path.

The noop tool is not part of the agent's base tool schema. It is appended to the tool list passed to the LLM client (`call_tools = self.tool_schemas + [NOOP_TOOL_SCHEMA]`) only on loops where `noop_gate_before_final and noop_gate_satisfied` both hold, i.e. only after the gate has already fired once. Every loop before the trigger therefore sees the identical tool surface as the NO-GATE arm, so all pre-trigger LLM calls replay from the same deterministic response cache across arms; only the branch point and what follows it differ. The same principle holds for NOOP's tool result: calling noop is intercepted internally (`if fname == "noop" and self.noop_gate_before_final`) and answered with the fixed string above without ever reaching the caller-supplied dispatcher, so it carries no retrieved tokens and no information about the corpus.

A.3 Manifest, preflight and audit checklist

Every experiment freezes a manifest before execution (arms, ordered question IDs with their sha256 hashes, model snapshots, cost ceilings, primary and secondary contrasts, and the sha256 of every implementation file the run depends on) and is closed by an automated audit whose checks must all pass before a result is marked `citable`. The checks are those of the `trace_audit` block of the frozen replication analysis (`results/_delay_gate/iclr27_core_replication_3data_n400_s42.json`); the same categories recur, per arm and per dataset, in every ladder-family analysis in this paper (§C):

They fall into five categories, each instantiated per arm and per dataset: (1) *manifest integrity* — any post-freeze amendment's declared base hash matches the manifest on disk, the frozen question-ID list's sha256 matches what the run consumed, the replication panel has zero overlap with the ladder panel and any prior smoke, and the manifest hash embedded in the result matches the file; (2) *execution fidelity* — every arm processed the exact frozen ID sequence in order with no duplicates or drops, no row carries a top-level error, every row carries the schema fields the analyzer needs, and a hash over the raw per-row status/answer fields detects post-hoc edits; (3) *arm configuration* — the recorded gate flags match the arm's manifest declaration (guarding against a mislabeled arm), the resolved agent and judge strings match the pinned dated snapshots, and the arm read from its own cache directory; (4) *implementation hashes* — the sha256 of every implementation file the analysis depends on (batch runner, agent, context, LLM client and cache, judge) matches the file on disk at analysis time; (5) *cost ceiling* — actual spend did not exceed the per-arm ceiling.

An analysis is marked `status: verified` and `citable_accuracy_result: true` only when `trace_audit.all_pass` is `true` and every arm reached its expected row count (`analysis_stage: final`); a preflight of one record per cell must pass this same checklist before the full run is authorized to spend. For experiments whose manifest declares explicit primary contrasts and has no DELAY arm (the replication), the analysis script records `decision: primary_contrasts_supported` and sets the legacy `g4_pass` field (defined only for the Delay-Gate pilots) to null; no number in this paper depends on those two fields.

Arm identifiers. Generated tables use the internal arm ids: `norg` = NO-GATE; `anytool` = ANY-TOOL; `sham_read` = SHAM-READ; `delay_v2` = DELAY (the structured message used in the main text); `delay` = the superseded open-ended delay (v1, §4.1); `noop_once` = NOOP; `rg_once` = READ-ONCE; `rg` = READ.

B Uncertainty for the original ladder contrasts

The main text reports selected intervals alongside point estimates and discordant counts. This appendix covers the original ladder; the format studies and Qwen factorial are reported in Appendix K (Tables S22 and S25). Table S1 lists the 95% stratified question-cluster bootstrap interval (10,000 draws, seed fixed in the manifests) and discordant-pair counts for the ladder, medium-effort, any-tool and sham-read contrasts quoted there; the post-hoc restate control, whose frozen summary did not include bootstrap draws, is documented with discordant counts in Appendix D. Exact McNemar p-values are in the analysis JSONs. "P" marks a contrast that was a primary of that experiment's frozen manifest (each experiment had its own primary family), "S" a secondary or descriptive contrast; the original ladder's focal contrast is READ-ONCE−NOOP, and the any-tool, restate and sham-read

arms are post-hoc controls. Intervals for the prevalence map (exact Clopper–Pearson) are in the prevalence table of this appendix; Figure S1 shows the ladder contrasts' intervals graphically.

Table S1: 95% stratified question-cluster bootstrap intervals (10,000 draws) and fixes/breaks for the ladder, medium-effort, any-tool and sham-read contrasts quoted in the main text; the retired open-ended v1 delay appears only through the one contrast the audit lesson of §4.1 quotes. The post-hoc restate control is reported separately in Appendix D. Generated by `scripts/build_main_contrast_ci_table.py` from the frozen analysis JSONs.

Panel	Contrast	Est. (pp)	95% CI	Fix/break	P/S
Panel (n=600)	READ − READ-ONCE	+0.33	[+0.00, +0.83]	2/0	P
Panel (n=600)	READ-ONCE − NOOP	+3.83	[+2.00, +5.67]	28/5	P
Panel (n=600)	DELAY_{v1} − NO-GATE	−6.00	[−8.67, −3.33]	18/54	S
Panel (n=600)	READ − NO-GATE	+4.83	[+3.00, +6.83]	34/5	S
Panel (n=600)	READ-ONCE − NO-GATE	+4.50	[+2.67, +6.50]	32/5	S
Panel (n=600)	NOOP − NO-GATE	+0.67	[−0.83, +2.17]	13/9	S
Delay-v2 panel (n=600)	DELAY − NO-GATE	+1.17	[−0.33, +2.83]	16/9	P
Delay-v2 panel (n=600)	READ-ONCE − DELAY	+3.33	[+1.33, +5.33]	29/9	P
Delay-v2 panel (n=600)	NOOP − DELAY	−0.50	[−2.33, +1.33]	15/18	S
Delay-v2 panel (n=600)	READ-ONCE − NOOP	+3.83	[+2.00, +5.67]	28/5	S
Delay-v2 panel (n=600)	READ − READ-ONCE	+0.33	[+0.00, +0.83]	2/0	S
Delay-v2 panel (n=600)	READ-ONCE − NO-GATE	+4.50	[+2.67, +6.50]	32/5	S
Delay-v2 panel (n=600)	NOOP − NO-GATE	+0.67	[−0.83, +2.17]	13/9	S
Repl. core (n=1,200)	READ-ONCE − NOOP	+3.08	[+1.67, +4.58]	60/23	P
Repl. core (n=1,200)	READ-ONCE − NO-GATE	+5.33	[+3.83, +6.83]	78/14	S
Repl. core (n=1,200)	NOOP − NO-GATE	+2.25	[+0.92, +3.58]	47/20	S
Repl. full ladder (n=1,200)	READ-ONCE − DELAY	+0.25	[−1.17, +1.75]	40/37	P
Repl. full ladder (n=1,200)	READ − READ-ONCE	+0.58	[+0.17, +1.08]	8/1	P
Repl. full ladder (n=1,200)	DELAY − NO-GATE	+5.08	[+3.58, +6.58]	74/13	S
Repl. full ladder (n=1,200)	NOOP − DELAY	−2.83	[−4.17, −1.50]	16/50	S
Repl. full ladder (n=1,200)	READ-ONCE − NOOP	+3.08	[+1.67, +4.58]	60/23	S
Repl. full ladder (n=1,200)	READ-ONCE − NO-GATE	+5.33	[+3.83, +6.83]	78/14	S
Repl. full ladder (n=1,200)	NOOP − NO-GATE	+2.25	[+0.92, +3.58]	47/20	S
Repl. full ladder (n=1,200)	READ − NO-GATE	+5.92	[+4.33, +7.50]	86/15	S
Medium effort 2-arm (n=600)	READ-ONCE − NO-GATE	+0.17	[−0.33, +0.67]	2/1	P
Any-tool panel (n=600)	ANY-TOOL − NO-GATE	+1.50	[−0.33, +3.33]	21/12	S
Any-tool panel (n=600)	ANY-TOOL − DELAY	+0.33	[−1.83, +2.50]	24/22	S
Any-tool panel (n=600)	ANY-TOOL − noop	+0.83	[−1.33, +3.00]	25/20	S
Any-tool panel (n=600)	READ-ONCE − ANY-TOOL	+3.00	[+0.83, +5.17]	31/13	S
Any-tool repl. (n=1,200)	ANY-TOOL − NO-GATE	+3.58	[+1.83, +5.33]	80/37	S
Any-tool repl. (n=1,200)	ANY-TOOL − DELAY	−1.50	[−3.17, +0.17]	40/58	S
Any-tool repl. (n=1,200)	ANY-TOOL − noop	+1.33	[−0.25, +3.00]	58/42	S
Any-tool repl. (n=1,200)	READ-ONCE − ANY-TOOL	+1.75	[+0.17, +3.33]	57/36	S
Any-tool pooled (n=1,800)	ANY-TOOL − NO-GATE	+2.89	[+1.56, +4.22]	101/49	S
Any-tool pooled (n=1,800)	ANY-TOOL − DELAY	−0.89	[−2.22, +0.44]	64/80	S
Any-tool pooled (n=1,800)	ANY-TOOL − noop	+1.17	[−0.17, +2.50]	83/62	S
Any-tool pooled (n=1,800)	READ-ONCE − ANY-TOOL	+2.17	[+0.89, +3.44]	88/49	S
Sham-read panel (n=600)	READ-ONCE − SHAM-READ	+2.50	[+0.83, +4.17]	21/6	S
Sham-read panel (n=600)	SHAM-READ − noop	+1.33	[−0.50, +3.17]	20/12	S
Sham-read panel (n=600)	SHAM-READ − DELAY	+0.83	[−1.17, +2.83]	21/16	S
Sham-read panel (n=600)	SHAM-READ − NO-GATE	+2.00	[+0.33, +3.83]	21/9	S
Sham-read repl. (n=1,200)	READ-ONCE − SHAM-READ	+0.42	[−0.75, +1.58]	27/22	S
Sham-read repl. (n=1,200)	SHAM-READ − noop	+2.67	[+1.25, +4.17]	56/24	S
Sham-read repl. (n=1,200)	SHAM-READ − DELAY	−0.17	[−1.58, +1.25]	36/38	S
Sham-read repl. (n=1,200)	SHAM-READ − NO-GATE	+4.92	[+3.42, +6.42]	74/15	S
Sham-read pooled (n=1,800)	READ-ONCE − SHAM-READ	+1.11	[+0.17, +2.06]	48/28	S
Sham-read pooled (n=1,800)	SHAM-READ − noop	+2.22	[+1.06, +3.39]	76/36	S
Sham-read pooled (n=1,800)	SHAM-READ − DELAY	+0.17	[−1.00, +1.28]	57/54	S
Sham-read pooled (n=1,800)	SHAM-READ − NO-GATE	+3.94	[+2.78, +5.11]	95/24	S
Combined HotpotQA (n=600)	READ-ONCE − NOOP	+3.33	[+1.50, +5.33]	27/7	S

Table S1 (continued)

Panel	Contrast	Est. (pp)	95% CI	Fix/break	P/S
Combined 2Wiki (n=600)	READ-ONCE − NOOP	+1.17	[−0.67, +3.00]	20/13	S
Combined MuSiQue (n=600)	READ-ONCE − NOOP	+5.50	[+3.33, +7.83]	41/8	S
Combined pooled (n=1,800)	READ-ONCE − NOOP	+3.33	[+2.22, +4.50]	—	S
Combined HotpotQA (n=600)	READ-ONCE − NO-GATE	+2.67	[+1.00, +4.33]	22/6	S
Combined 2Wiki (n=600)	READ-ONCE − NO-GATE	+4.50	[+2.50, +6.67]	35/8	S
Combined MuSiQue (n=600)	READ-ONCE − NO-GATE	+8.00	[+5.67, +10.50]	53/5	S
Combined pooled (n=1,800)	READ-ONCE − NO-GATE	+5.06	[+3.89, +6.28]	—	S
Combined HotpotQA (n=600)	NOOP − NO-GATE	−0.67	[−2.50, +1.17]	13/17	S
Combined 2Wiki (n=600)	NOOP − NO-GATE	+3.33	[+1.83, +5.00]	22/2	S
Combined MuSiQue (n=600)	NOOP − NO-GATE	+2.50	[+0.67, +4.50]	25/10	S
Combined pooled (n=1,800)	NOOP − NO-GATE	+1.72	[+0.72, +2.78]	—	S
Combined HotpotQA (n=600)	DELAY − NO-GATE	+1.33	[−0.17, +3.00]	16/8	S
Combined 2Wiki (n=600)	DELAY − NO-GATE	+4.33	[+2.50, +6.33]	30/4	S
Combined MuSiQue (n=600)	DELAY − NO-GATE	+5.67	[+3.33, +8.00]	44/10	S
Combined pooled (n=1,800)	DELAY − NO-GATE	+3.78	[+2.61, +4.94]	—	S
Combined HotpotQA (n=600)	READ-ONCE − DELAY	+1.33	[−0.50, +3.17]	20/12	S
Combined 2Wiki (n=600)	READ-ONCE − DELAY	+0.17	[−1.67, +2.00]	16/15	S
Combined MuSiQue (n=600)	READ-ONCE − DELAY	+2.33	[+0.00, +4.67]	33/19	S
Combined pooled (n=1,800)	READ-ONCE − DELAY	+1.28	[+0.11, +2.50]	—	S
Combined HotpotQA (n=600)	NOOP − DELAY	−2.00	[−3.83, −0.17]	10/22	S
Combined 2Wiki (n=600)	NOOP − DELAY	−1.00	[−2.50, +0.33]	6/12	S
Combined MuSiQue (n=600)	NOOP − DELAY	−3.17	[−5.50, −1.00]	15/34	S
Combined pooled (n=1,800)	NOOP − DELAY	−2.06	[−3.17, −1.00]	—	S
Combined HotpotQA (n=600)	READ − READ-ONCE	+0.17	[+0.00, +0.50]	1/0	S
Combined 2Wiki (n=600)	READ − READ-ONCE	+0.33	[+0.00, +0.83]	2/0	S
Combined MuSiQue (n=600)	READ − READ-ONCE	+1.00	[+0.17, +2.00]	7/1	S
Combined pooled (n=1,800)	READ − READ-ONCE	+0.50	[+0.17, +0.89]	—	S
Combined HotpotQA (n=600)	READ − NO-GATE	+2.83	[+1.17, +4.67]	23/6	S
Combined 2Wiki (n=600)	READ − NO-GATE	+4.83	[+2.67, +7.00]	37/8	S
Combined MuSiQue (n=600)	READ − NO-GATE	+9.00	[+6.50, +11.67]	60/6	S
Combined pooled (n=1,800)	READ − NO-GATE	+5.56	[+4.33, +6.83]	—	S

B.1 CLAIM MAP AND MULTIPLICITY

Primary status belongs to an experiment and its frozen manifest, not to a contrast selected for the paper's headline. The ladder draws on four experiment-level primary families (Table S2, first blocks; generated from the immutable manifests and analysis JSONs by `scripts/build_claim_map_table.py`):

- *Discovery ladder* ($n = 600$): READ−READ-ONCE and READ-ONCE−NOOP, Holm over two tests. The first is unsupported ($p = 0.500$); the second is supported (raw $p = 6.6 \times 10^{-5}$, adjusted 1.3×10^{-4}).
- *Structured-delay replacement* (the same $n = 600$ questions): DELAY−NO-GATE and READ-ONCE−DELAY, Holm over two tests; adjusted $p = 0.230$ and 0.0033. This is an additional control experiment on the original panel, not a replication.
- *Core replication* ($n = 1{,}200$): READ-ONCE−NOOP is the single primary ($p = 6.0 \times 10^{-5}$; no within-experiment correction).
- *Full-ladder replication* (the same $n = 1{,}200$ questions): READ-ONCE−DELAY and READ−READ-ONCE, Holm over two tests. Neither is supported: adjusted $p = 0.820$ and 0.0781 (the latter's raw $p = 0.0391$). Replayed core arms retain their original experiment's status.

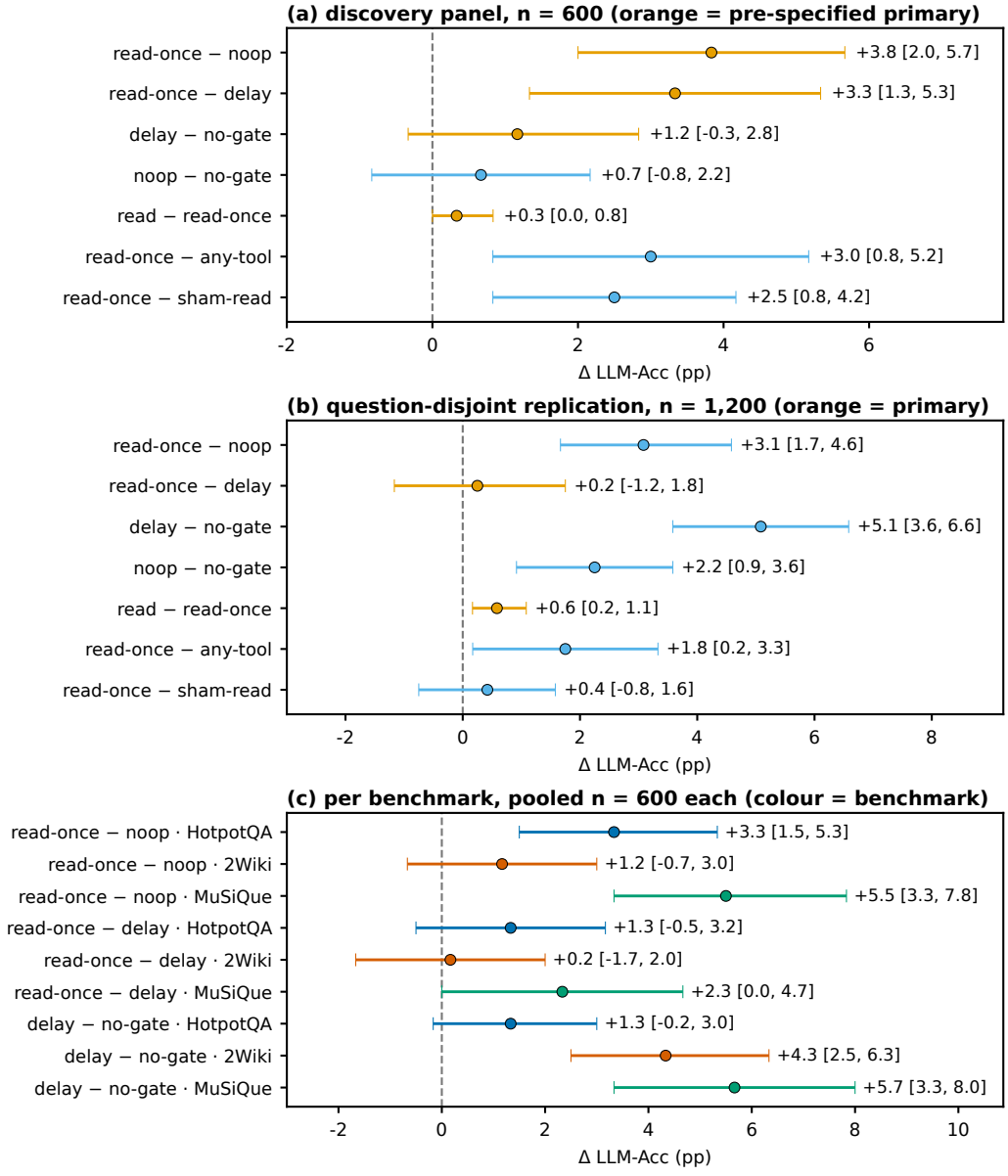

Figure S1: Forest plot of seven focal contrasts: (a) panel, (b) replication and (c) ladder contrasts by benchmark. Points are paired ITT differences; bars are 95% stratified question-cluster bootstrap intervals. Orange marks contrasts pre-specified as primary in the relevant experiment manifest; multiple experiments contribute to each panel. Blue marks the remaining contrasts, and panel (c) uses colour for benchmark.

The replicated read-over-noop result therefore survives its original correction, while the pre-specified read-over-delay replication does not. Nominal pointwise 95% intervals are not simultaneous intervals and need not agree with multiplicity-adjusted decisions.

The pooled ($n = 1{,}800$) blocks of Table S2 are later reporting choices, not new confirmatory tests or replacements for the original primaries: Holm is applied to the eight pooled ladder contrasts and, separately, to the five pooled side-control contrasts, whose manifests declared no confirmatory primary. Grouping all five side controls into one family is the conservative choice; splitting them would only lower the adjusted values. In that family ANY-TOOL−NOOP has adjusted $p = 0.096$ and READ-ONCE−SHAM-READ adjusted $p = 0.057$; both remain diagnostic, and their non-significance does not establish equivalence. The medium-effort experiment retains its own single primary in the same table. The original v1 delay's construct failure is retained in Appendix I.2; its results are not promoted by the replacement control.

Table S2: Original primary families and later pooled reporting. The first five blocks preserve every primary in the corresponding verified experiment, including the structured-delay follow-up and the separate medium-effort study. Exact McNemar p-values are recomputed from frozen paired counts; adjusted p-values use each original family. The final two blocks are later descriptive pooled families, not pre-registered pooled primaries. P/S in Table S1 refers to the named source experiment; a contrast may have different roles in different experiments. Nominal 95% CIs are not simultaneous Holm intervals. Retired delay-v1 results are audit diagnostics only and are absent from this ledger.

Contrast	Role	Est. (pp)	Raw p	Adjusted p
Original primary family: Original ladder panel ($n = 600$); Holm over the original 2 primaries.				
READ − READ-ONCE	original P	$+0.33$	0.5	0.5
READ-ONCE − NOOP	original P	$+3.83$	6.6×10^{-5}	1.3×10^{-4}
Original primary family: Structured-delay follow-up panel ($n = 600$); Holm over the original 2 primaries.				
DELAY − NO-GATE	original P	$+1.17$	0.23	0.23
READ-ONCE − DELAY	original P	$+3.33$	0.00166	0.00332
Original primary family: Core replication ($n = 1,200$); single primary (no multiplicity adjustment).				
READ-ONCE − NOOP	original P	$+3.08$	6.0×10^{-5}	—
Original primary family: Full-ladder replication ($n = 1,200$); Holm over the original 2 primaries.				
READ-ONCE − DELAY	original P	$+0.25$	0.82	0.82
READ − READ-ONCE	original P	$+0.58$	0.0391	0.0781
Original primary family: Medium-effort boundary ($n = 600$); single primary (no multiplicity adjustment).				
READ-ONCE − NO-GATE	original P	$+0.17$	1	—
Later pooled ladder reporting ($n = 1{,}800$; descriptive); editorial Holm family over all 8 rows; does not replace any original family.				
READ-ONCE − NOOP	pooled D	$+3.33$	2.1×10^{-8}	1.0×10^{-7}
READ − READ-ONCE	pooled D	$+0.50$	0.0117	0.0234
READ-ONCE − NO-GATE	pooled D	$+5.06$	9.1×10^{-17}	6.4×10^{-16}
READ − NO-GATE	pooled D	$+5.56$	1.4×10^{-18}	1.1×10^{-17}
NOOP − NO-GATE	pooled D	$+1.72$	0.00134	0.00401
DELAY − NO-GATE	pooled D	$+3.78$	6.0×10^{-11}	3.6×10^{-10}
READ-ONCE − DELAY	pooled D	$+1.28$	0.0398	0.0398
NOOP − DELAY	pooled D	-2.06	2.6×10^{-4}	0.00102
Later pooled side-control reporting ($n = 1{,}800$; descriptive); editorial Holm family over 5 rows; each side-control manifest declared no confirmatory primary.				
READ-ONCE − ANY-TOOL	pooled D	$+2.17$	0.00109	0.00327
ANY-TOOL − NOOP	pooled D	$+1.17$	0.0964	0.0964
READ-ONCE − SHAM-READ	pooled D	$+1.11$	0.0286	0.0573
SHAM-READ − NOOP	pooled D	$+2.22$	2.0×10^{-4}	7.9×10^{-4}
RESTATE − NO-GATE	pooled D	$+3.28$	1.1×10^{-14}	5.5×10^{-14}

C FULL PER-DATASET TABLES

Tables S3–S6 are generated by `scripts/build_iclr_appendix_tables.py` directly from the frozen analysis JSONs and are never hand-typed; they use the paper's arm names (the frozen JSONs' run identifiers are `norg` = NO-GATE, `noop_once` = NOOP, `rg_once` = READ-ONCE, `rg` = READ, `delay_v2` = DELAY and `delay` = the retired open-ended v1 delay). Per-dataset cells report the point estimate (net accuracy points from fixes−breaks) with fixes/breaks in parentheses. In Tables S3–S5 per-dataset intervals are not computed, so only the pooled column carries one, matching what §4 cites; the combined analysis of Table S6 reports per-dataset intervals as well.

Table S3: Mechanism-ladder panel ($n = 200$/dataset, $n = 600$ pooled): all preregistered primary and secondary contrasts, ITT, per dataset and pooled. 95% stratified cluster-bootstrap intervals. Per-dataset cells are point estimates (net fixes−breaks, in pp) with fixes/breaks in parentheses; the pooled column additionally reports the 95% stratified cluster-bootstrap interval.

Contrast	HotpotQA	2WikiMultihop	MuSiQue	Pooled (pp, 95% CI)
read − read-once	+0.00 (0/0)	+0.50 (1/0)	+0.50 (1/0)	+0.33 [+0.00, +0.83]
read-once − noop	+1.50 (6/3)	+3.00 (7/1)	+7.00 (15/1)	+3.83 [+2.00, +5.67]
noop − delay (v1, retired)	+7.00 (16/2)	+4.00 (13/5)	+9.00 (26/8)	+6.67 [+4.00, +9.33]
read-once − delay (v1, retired)	+8.50 (19/2)	+7.00 (18/4)	+16.00 (35/3)	+10.50 [+7.83, +13.33]
read − delay (v1, retired)	+8.50 (19/2)	+7.50 (18/3)	+16.50 (36/3)	+10.83 [+8.17, +13.50]
delay (v1, retired) − no-gate	-8.50 (1/18)	-3.50 (6/13)	-6.00 (11/23)	-6.00 [-8.67, -3.33]
read − no-gate	+0.00 (4/4)	+4.00 (9/1)	+10.50 (21/0)	+4.83 [+3.00, +6.83]
read-once − no-gate	+0.00 (4/4)	+3.50 (8/1)	+10.00 (20/0)	+4.50 [+2.67, +6.50]
noop − no-gate	-1.50 (3/6)	+0.50 (1/0)	+3.00 (9/3)	+0.67 [-0.83, +2.17]

Table S4: Delay-v2 experiment (same 600-question panel, adds the structured DELAY arm): all preregistered contrasts, per dataset and pooled. 95% stratified cluster-bootstrap intervals. Per-dataset cells are point estimates (net fixes−breaks, in pp) with fixes/breaks in parentheses; the pooled column additionally reports the 95% stratified cluster-bootstrap interval.

Contrast	HotpotQA	2WikiMultihop	MuSiQue	Pooled (pp, 95% CI)
delay − no-gate	-1.50 (0/3)	+0.50 (3/2)	+4.50 (13/4)	+1.17 [-0.33, +2.83]
read-once − delay	+1.50 (6/3)	+3.00 (8/2)	+5.50 (15/4)	+3.33 [+1.33, +5.33]
noop − delay	+0.00 (4/4)	+0.00 (3/3)	-1.50 (8/11)	-0.50 [-2.33, +1.33]
delay − delay (v1, retired)	+7.00 (15/1)	+4.00 (13/5)	+10.50 (25/4)	+7.17 [+4.67, +9.67]
read-once − noop	+1.50 (6/3)	+3.00 (7/1)	+7.00 (15/1)	+3.83 [+2.00, +5.67]
read − read-once	+0.00 (0/0)	+0.50 (1/0)	+0.50 (1/0)	+0.33 [+0.00, +0.83]
read-once − no-gate	+0.00 (4/4)	+3.50 (8/1)	+10.00 (20/0)	+4.50 [+2.67, +6.50]
noop − no-gate	-1.50 (3/6)	+0.50 (1/0)	+3.00 (9/3)	+0.67 [-0.83, +2.17]
delay (v1, retired) − no-gate	-8.50 (1/18)	-3.50 (6/13)	-6.00 (11/23)	-6.00 [-8.67, -3.33]

Table S5: Preregistered replication ($n = 400$/dataset, $n = 1{,}200$ pooled, disjoint from the panel): all contrasts, per dataset and pooled. 95% stratified cluster-bootstrap intervals. Per-dataset cells are point estimates (net fixes−breaks, in pp) with fixes/breaks in parentheses; the pooled column additionally reports the 95% stratified cluster-bootstrap interval.

Contrast	HotpotQA	2WikiMultihop	MuSiQue	Pooled (pp, 95% CI)
read-once − noop	+4.25 (21/4)	+0.25 (13/12)	+4.75 (26/7)	+3.08 [+1.67, +4.58]
read-once − no-gate	+4.00 (18/2)	+5.00 (27/7)	+7.00 (33/5)	+5.33 [+3.83, +6.83]
noop − no-gate	-0.25 (10/11)	+4.75 (21/2)	+2.25 (16/7)	+2.25 [+0.92, +3.58]

Table S6: Combined per-dataset estimates (panel + replication pooled, $n = 600$ per benchmark, $n = 1{,}800$ pooled). 95% stratified cluster-bootstrap intervals.

Contrast (pp)	HotpotQA	2Wiki	MuSiQue	Pooled (n=1,800)
read-once − noop	+3.33 [+1.50, +5.33]	+1.17 [-0.67, +3.00]	+5.50 [+3.33, +7.83]	+3.33 [+2.22, +4.50]
read-once − no-gate	+2.67 [+1.00, +4.33]	+4.50 [+2.50, +6.67]	+8.00 [+5.67, +10.50]	+5.06 [+3.89, +6.28]
noop − no-gate	-0.67 [-2.50, +1.17]	+3.33 [+1.83, +5.00]	+2.50 [+0.67, +4.50]	+1.72 [+0.72, +2.78]
delay − no-gate	+1.33 [-0.17, +3.00]	+4.33 [+2.50, +6.33]	+5.67 [+3.33, +8.00]	+3.78 [+2.61, +4.94]
read-once − delay	+1.33 [-0.50, +3.17]	+0.17 [-1.67, +2.00]	+2.33 [+0.00, +4.67]	+1.28 [+0.11, +2.50]
noop − delay	-2.00 [-3.83, -0.17]	-1.00 [-2.50, +0.33]	-3.17 [-5.50, -1.00]	-2.06 [-3.17, -1.00]
read − read-once	+0.17 [+0.00, +0.50]	+0.33 [+0.00, +0.83]	+1.00 [+0.17, +2.00]	+0.50 [+0.17, +0.89]
read − no-gate	+2.83 [+1.17, +4.67]	+4.83 [+2.67, +7.00]	+9.00 [+6.50, +11.67]	+5.56 [+4.33, +6.83]

Tables S7–S8 report additional trigger-row behavioral diagnostics for the original ladder: hand-back rate, the share of trigger rows making zero post-trigger tool calls, the answer-churn decomposition relative to the rejected first answer, and accuracy conditional on *not* hand-backing. NO-GATE has no churn columns because churn is defined relative to the NO-GATE answer on the same question.

Table S7: Trigger-row behavior on the 600-question panel (pooled, all six arms including DELAY_V2), $n = 197$ trigger rows per arm. W/C = wrong/correct under the judge; W→C and C→W count trigger rows whose outcome changes relative to NO-GATE; Acc|non-hand-back = accuracy excluding hand-back rows.

Arm	Hand-back	Zero post-tool	Net churn (W→C−C→W)	C→W	W→C	Acc|non-hand-back	n
no-gate	0.0%	100.0%	—	—	—	33.0%	197
delay (v1, retired)	70.6%	78.7%	-36	54	18	36.2%	197
delay	4.1%	87.8%	7	9	16	38.1%	197
noop	12.7%	0.0%	4	9	13	39.5%	197
read-once	5.1%	3.6%	27	5	32	49.2%	197
read	3.6%	2.0%	29	5	34	49.5%	197

Table S8: Trigger-row behavior on the replication panel (pooled, all five arms; NO-GATE/NOOP_ONCE/RG_ONCE replayed from the three-arm core), $n = 488$ trigger rows per arm. W/C = wrong/correct under the judge; W→C and C→W count trigger rows whose outcome changes relative to NO-GATE; Acc|non-hand-back = accuracy excluding hand-back rows.

Arm	Hand-back	Zero post-tool	Net churn (W→C−C→W)	C→W	W→C	Acc|non-hand-back	n
no-gate	0.2%	100.0%	—	—	—	32.4%	488
delay	3.1%	89.1%	61	13	74	46.3%	488
noop	11.3%	0.2%	27	20	47	42.3%	488
read-once	5.3%	2.0%	64	14	78	48.1%	488
read	2.9%	1.0%	71	15	86	48.3%	488

D THE INTERRUPTION EFFECT: HETEROGENEITY, MODERATOR, AND VALENCE-CONTROL DIAGNOSTICS

The structured delay (DELAY) was worth $+1.17$pp on the 600-question panel and $+5.08$pp on the 1,200-question replication; the READ-ONCE−DELAY increment correspondingly fell from $+3.33$ to $+0.25$pp, so the replication's pre-specified read-once-vs-delay primary was not met (the panel $+3.33$pp [1.33, 5.33] and pooled $+1.28$pp [0.11, 2.50] intervals exclude zero; the replication interval $[-1.17, 1.75]$ does not). Offline diagnostics on the raw trigger rows (197 panel, 488 replication; recomputed from `per_question.jsonl`) bound what the difference is *not*:

- *Exposure.* Trigger rates rose (HotpotQA/2Wiki/MuSiQue 16.5/21.0/61.0% → 30.0/26.8/65.3%), but the wrong-share of triggers was nearly identical (67.0% vs. 67.6%).
- *Delay-arm behaviour.* Hand-back rate 4.1% (8/197) vs. 3.1% (15/488); roughly nine in ten delay trigger rows made no further tool call before answering on both panels.

Table S9: **Behavior after the first gate message.** Percentages over 197 panel / 488 replication trigger rows per arm. Hand-back: the frozen regex proxy on the first 200 answer characters (it can also match substantive answers). Tool-free: no post-trigger tool call; read: at least one successful post-trigger read. NO-GATE receives no message; its hand-back rate is on the accepted baseline answer.

	Panel ($n = 197$)				Replication ($n = 488$)			
Arm	Hand-back	Tool-free	Read	Acc.	Hand-back	Tool-free	Read	Acc.
NO-GATE	0.0	—	—	33.0	0.2	—	—	32.4
DELAY	4.1	87.8	5.6	36.5	3.1	89.1	5.5	44.9
RESTATE	1.5	93.9	2.0	40.1	0.2	92.6	3.5	41.6
NOOP	12.7	0.0	6.1	35.0	11.3	0.2	5.9	37.9
ANY-TOOL	2.5	0.5	18.8	37.6	2.5	0.8	24.2	41.2
SHAM-READ	9.1	3.6	90.9	39.1	7.8	2.0	90.2	44.5
READ-ONCE	5.1	3.6	90.9	46.7	5.3	2.0	90.2	45.5
READ	3.6	2.0	95.9	47.7	2.9	1.0	97.3	46.9

- *Outcome conditional on trigger.* Accuracy on trigger rows, panel → replication: NO-GATE 33.0→32.4%, DELAY 36.5→44.9%, NOOP 35.0→37.9%, READ-ONCE 46.7→45.5%, READ 47.7→46.9%. Per-trigger net gain over NO-GATE: DELAY $+3.6 \rightarrow +12.5$pp (16/9 → 74/13 fixes/breaks), NOOP $+2.0 \rightarrow +5.5$, READ-ONCE $+13.7 \rightarrow +13.1$, READ $+14.7 \rightarrow +14.5$. By dataset the delay's per-trigger net is HotpotQA $-9.1 \rightarrow +9.2$ (33 vs. 120 triggers), 2Wiki $+2.4 \rightarrow +23.4$ (42 vs. 107; 3/2 vs. 27/2 fixes/breaks), MuSiQue $+7.4 \rightarrow +9.6$ (122 vs. 261).
- *Cost.* Mean loops per question NO-GATE/DELAY/READ-ONCE 3.20/3.60/4.05 (panel) and 3.06/3.54/4.09 (replication); mean retrieved tokens 10.2k/10.5k/12.2k and 9.6k/9.9k/11.7k.

The panel HotpotQA and 2Wiki delay cells contain only 33 and 42 triggers, so small-cell variation may contribute to the gap. An offline containment diagnostic (`scripts/analyze_rethink_moderator.py`) nevertheless finds nearly identical gold-answer containment in pre-trigger retrieved-but-unread chunks across panels: 59.4% of 197 vs. 58.6% of 488 triggers, and 52.3% vs. 50.6% among wrong triggers. The delay's per-trigger net gain is also not concentrated where the gold answer had already been retrieved (panel $+4.3$ vs. $+2.5$pp; replication $+10.8$ vs. $+14.9$pp, gold-in vs. not). The read demand shows the same weak pattern ($+15.4/+11.3$ and $+11.2/+15.8$). Thus this measured moderator does not explain the heterogeneity; parametric knowledge or question re-parsing remain possible explanations.

We also ran a post-hoc valence-matched control (`iclr27_restate_valence_2panel_s42`; 1,800 paired rows, $2.53). It uses the same one-shot rejection framing but asks only to restate the final answer. Relative to NO-GATE, it gains $+3.28$pp pooled: $+2.33$ on the panel (14 fixes, 0 breaks; $+7.1$ per trigger) and $+3.75$ on replication (49/4; $+9.2$ per trigger), compared with the structured delay's $+3.78$pp pooled. The restate point estimate is close to the structured interruption estimate. Shared rejection/deferral framing is a possible explanation, but the messages also differ in wording; their ratio does not identify a causal contribution of that framing. The delay-minus-restate difference is $\sim +0.5$pp pooled and remains panel-dependent (-1.17 panel, $+1.33$ replication). Construct checks pass (hand-back 1.5%/0.2% under the frozen first-200-character rule; zero one-shot violations; ∼93% of trigger rows tool-free after the message). The RESTATE arms ran under a later implementation freeze than the frozen DELAY/NO-GATE arms. The agent code gained the restate branch, but gate logic, judge and cache-layer hashes are identical, and non-trigger prefixes replay byte-identically from copied caches; residual drift is therefore confined to post-trigger behavior.

E READ-QUALITY DETAILS

E.1 PROXIES

§L.4 separates *a read occurred* from *useful evidence was inspected* using the first read event in each trigger row's post-trigger tool-trace slice. Two proxies are computed on that chunk's text: (i) *gold containment* — whether the normalized chunk text contains the normalized gold answer string; (ii) *entity overlap* — the fraction of the question's extracted entities (the project's pinned regex entity extractor, the same extractor validated in Appendix J.2) that appear in the chunk text. A read is

low-proxy-relevance when its entity overlap is below 0.2 *and* it does not contain the gold answer. This proxy can miss useful intermediate-hop evidence and does not establish irrelevance. Rows with zero extractable question entities are excluded from the overlap and low-proxy-relevance denominators (reported separately, not folded into either side). A failed read (invalid `evidence_id`, `status != "ok"`) is counted as a compliance attempt but has no resolvable chunk text and is excluded from the relevance denominators rather than scored as irrelevant. The voluntary-read comparison population is every NO-GATE row that read at least one chunk on its own, without any gate ever firing on it. Table S10 is the pooled panel summary cited in §L.4; Tables S11 and S12 give the per-arm, per-dataset proxies on the panel and the replication.

Table S10: Read quality on trigger rows (panel). "Voluntary" = reads the same controller makes on non-trigger rows without a gate.

First read	READ (forced)	READ-ONCE (forced)	voluntary
top-1 of latest search	30.2%	30.2%	62.8%
contains gold answer	50.3%	50.8%	65.0%
low-proxy-relevance (low entity overlap, no gold containment)	20.1%	20.7%	8.7%
accuracy | gold-containing read	73.7%	74.7%	80.9%
accuracy | gold-non-containing read	25.5%	26.1%	57.4%
accuracy with no demanded read (same questions, no intervention)	33.0%		—
accuracy with no demanded read (same questions, DELAY; 5.6% read)	36.5%		—

Table S11: Read quality on trigger rows, by dataset and pooled (mechanism-ladder panel, $n = 197$ trigger rows). 'Low-proxy-relevance': gold answer absent and question-entity overlap below 0.2; this does not establish irrelevance. acc|rel./acc|irrel.: accuracy when the first read does / does not contain the gold answer.

Arm	Dataset	n trigger	n w/ read	gold-contain	mean overlap	low-proxy-relevance	acc|rel. / acc|irrel.
read	HotpotQA	33	33	69.7%	0.523	12.1%	82.6% / 30.0%
read	2WikiMultihop	42	42	52.4%	0.421	38.1%	72.7% / 55.0%
read	MuSiQue	122	114	43.9%	0.553	15.8%	70.0% / 15.6%
read-once	HotpotQA	33	33	69.7%	0.523	12.1%	82.6% / 30.0%
read-once	2WikiMultihop	42	40	50.0%	0.403	40.0%	70.0% / 55.0%
read-once	MuSiQue	122	106	45.3%	0.542	16.0%	72.9% / 15.5%
read	**pooled**	197	189	50.3%	0.513	20.1%	73.7% / 25.5%
read-once	**pooled**	197	179	50.8%	0.502	20.7%	74.7% / 26.1%
voluntary (norg, non-trigger)	**pooled**	403	403	65.0%	0.721	8.7%	80.9% / 57.4%

Table S12: Read quality on trigger rows, by dataset and pooled (replication panel, $n = 488$ trigger rows). 'Low-proxy-relevance': gold answer absent and question-entity overlap below 0.2; this does not establish irrelevance. acc|rel./acc|irrel.: accuracy when the first read does / does not contain the gold answer.

Arm	Dataset	n trigger	n w/ read	gold-contain	mean overlap	low-proxy-relevance	acc|rel. / acc|irrel.
read-once	HotpotQA	120	116	72.4%	0.730	6.0%	79.8% / 50.0%
read-once	2WikiMultihop	107	102	54.9%	0.586	20.6%	73.2% / 41.3%
read-once	MuSiQue	261	222	37.6%	0.490	17.6%	54.2% / 21.0%
read-once	**pooled**	488	440	50.8%	0.580	15.3%	68.6% / 29.6%
voluntary (norg, non-trigger)	**pooled**	712	712	66.9%	0.750	6.5%	82.1% / 57.6%

E.2 BLINDED TWO-MODEL ANNOTATION AUDIT

Two blinded annotation audits (two different AI systems as raters, author-adjudicated) validate the automated proxies used throughout the paper. Both use the same protocol: a rater sees only the question, the gold answer, and the text to be labeled (arm, dataset and run identity are withheld), rows are not reorderable, and labels are matched back by a stable `item_id`.

- **Hand-back detector** (`docs/experiments/manual_audit/hand_back_sheet.md`), $n = 60$: for each accepted final answer, label `HAND_BACK` (the answer only acknowledges, asks a clarifying question, or asks what to do next) or `ANSWER` (a substantive answer, even if wrong).

This validates the frozen HAND_BACK_PATTERN regex behind every hand-back rate reported in §4.1, §4 and Appendix C.

- **Read-relevance proxies** (docs/experiments/manual_audit/read_relevance_sheet.md), $n = 60$: for each trigger row, the rater sees the question, the gold answer and the full text of the chunk returned by the first induced read, and labels it RELEVANT (contains the answer or a fact that directly supports it), PARTIAL (a question or answer entity is present but not the answer) or IRRELEVANT. This validates the gold-containment and entity-overlap proxies behind §L.4; items are stratified across arms and benchmarks and drawn by a seeded sampler (scripts/build_manual_audit_sheets.py); scoring is by scripts/score_manual_audit.py against hidden detector keys.

Results (two independent raters, adjudicated). Both sheets ($n = 60$ each, blinded: arm, dataset and detector verdicts hidden) were labeled by two independent raters — a first pass by an AI assistant reading the full chunk text, and a second, blind pass by a different AI system that saw neither the first labels nor any detector output — and adjudicated by the authors with a pre-stated rule: disagreements resolve to the *stricter* label (the one less favorable to the paper's read-quality claims). Inter-rater agreement before adjudication: hand-back $\kappa = 0.965$ (59/60), read-relevance $\kappa = 0.799$ (52/60). Final adjudicated labels: hand-back 36 HAND_BACK / 24 ANSWER; read-relevance 19 RELEVANT / 27 PARTIAL / 14 IRRELEVANT. Detector validity against the final labels: the hand-back regex (§4.1) has accuracy 0.83, $\kappa = 0.67$ — its misses are turns that *announce a procedure* without answering (8/60), which the raters count as hand-backs, so the regex under-counts; the gold-containment proxy (§L.4) has accuracy 0.82, $\kappa = 0.58$ as a binary relevance detector (precision 0.70, recall 0.74 for RELEVANT); the entity-overlap proxy $\kappa = 0.54$ ($n = 56$); and the three-way proxy bucket $\kappa = 0.53$. The raters call a chunk relevant more often than either proxy when it pins a sub-hop of a multi-hop question without containing the final answer string, so the proxies are conservative for "useful evidence inspected." Files: results/_manual_audit/ (scores, inter-rater and adjudication records) and the filled sheets in docs/experiments/manual_audit/.

F EXPERIMENT INVENTORY AND SAMPLE SIZES

Table S13 lists every paid run generated for this submission before the answer-obligation studies (Appendix K itemises those). The 15,000 trajectories of §3 are the five ladder arms and three side controls on the 1,800 matched IDs ($8 \times 1{,}800 = 14{,}400$) plus the 600 retired v1-delay trajectories; the medium-effort and second-family runs are released in addition. Every agent run was initially judged by gpt-5-mini and executed under a frozen manifest with per-arm cost guards; the final row is a targeted Gemini judge substitution. Costs are new API spend including judge calls. The pooled decomposition rests on the first four rows; medium-effort, valence-control and second-family rows support the scoped claims cited at their point of use; the rest are prevalence or boundary cells under the sampling rule of §3.

Table S13: Experiment inventory: every paid run generated for this submission before the answer-obligation studies (those are itemised in Appendix K). n is per benchmark; "panel IDs" = the ladder panel's seeded first-200 question IDs per benchmark, targeted; where rows were filtered or a run was stopped, the n column gives the valid denominator.

Run	Controller	Benchmarks	n (total)	Arms	Cost
Mechanism ladder	gpt-5-mini minimal	HotpotQA, 2Wiki, MuSiQue	200 (600)	NO-GATE, NOOP, READ-ONCE, READ, DELAY (v1, retired)	$1.55
Delay-v2	gpt-5-mini minimal	same 600 questions	200 (600)	DELAY (structured)	$0.91
Replication core	gpt-5-mini minimal	same, disjoint panel	400 (1,200)	NO-GATE, NOOP, READ-ONCE	$15.48
Replication full ladder	gpt-5-mini minimal	same 1,200 questions	400 (1,200)	DELAY, READ (+3 replayed)	$2.89
Medium-effort 2-arm	gpt-5-mini medium	600 panel questions	200 (600)	NO-GATE, READ-ONCE	$24.46
Luna prevalence	gpt-5.6-luna (effort none)	HotpotQA, 2Wiki, MuSiQue (panel IDs)	200 (600)	NO-GATE	$6.78
Qwen3.7-plus prevalence	qwen3.7-plus-2026-05-26	HotpotQA (panel IDs)	200	NO-GATE	$3.35
Qwen3.7-plus prevalence	same	2Wiki, MuSiQue (panel IDs; 2 and 22 filtered rows)	198, 178 valid	NO-GATE	$15.56
Qwen3.7-flash scout	qwen3.7-flash-2026-07-15	HotpotQA; MuSiQue provider-stopped	99 valid; 28 attempted (24 valid)	NO-GATE	$1.01
Medical / Novel scouts	gpt-5-mini minimal	GraphRAG-Bench Medical / Novel (Xiang et al., 2025)	200 / 200	NO-GATE	$0.65 + $0.66
RESTATE valence control	gpt-5-mini minimal	both panels (same IDs)	200+400 (1,800)	RESTATE	$2.53
Any-tool control	gpt-5-mini minimal	both panels (same IDs)	200+400 (1,800)	ANY-TOOL	$5.74
Sham-read control	gpt-5-mini minimal	both panels (same IDs)	200+400 (1,800)	SHAM-READ	$3.80
Second-family core	Qwen2.5-7B (local vLLM)	HotpotQA, 2Wiki, MuSiQue (panel IDs)	200 (600; 541 valid pairs)	NO-GATE, NOOP, READ-ONCE	$0.24 (judge; agent local)
Cross-family judge re-check	Gemini 2.5 Pro (judge only)	678 discordant answers	—	—	$5.98
Total					**$91.59**

G PREVALENCE, CROSS-FAMILY BOUNDARY, AND DOMAIN/MODEL SCOUTS

G.1 FULL PREVALENCE MAP

Controller provenance. Provider documentation identifies the API controllers and their available snapshots: gpt-5-mini (OpenAI, 2026a), gpt-5.6-luna (OpenAI, 2026b), Qwen3.7 (Alibaba Cloud, 2026) and Gemini 2.5 Flash (Google, 2026). Prices and snapshot identifiers used for the cost figures are frozen in `experiments/model_catalog.toml` with their access dates.

Table S14 is the complete controller×dataset zero-read prevalence table behind Figure 3, including pooled rows and every registry citability flag. Its source is `results/_framework_tier1/capability_boundary.json`, built by `scripts/build_capability_boundary.py`.

Table S14: Full zero-read prevalence table: 32 rows including pooled summaries, descending by strict rate. *Strict* counts every zero-read final; *loop-cap* is the subset forced at the loop limit, which no final-time gate can intercept (— indicates unavailable per-question rows with the strict count carried); *interceptable* is their difference; Figure 3 plots the 22 non-pooled controller×dataset rows, excluding the ladder-panel duplicate rows and Qwen3.7-flash. Exact Clopper–Pearson 95% intervals are shown. Counts are strict/loop-cap/n. All rows are registry-verified (§G.1).

Controller	Setting	Dataset	strict/loop-cap/n	strict rate, % [95% CI]	interceptable, % [95% CI]
qwen2.5-3b-instruct	3B	2Wiki	187/3/200	93.5 [89.1, 96.5]	92.0 [87.3, 95.4]
qwen2.5-3b-instruct	3B	pooled	492/8/600	82.0 [78.7, 85.0]	80.7 [77.3, 83.8]
qwen2.5-3b-instruct	3B	MuSiQue	160/5/200	80.0 [73.8, 85.3]	77.5 [71.1, 83.1]
qwen2.5-3b-instruct	3B	HotpotQA	145/0/200	72.5 [65.8, 78.6]	72.5 [65.8, 78.6]
gpt-5-mini-minimal	minimal	MuSiQue	655/0/1000	65.5 [62.5, 68.4]	65.5 [62.5, 68.4]
gpt-5-mini-minimal (ladder panel)	minimal	MuSiQue	122/0/200	61.0 [53.9, 67.8]	61.0 [53.9, 67.8]
gemini-2.5-flash	budget 1024	HotpotQA	580/0/1000	58.0 [54.9, 61.1]	58.0 [54.9, 61.1]
gpt-5-mini-minimal (medical scout)	minimal	Medical	111/0/200	55.5 [48.3, 62.5]	55.5 [48.3, 62.5]
gemini-2.5-flash	budget 0	HotpotQA	523/0/1000	52.3 [49.2, 55.4]	52.3 [49.2, 55.4]
gpt-5-mini-minimal	minimal	pooled	1137/0/3000	37.9 [36.2, 39.7]	37.9 [36.2, 39.7]
gpt-5-mini-minimal (ladder panel)	minimal	pooled	197/0/600	32.8 [29.1, 36.8]	32.8 [29.1, 36.8]
qwen2.5-7b-instruct	7B	2Wiki	59/0/200	29.5 [23.3, 36.3]	29.5 [23.3, 36.3]
qwen2.5-7b-instruct	7B	pooled	167/0/600	27.8 [24.3, 31.6]	27.8 [24.3, 31.6]
qwen2.5-7b-instruct	7B	HotpotQA	55/0/200	27.5 [21.4, 34.2]	27.5 [21.4, 34.2]
qwen2.5-7b-instruct	7B	MuSiQue	53/0/200	26.5 [20.5, 33.2]	26.5 [20.5, 33.2]
gpt-5-mini-minimal	minimal	2Wiki	264/0/1000	26.4 [23.7, 29.2]	26.4 [23.7, 29.2]
gpt-5-mini-minimal	minimal	HotpotQA	218/0/1000	21.8 [19.3, 24.5]	21.8 [19.3, 24.5]
gpt-5-mini-minimal (ladder panel)	minimal	2Wiki	42/0/200	21.0 [15.6, 27.3]	21.0 [15.6, 27.3]
qwen3.7-flash-2026-07-15	flash, non-thinking	HotpotQA	18/1/99	18.2 [11.1, 27.2]	17.2 [10.3, 26.1]
gpt-5-mini-minimal (ladder panel)	minimal	HotpotQA	33/0/200	16.5 [11.6, 22.4]	16.5 [11.6, 22.4]
gpt-5-mini-minimal (novel scout)	minimal	Novel	17/0/200	8.5 [5.0, 13.3]	8.5 [5.0, 13.3]
gpt-5.6-luna	effort none	2Wiki	15/9/200	7.5 [4.3, 12.1]	3.0 [1.1, 6.4]
gpt-4o-mini	n/a	MuSiQue	64/17/1000	6.4 [5.0, 8.1]	4.7 [3.5, 6.2]
gpt-5.6-luna	effort none	pooled	32/15/600	5.3 [3.7, 7.4]	2.8 [1.7, 4.5]
gpt-5.6-luna	effort none	HotpotQA	9/0/200	4.5 [2.1, 8.4]	4.5 [2.1, 8.4]
gpt-5.6-luna	effort none	MuSiQue	8/6/200	4.0 [1.7, 7.7]	1.0 [0.1, 3.6]
qwen3.7-plus-2026-05-26	plus, non-thinking	MuSiQue	7/6/178	3.9 [1.6, 7.9]	0.6 [0.0, 3.1]
gpt-4o-mini	n/a	pooled	105/23/3000	3.5 [2.9, 4.2]	2.7 [2.2, 3.4]
gpt-4o-mini	n/a	2Wiki	28/6/1000	2.8 [1.9, 4.0]	2.2 [1.4, 3.3]
qwen3.7-plus-2026-05-26	plus, non-thinking	HotpotQA	3/0/200	1.5 [0.3, 4.3]	1.5 [0.3, 4.3]
gpt-4o-mini	n/a	HotpotQA	13/0/1000	1.3 [0.7, 2.2]	1.3 [0.7, 2.2]
qwen3.7-plus-2026-05-26	plus, non-thinking	2Wiki	2/1/198	1.0 [0.1, 3.6]	0.5 [0.0, 2.8]

Figure 3 plots the *interceptable* rate: strict zero-read finals minus forced finals at the loop cap (zero reads after the maximum number of loops; `scripts/build_interceptable_prevalence.py`), which a final-time gate cannot intercept. Table S14 lists strict, loop-cap and interceptable counts for every cell. Loop-cap finals are absent at minimal effort on the ladder panels (0 of 685 zero-read rows) but make up a share of zero-read rows for some controllers (gpt-4o-mini MuSiQue 17/64; gpt-5.6-luna MuSiQue 6/8 and 2Wiki 9/15; Qwen3.7-plus MuSiQue 6/7 and 2Wiki 1/2; most zero-read rows of gpt-5-mini at medium effort; §H.1). All rows shown are verified/citable in the run registry. The gpt-5.6-luna $n = 200$ cells supersede the earlier $n = 50$ smokes.

G.2 Cross-family boundary cells and the payoff identity

Table S15 evaluates the trigger-payoff identity of §6.2 on four cells. The two in-harness rows are the gpt-5-mini ladder cell and the Qwen2.5-7B second-family core (Appendix H.2), both on the panel IDs in the released scaffold and comparable to each other. The two archived rows (Gemini 2.5 Flash, Qwen3.5-9B) are boundary diagnostics only and are *not* matched to them: (a) harness — they ran through separate standalone scripts; (b) trigger — their zero-read rows are induced by a `thinking_budget`/thinking-mode intervention layered on their own baselines, so the gate there rescues a different failure population than in the causal core; (c) mechanism — a forced-read-until-satisfied loop rather than the one-shot READ-ONCE rung; (d) loop-cap saturation — up to 41% of the Qwen3.5 gate rows hit an observed loop cap, so some "recoveries" may be forced answers rather than settled ones; (e) judge pinning — the undated `gpt-5-mini` string rather than the dated snapshot; (f) coverage — HotpotQA only. In the run registry the Gemini NO-GATE baseline is verified/citable, the Gemini gate follow-up is archived (a partial targeted follow-up, not a full paired cell), and the Qwen3.5 run group is archived ("exploratory thinking-budget and parser smokes; not a complete matched experiment"). We report them because each independently instantiates a high-prevalence,

mostly-correct-skip regime and the identity holds exactly on each (decomposition check `passed: true` in the source JSON), but we do not pool them and promote neither to a causal claim; the full cell detail is in `results/_framework_tier1/cross_family_boundary.json`.

Table S15: The trigger-payoff identity: $\Delta_{\text{trig}} = w\,r - (1-w)\,d$. The first two rows are in-harness ladder cells on the panel IDs, recomputed from raw trajectories by `scripts/analyze_second_family_payoff.py`, whose self-check reproduces the gpt-5-mini row before emitting any other. Gemini and Qwen3.5 are boundary diagnostics on archived targeted-gate rows (induced triggers, different loop); see the caveats above.

Controller (dataset)	triggers	$1-w$ (skips already right)	r	d	Δ_{trig} [95% CI]
In-harness ladder cells (same scaffold, panel IDs; comparable to each other)					
gpt-5-mini minimal, READ-ONCE (3 benchmarks)	197	33.0%	24.2%	7.7%	+13.7 [8.1, 19.3]
Qwen2.5-7B, READ-ONCE (3 benchmarks)	114	48.2%	6.8%	9.1%	−0.9 [−6.1, 4.4]
Archived boundary diagnostics (different harness and trigger; not matched to the above)					
Gemini 2.5 Flash, $\theta = 1024$ (HotpotQA)	580	62.9%	29.3%	14.2%	+1.9 [−1.7, 5.5]
Qwen3.5-9B thinking (HotpotQA)	100	51.0%	20.4%	17.6%	+1.0 [−8.0, 9.0]

Gemini break decomposition (representative detail). On the Gemini cell, roughly half of the breaks (correct-under-NO-GATE, wrong under the archived gate follow-up) are a direct zero-read transition (HotpotQA 41.9%) and roughly half retain the same read count but produce a different wrong answer (50.0%). A representative break: for *"What is the genus of the viral disease that has symptoms such as fever, chills, loss of appetite, nausea, muscle pains, and headaches, and has a chance of causing liver damage?"* (gold: *Flavivirus*), the thinking-disabled trajectory searches, reads a chunk linking yellow-fever symptoms to the genus, and answers correctly; with a 1024-token thinking budget the model emits its internal thinking tokens, searches once, and finalizes with zero reads, restating the symptoms without ever naming the genus.

G.3 Domain and cross-model scouts

Promotion rule and why its thresholds are what they are. A scout or prevalence cell is promoted to a full three-benchmark ladder only if its trigger (zero-read) prevalence is $\geq 15\%$ *and* the wrong-share of its triggers is $\geq 40\%$. These thresholds are screening heuristics, not payoff predictions or power guarantees for an unseen controller. In measured regimes, the trigger-payoff identity gives the observed break-even wrong-share $w^* = d/(r+d)$: 24% for gpt-5-mini ($r = 24.2$, $d = 7.7$) and 33% for Gemini 2.5 Flash ($r = 29.3$, $d = 14.2$). We set the 40% screen above both observed values while requiring enough opportunities for the gate to matter; a 15% trigger rate yields at least 30 trigger rows per 200-question benchmark. The 600-question core is sized for the payoff identity rather than for an ITT test: at 100–200 triggers it yields intervals of about ±5–6pp on Δ_{trig} (Table S15), which is what the deployment decision reads. Scouts use $n = 200$ so the threshold decisions are better resolved than at $n = 100$, where the Medical wrong-share and Novel trigger-rate intervals both crossed their respective screens.

Why a cell may not have $n = 200$. New prevalence and scout cells target the ladder panel's seeded first-200 IDs per benchmark. A cell is reported at a different denominator for exactly one of four reasons, each recorded in the run registry: (i) deterministic provider content-filter rows retained as errors (Qwen3.7-plus: 198 and 178 valid rows on 2Wiki and MuSiQue); (ii) a provider-side stop under the cost ceiling (Qwen3.7-flash MuSiQue, 24 valid rows, excluded from the map); (iii) local tool-call error rows retained rather than re-run (Qwen2.5-7B core: 541 valid paired rows of 600); (iv) inherited cells from the earlier grid, reported at their original $n = 1{,}000$. Rows replayed from a smaller earlier smoke on the same IDs form part of the $n = 200$ cell and are not a deviation.

Medical (verified, complete). A GraphRAG-Bench Medical fact-retrieval scout ($n = 200$ on the seeded first-200 IDs, gpt-5-mini minimal; the first 100 rows replay the original scout; `iclr27_domain_scout_medical_norg_n200_s42`) yields zero-read $111/200 = 55.5\%$ [48.3, 62.5]%, with $35/111 = 31.5\%$ [23.0, 41.0]% wrong among zero-read rows (share already correct 68.5%); LLM-Acc is 68.5%. The pre-specified promotion rule requires both trigger $\geq 15\%$ and wrong-share $\geq 40\%$. Medical meets the former but not the latter on its point estimate, so it is recorded as a

high-prevalence, mostly-correct-skip boundary point and is not promoted. No gate arm was run in this scout, so it estimates no Medical payoff; the 800-question study of §5 does (Table 4).

Novel (verified, complete). A GraphRAG-Bench Novel-domain scout ($n = 200$, first 100 rows replayed from the original scout; `iclr27_domain_scout_novel_norg_n200_s42`): zero-read $17/200 = 8.5\%$ [5.0, 13.3]%; wrong among zero-read $11/17 = 64.7\%$ (skip-worked 35.3%); LLM-Acc 54.5%. Wrong-share was met but trigger prevalence was not (the 15% threshold lies outside the interval), so Novel is a boundary point, not a ladder.

Qwen3.7-plus ($n = 200$ cell, descriptive). A HotpotQA NO-GATE prevalence cell (`iclr27_qwen37plus_norg_hotpot_n200_s42`; qwen3.7-plus-2026-05-26 via DashScope, non-thinking mode; seed-42 first 200 questions, first 50 rows replayed from the $n = 50$ smoke, which it supersedes; $3.35 under an amended $4.00 ceiling): zero-read $3/200 = 1.5\%$ [0.3, 4.3]%, all three skips correct (wrong-share 0%), LLM-Acc 90.5%, 4.4 loops and 1.2 reads per question. The earlier paired $n = 50$ smoke (NO-GATE vs. RG, both 92.0%, 0 fixes / 0 breaks) remains a compatibility check only. The 2WikiMultihop and MuSiQue cells (`iclr27_qwen37plus_norg_2data_n200_s42`, same IDs and settings; completed under an amendment after a first stop at the original guards) attempted all 200 rows each and retain deterministic provider content-filter rows as errors (DashScope `data_inspection_failed`; 2 rows on 2Wiki, 22 on MuSiQue — retrying them after a balance top-up reproduced the same code, so the filter, not billing, binds): 2Wiki $2/198 = 1.0\%$ [0.1, 3.6]% (wrong-share 0%; LLM-Acc 86.4%) and MuSiQue $7/178 = 3.9\%$ [1.6, 7.9]% (six of seven wrong, six of seven loop-cap exits — interceptable $1/178$; LLM-Acc 59.6%); MuSiQue trajectories cost $0.05 per row, $2.3\times$ the HotpotQA rate. No benchmark meets the promotion rule. Strict zero-read rates span 1.0–3.9%; after excluding loop-cap exits, the prevalence map reports 0.5–1.5%.

gpt-5.6-luna ($n = 200$ cells, descriptive). Three NO-GATE prevalence cells at reasoning effort `none` (`iclr27_luna_norg_2data_n200_s42` and `. . . _2wiki_n200_s42`; the ladder panel's seeded first-200 IDs per benchmark; HotpotQA/MuSiQue first 50 rows replayed from the earlier $n = 50$ smokes, which these cells supersede): zero-read $9/200 = 4.5\%$ [2.1, 8.4]% on HotpotQA (wrong-share of skips 22%; LLM-Acc 82.5%), $15/200 = 7.5\%$ [4.3, 12.1]% on 2WikiMultihop (9 of the 15 loop-cap exits; wrong-share 33%; LLM-Acc 70.5%) and $8/200 = 4.0\%$ [1.7, 7.7]% on MuSiQue (wrong-share 100%, all eight skips wrong; LLM-Acc 50.0%); pooled $32/600 = 5.3\%$. None of the three cells meets the pre-specified promotion rule (prevalence $\geq 15\%$ and wrong-share $\geq 40\%$), so Luna remains a low-prevalence boundary point with no ladder. Six of the eight MuSiQue zero-read rows are loop-cap exits (ten searches, no read, forced final) rather than premature `final` calls, so the interceptable trigger rate there is $2/200$; all six loop-cap rows are wrong.

Qwen3.7-flash (provider-stopped partial cell). Reported here for provider compatibility, not as a prevalence cell, and omitted from Figure 3: a two-dataset NO-GATE scout (`iclr27_qwen37flash_scout_2data_norg_n100_s42`; non-thinking, DashScope International endpoint, $1.01) gives HotpotQA zero-read 18/99 = 18.2% (exact 95% CI [11.1, 27.2]%) with a wrong-share of triggers of 16.7% and LLM-Acc 85%, i.e. the rule for promoting a ladder (trigger $\geq 15\%$ and wrong-share $\geq 40\%$) is not met. The MuSiQue cell stopped at 28 rows under its $0.50 ceiling (conservative upper-tier pricing) with 4 provider content-inspection errors (`DataInspectionFailed`) on Wikipedia chunks; with 24 valid rows it is underpowered and, because the provider filter is deterministic, paired arms would be broken on this endpoint. Qwen3.7-flash therefore appears only in the appendix table as a HotpotQA compatibility point. It is excluded from the prevalence figure and its 22-cell count, and it is not a cross-family causal cell.

H Cost frontier and the second-family core

H.1 Cost frontier (inherited grid)

Should one raise reasoning effort rather than gate? Figure S2 places the cells on an accuracy-vs-cost plane (gpt-5-mini, pooled over the three benchmarks, $n = 1{,}000$ per dataset; Read-Gate costs are replay-corrected, because cache-served calls were recorded at $0; see the replay-corrected costs below): no intervention at minimal effort reaches 59.4% at $0.0064 per question; adding the Read-Gate reaches 65.4% ($+5.97$pp) at $0.0095, about $1.5\times$ the no-gate cost, while medium effort reaches

82.2% at $0.036. The gate recovers 26.3% of the minimal-to-medium accuracy gap while costing 26.3% as much in total per-question spend. Relative to the minimal no-gate baseline, its incremental cost is 10.5% of the minimal-to-medium cost increment (accuracy per dollar 69 vs. 23 for medium; 93 for no intervention). This is a within-controller trade: gated minimal remains 16.9pp below medium. The gpt-4o-mini no-gate point is both cheaper ($0.00796/question) and more accurate (66.1%) than gated minimal. No joint uncertainty or test of economic dominance was computed, and the costs retain the frozen proxy definition below. At medium effort the one-shot read demand has a small observed effect: +0.17pp on the same 600 panel questions (2 fixes, 1 break; per benchmark 0.0, 0.0, +0.5), because premature finals become rare (20 of 600 trigger; most zero-read rows there are loop-cap exits no final-time gate can intercept).

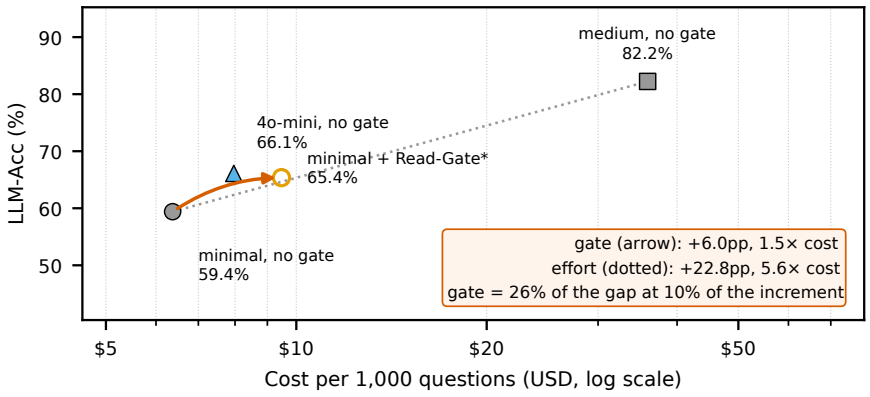

Figure S2: Accuracy vs. cost per question on the inherited grid (gpt-5-mini at minimal and medium effort, with gpt-4o-mini as a controller alternative; each pooled over three benchmarks; $n = 1{,}000$ per dataset; hollow markers are replay-corrected Read-Gate cells; the dotted line is the minimal-to-medium effort trade, and the box restates both trades from the plotted cells). The medium-effort gate estimate is the clean $n = 600$ paired cell reported below and in §8, and is not plotted. Costs are per-question totals: 0.0095 vs. 0.0359 is 26.3%; the gate's incremental cost over minimal no-gate is 10.5% of the minimal-to-medium increment.

Medium effort, clean paired estimate ($n = 600$). The medium-effort statement in §8 comes from `iclr27_medium_gate_2arm_3data_n200_s42`: gpt-5-mini at `reasoning_effort=medium`, NO-GATE vs. READ-ONCE on the 600 ladder-panel questions ($24.46). ITT READ-ONCE−NO-GATE = +0.17pp, 95% CI $[-0.33, +0.67]$ (2 fixes, 1 break; McNemar $p = 1$); per benchmark 0.00, 0.00, +0.50. The one-shot demand fired on only 20 of 600 rows (1.0/2.5/6.5%); a further 46 zero-read rows were loop-cap forced finals (ten searches, no read, no `final` call), which a final-time gate cannot intercept, and the trigger definition used by every analysis in this paper excludes them (none occur at minimal effort). NO-GATE accuracy 90.0/86.0/68.0% at $0.040 per question. This supersedes the earlier paired $n = 100$ medium+RG probe (−3.67pp), which is retained in the artifact only; no inference in this paper uses it.

Table S16 is the full per-dataset detail behind Figure S2, transcribed from the `proxy $` column of `docs/experiments/cost_accuracy_frontier.md` (itself generated offline by `scripts/build_cost_accuracy_frontier.py` from the per-question result files cited in its own provenance section; no number below is re-derived).

Table S16: Cost–accuracy detail by dataset and scenario, gpt-5-mini unless noted. RG = Read-Gate; LLM-Acc = accuracy under the gpt-5-mini judge; Acc/$ = LLM-Acc per dollar per question ($\times 1000$). Costs are the replay-corrected values used by Figure S2 and the main text (recorded costs, which understate the Read-Gate cells, and the correction method are in the artifact's `cost_accuracy_frontier.md`). All rows are the full $n = 1{,}000$ population.

Dataset	Scenario	n	LLM-Acc	mean loops	mean reads	corrected $/q	Acc/$ ($\times 1000$)
HotpotQA	minimal, no gate	1000	79.6%	3.07	0.81	$0.00521	152.7
HotpotQA	minimal, +RG	1000	82.8%	3.63	1.03	$0.00620	133.5
HotpotQA	medium, no gate	1000	90.8%	4.07	1.12	$0.01698	53.5
HotpotQA	gpt-4o-mini, no gate	1000	81.6%	4.46	2.00	$0.00567	143.9
2WikiMultihop	minimal, no gate	1000	64.4%	3.54	0.77	$0.00823	78.3
2WikiMultihop	minimal, +RG	1000	69.7%	4.32	1.06	$0.01083	64.4
2WikiMultihop	medium, no gate	1000	88.8%	5.34	1.14	$0.03389	26.2
2WikiMultihop	gpt-4o-mini, no gate	1000	70.1%	4.86	2.36	$0.00802	87.4
MuSiQue	minimal, no gate	1000	34.2%	2.90	0.35	$0.00568	60.2
MuSiQue	minimal, +RG	1000	43.6%	4.97	1.00	$0.01139	38.3
MuSiQue	medium, no gate	1000	67.1%	7.02	1.16	$0.05693	11.8
MuSiQue	gpt-4o-mini, no gate	1000	46.6%	6.47	2.55	$0.01020	45.7

Replay-corrected costs and pooled Pareto reading. Recorded per-row cost is $0 whenever a call is served from the deterministic cache, and the Read-Gate cells re-use the no-gate cells' prefixes: 78%/74%/35% of minimal+RG rows (HotpotQA/2Wiki/MuSiQue) have $0 recorded cost, against 0–10% for the no-gate cells, so recorded spend understates the gate's cost. `cost_accuracy_frontier.md` (section 3) documents the correction: $0 rows are replaced by the matched no-gate row's real cost (their traces are byte-identical). Pooled: minimal no-gate 59.4% at $0.00637/q (93.2 acc-pts/$$\times 1000$); minimal+RG 65.4% at $0.00947/q (69.0); medium no-gate 82.2% at $0.03594/q (22.9); gpt-4o-mini no-gate 66.1% at $0.00796/q (83.0). Minimal+RG costs about $1.5\times$ minimal no intervention and about a quarter of medium effort. Its accuracy-per-dollar ratio exceeds medium's but is below those of minimal no-gate and gpt-4o-mini. This ratio does not select a universally best deployment: an accuracy target, latency and the cost proxy matter.

H.2 Second-family core (Qwen2.5-7B, exploratory)

A 3-arm core (NO-GATE/NOOP/READ-ONCE; `iclr27_qwen25_7b_core3_3data_n200_s42`) on Qwen2.5-7B-Instruct (Qwen Team, 2024), served locally in the same harness on the ladder panel's seeded IDs ($n = 200$ per benchmark; 6–29 local tool-call error rows per cell retained; 541 valid paired rows, 114 triggers $= 21\%$; matched-trigger and one-shot audits pass; $0.24 judge cost, agent inference local), gives a directionally consistent but non-significant exploratory estimate of the evidence-over-action contrast: $+0.74$pp ITT (9 fixes vs. 5 breaks; two-sided sign test $p = 0.42$; per-trigger $+3.51$pp), positive on HotpotQA ($+1.04$) and 2WikiMultihop ($+1.13$) and zero on MuSiQue, while neither one-shot arm beats no intervention (NOOP$-$NO-GATE $= -0.92$pp, 0 fixes vs. 5 breaks; READ-ONCE$-$NO-GATE $= -0.18$pp, 4 vs. 5). The point estimate neither confirms nor refutes the gpt-5-mini result, and we never pool it with the gpt-5-mini cells. Evaluating the payoff identity on this cell (Table S15; `scripts/analyze_second_family_payoff.py`, on the 114 triggers common to all three arms; this run's own no-gate arm triggers on $114/541 = 21.1\%$, distinct from the 27.8% legacy prevalence cell for the same controller in Figure 3, which is a different run) puts both of its terms against the gate: 48.2% of the 7B's skipped answers were already correct, against 33.0% for the primary controller, and it recovers on only $r = 6.8\%$ of the wrong ones against 24.2%, while disrupting $d = 9.1\%$ of the correct ones against 7.7%. The product is $\Delta_{\text{trig}} = -0.9$pp $[-6.1, 4.4]$. Two caveats bound what this explains: (w, r, d) were computed for this controller only after its arms had run, so the identity accounts for the null rather than predicting it, and r is a property of the controller-plus-corpus cell, not a transferable constant. The practical consequence is recorded in §6 — the pre-specified promotion screen reads prevalence and wrong-share, both of which this cell clears, and neither carries the information that decided the outcome.

I ADAPTIVE-GATE NEGATIVE RESULT AND THE V1 DELAY ARTIFACT

I.1 ADAPTIVE GATE: FULL OFFLINE EVALUATION

This is a frozen, retrospective, zero-API evaluation (Table S17; `results/_adaptive_gate/offline_v1.json`, `docs/experiments/adaptive_gate_offline_v1.md`): 1,946 paired attempted-finalization rows (1,497 unique questions) across the three benchmarks, with saved NO-GATE and READ-GATE arms as independent paired rollouts (not two continuations cloned from one prefix, so per-question deltas are noisy realized contrasts, not an oracle individual-treatment label). A logistic policy is fit under leave-one-dataset-out (primary) and leave-one-model-family-out (cross-family) splits, selecting an intervention threshold under an explicit utility correct $- 0.005 \times$ extra_loops $- 0.0005 \times$ extra_retrieved_tokens$/1000$.

Table S17: Adaptive-gate policies, leave-one-dataset-out (primary), macro-averaged over folds. *never*/*always*: fixed no-gate / gate-on-every-trigger policies; *loop_heuristic* and *repetition_heuristic*: threshold rules on the loop count and on the count of repeated queries and repeated returned chunks, thresholds selected on the training folds; at the selected thresholds both fire on every trigger and coincide with *always*.

Policy	Accuracy	Utility	Gate rate	Extra loops	Extra retr. tokens
never	0.341	0.341	0.000	0.00	0.0
always	0.433	0.416	1.000	3.04	2371.3
adaptive	0.432	0.418	0.856	2.59	2128.4
loop_heuristic	0.433	0.416	1.000	3.04	2371.3
repetition_heuristic	0.433	0.416	1.000	3.04	2371.3
rate-matched random (100 seeds)	0.418	0.404	0.855	2.60	2019.7

Pre-specified decision: FAIL. Clustered bootstrap 95% CI for adaptive minus the better fixed policy's utility is $[-0.0046, +0.0088]$ (crosses zero); loop reduction vs. ALWAYS is 14.6% and retrieved-token reduction is 10.2%, both below the pre-specified 20% success criterion. Of the seven frozen checks, three fail: `bootstrap_ci_beats_better_fixed`, `loops_reduced_at_least_20pct`, and `tokens_reduced_at_least_20pct`; four pass, including `point_utility_beats_both_fixed` and `beats_rate_matched_random` — the adaptive policy is not *worse* than either fixed policy at the point estimate, it simply does not clear the pre-specified bar for replacing them. **Cross-family transfer is weak**: under leave-one-model-family-out splitting (holding out OpenAI or Qwen2.5 entirely), macro adaptive utility falls to 0.341 — statistically indistinguishable from NEVER (0.341) and well below ALWAYS (0.416) — i.e. the learned policy collapses toward doing nothing when the target family is unseen during fitting, rather than generalizing the always-gate rule it approximates in-family. Failure action (as recorded): do not use adaptive gating as the ICLR method headline; retain as a transparent retrospective boundary result in this appendix.

I.2 THE V1 DELAY ARTIFACT

The original open-ended delay control (`NEUTRAL_DELAY_MESSAGE`, Appendix A.1) was flagged in a second-opinion review and verified by us on the raw rows (`iclr27_mechanism_ladder_3data_n200_s42_behavior.json`): after the message, the harness's plain-final detection accepts the model's next tool-call-free turn as the final answer, and the open-ended wording elicited a conversational hand-back (*"Understood — what would you like me to do next?"*) on $139/197 = 70.6\%$ of trigger rows under the frozen `HAND_BACK_PATTERN`, against 0–12.7% for every other arm; $155/197$ rows made no post-trigger tool call and none retained the rejected answer's text. Relative to the rejected NO-GATE answer, the v1 delay destroyed 54 correct answers and rescued 18 (net -36), where READ and READ-ONCE net $+29/+27$ and NOOP $+4$ on the same rows (Table S7 gives the structured replacement's footprint: hand-back 4.1%, net churn $+7$). Consequences, recorded without altering any frozen result: the v1 contrasts DELAY $-$ NO-GATE $= -6.00$pp and NOOP $-$ DELAY $= +6.67$pp are artifact-contaminated and are not evidence for a deliberation-harm or action-commitment mechanism; the v1 primary READ $-$ DELAY $= +10.83$pp remains protocol-valid as executed (every audit in Appendix A.3 passed) but is no longer a clean deliberation control. The headline instead uses the one-shot contrast READ-ONCE $-$ NOOP, already a primary of the separate ladder manifest; this avoids the failed v1

delay but does not establish format equivalence between read and noop. The structured DELAY was designed to close the gap and is checked against its own construct-validity gate (hand-back 4.1%) before any accuracy number from it is reported. Rescoring only the derailed rows with NO-GATE's own answer text would move the RG−Delay contrast from +10.83 to +3.50pp; this is an outcome-selected substitution sensitivity, not a corrected estimator or a bound on an identified mechanism.

I.3 READ–NOOP FORMAT SENSITIVITY

We apply the frozen `HAND_BACK_PATTERN` to the first 200 characters of each final answer, including all paired questions in the ITT population. On the 600-question panel, 32 pairs have either arm flagged and account for 9 fixes and no breaks; the 568 unflagged pairs account for 19 fixes and 5 breaks. On replication, 78 flagged pairs account for 20 fixes and 1 break, and 1,122 unflagged pairs for 40 fixes and 22 breaks. In total, 28 of the net 60 improvements occur in flagged pairs and 32 in unflagged pairs. These are observed post-treatment strata: dropping flagged pairs would change the estimand and cannot identify a format-mediated share. The regex also flags some correct answers and is not a human adjudication (Appendix E.2). The noop return string's acknowledgement and its higher flag rate motivate a format-matched comparison; avoiding the v1 delay does not make this contrast immune to output-format effects.

J SUPPORTING EVIDENCE FROM THE $n = 1{,}000$ GRID

This section compresses supporting material from the earlier no-gate/persistent-gate grid on the same question pools (§3). It keeps only the tables and results the main text references; every number is unchanged from the original, audited measurement and is cited to `docs/experiments/paper_tables.md` and the underlying result files.

J.1 THE $n = 1{,}000$ NO-GATE / READ-GATE GRID

Table S18 is the full-population ($n = 1{,}000$/dataset) gpt-5-mini-minimal grid of which the mechanism-ladder panel is a nested subset re-measuring the same contrast (§4 uses the $n = 200$/$n = 400$ panels for the causal decomposition; this is the larger, earlier-measured population that established the effect exists before the ladder decomposed it).

Table S18: gpt-5-mini-minimal, $n = 1{,}000$/dataset, no-gate vs. Read-Gate. Ret. toks/Q counts retrieved evidence tokens; Corrections/Q is inferred from rejected-action loop turns.

Dataset	Condition	Reads/Q	Loops/Q	Corrections/Q	Ret. toks/Q	LLM-Acc
HotpotQA	No gate	0.81	3.07	0.00	8,227	79.6
HotpotQA	RG	1.03	3.63	0.25	9,011	82.8
2Wiki	No gate	0.77	3.54	0.00	12,721	64.4
2Wiki	RG	1.06	4.32	0.29	14,783	69.7
MuSiQue	No gate	0.35	2.90	0.00	9,711	34.2
MuSiQue	RG	1.00	4.97	0.89	14,192	43.6

Contain-Acc (string-containment secondary metric) moves in the same direction as LLM-Acc on every cell: HotpotQA 76.8 → 80.1; 2Wiki 72.3 → 75.9; MuSiQue 32.8 → 41.3.

J.2 ENTITY-MATCHER VALIDATION

The regex entity matcher behind the read-quality entity-overlap proxy (Appendix E.1) and §2.1's snippet-only/low-evidence subtypes is validated (Table S19) against an agent-labeled, author-verified sample ($n = 60$: 30 questions and 30 gold reference answers, stratified across HotpotQA and 2WikiMultihopQA; the labels were produced by the coding agent and every label was reviewed by an author, so this is not independent human annotation) under contains-match, the same rule used operationally.

Table S19: Entity-matcher validation against an agent-labeled, author-verified sample ($n = 60$), contains-match.

	Precision	Recall	F1
Overall	1.000	0.734	0.847
Single-token entities	—	0.148	—
Multi-word entities	—	0.966	—

Precision is 1.000 on this sample (no false positives among 69 predicted entities; 95% Wilson CI $[0.947, 1.000]$). The recall gap is concentrated almost entirely in single-token entities, which the regex cannot match by construction (it requires ≥ 2 consecutive capitalized tokens), compared with 0.966 recall on the multi-word entities it targets. These observations suggest that the matcher is more likely to omit than invent an entity, but the finite validation sample does not establish a literal lower bound for every reported rate.

J.3 Cross-family judge agreement

Judge substitution on the decision-relevant rows (new arms; a targeted sensitivity check, not a full re-judging). We re-judged, with Gemini 2.5 Pro and the byte-identical judge prompt, all 678 answers involved in a discordant paired outcome of the ladder's headline contrasts (`iclr27_crossfamily_rejudge_discordant`; raw agreement 91.9%, $\kappa = 0.831$; disagreements 49 vs. 6 with the cross-family judge stricter, 44 of the 49 on intervention arms). Substituting the Gemini labels on exactly those rows and recomputing the pooled ($n = 1{,}800$) contrasts changes no sign: DELAY−NO-GATE $+3.78 \to +3.33$; NOOP−DELAY $-2.06 \to -1.94$; READ-ONCE−NOOP $+3.33 \to +3.39$; READ−READ-ONCE $+0.50 \to +0.56$; READ-ONCE−DELAY $+1.28 \to +1.44$; READ-ONCE−NO-GATE $+5.06 \to +4.78$; READ−NO-GATE $+5.56 \to +5.33$ (`results/_delay_gate/iclr27_judge_substitution_contrasts.json`). The maximum observed shift of 0.45pp applies only to this partial substitution. Originally concordant outcomes could become discordant under another judge and were not exhaustively checked; the later any-tool, restate and sham-read arms are also outside this re-judging. The result is not a bound on replacing the judge for the complete study.

§3 cites cross-family agreement $\kappa = 0.831$ on the 678 discordant answers of this paper; the inherited grid's agreement, $\kappa = 0.924$, is measured by re-judging $n = 450$ stored (question, gold, prediction) triples (75 per cell across the three datasets $\times$ no-RG/RG) with Gemini 2.5 Pro under the verbatim judge prompt used throughout the project, stratified by the gpt-5-mini-judge agreement landscape between no-RG and RG on the same question (50% disagreement, 30% both-correct, 20% both-incorrect) so κ is informed by the cases most likely to flip across judges.

Table S20: Cross-judge agreement and stratum-reweighted LLM-Acc, $n = 450$ sample. "Paper" is gpt-5-mini on the full $n = 1{,}000$ cell; "Gemini-rw" is the Horvitz–Thompson-reweighted Gemini 2.5 Pro estimate on the 75-per-cell sample.

Cell	κ	Paper	Gemini-rw	Gap
HotpotQA, no-RG	0.940	79.6	75.9	−3.7
HotpotQA, RG	0.932	82.8	79.1	−3.7
2Wiki, no-RG	0.886	64.4	63.6	−0.8
2Wiki, RG	0.871	69.7	71.0	+1.3
MuSiQue, no-RG	0.970	34.2	34.2	0.0
MuSiQue, RG	0.860	43.6	41.5	−2.1
Pooled (all six)	**0.924**	—	—	—

All six cells have $\kappa \geq 0.80$; the reweighted Gemini estimates are within ± 3.7pp of the gpt-5-mini headline on every cell, and the RG-vs-no-RG Δ is direction-preserving and similar-magnitude across judges on all three datasets (e.g. MuSiQue $+9.4$pp under gpt-5-mini vs. $+7.3$pp under reweighted Gemini). These audited cells provide no evidence that the gpt-5-mini judge systematically favors gpt-5-mini-family generations, while not ruling out smaller or task-specific judge effects. gpt-5-mini

and Gemini disagree on only $17/450$ (3.8%) triples, roughly evenly split between Gemini catching a subject-mismatch or contradiction that gpt-5-mini missed and gpt-5-mini correctly accepting a paraphrase that Gemini penalizes on format strictness.

J.4 SNIPPET-ONLY AND LOW-EVIDENCE SUBTYPES

§2.1 mentions two heuristic subtypes beyond the strict zero-read signal; we define them here (Table S21): *snippet-only* (an answer entity appears only in a retrieved snippet, never in a read chunk) and *low-evidence* (question-entity coverage in read chunks below an 80% threshold). Both use the entity matcher validated in Appendix J.2 and are reported as diagnostics, not as the headline zero-read quantity. A *discipline failure* here is an incorrect trajectory assigned to one of these subtypes, using the first-match priority no-read, then snippet-only, then low-evidence. Each table row uses the resulting discipline failures in that cell as its denominator, so the three shares sum to 100%. Unlike the main-text zero-read prevalence, this taxonomy is conditioned on an incorrect answer.

Table S21: Discipline-failure subtype shares among incorrect trajectories, nine dataset×controller cells from the inherited grid. Percentages use the mutually exclusive discipline failures in each cell as their denominator.

Dataset	Controller	n (discipline failures)	No-read	Snippet-only	Low-evidence
HotpotQA	4o-mini	76	3.9	9.2	86.8
HotpotQA	gpt-5-mini minimal	133	69.9	1.5	28.6
HotpotQA	5mini-med	45	37.8	6.7	55.6
2Wiki	4o-mini	128	7.8	11.7	80.5
2Wiki	gpt-5-mini minimal	221	69.7	3.6	26.7
2Wiki	5mini-med	57	38.6	26.3	35.1
MuSiQue	4o-mini	278	15.5	17.6	66.9
MuSiQue	gpt-5-mini minimal	570	90.4	2.1	7.5
MuSiQue	5mini-med	217	56.7	22.6	20.7
Pooled (1,725 discipline failures)			**56.8**	**9.3**	**33.9**

Pooled over all nine cells, 56.8% of discipline failures are strict no-read finals (the primary, non-heuristic signal used throughout this paper), 9.3% take the prompt-permitted snippet-only path, and 33.9% are low-evidence; 43.2% of discipline failures ($745/1{,}725$) occur despite at least one read (snippet-only plus low-evidence combined) — the Read-Gate removes the zero-read floor, it does not guarantee adequate post-read inspection, which is exactly the gap §L.4 measures directly on the ICLR ladder's own trigger rows.

K THE ANSWER-OBLIGATION STUDIES (P1, P2, P3)

Arms and the answer obligation. All seven arms branch at the first zero-read final of the frozen no-gate trajectory; non-trigger questions retain the B record in every arm. N0 and R0 are the ladder's NOOP and READ-ONCE arms re-run under the format-study manifest (their gate messages are byte-identical to Appendix A.1). NF, RF, DF and SF append, once, two line feeds and the two-sentence instruction

> When finalizing after any required action, answer the original question. Do not return an acknowledgement, a procedural plan, or a question to the user as the final answer.

to the first-trigger user gate message of the noop demand, the read demand, the structured delay and the decoy read respectively. No prefix, system prompt or tool-result string is altered; the `noop` result remains "noop acknowledged. No operation was performed." The instruction's hash is pinned in both manifests (`iclr27_format_factorial_2panel_s43`, `iclr27_format_medical_external_n800_s43`).

Questions, prefixes and cache protocol. P1 uses the ordered question IDs of the ladder panel (600) and the replication panel (1,200) and replays each question's pinned prefix from the read-only response cache; a cache miss aborts before any provider call, and the first final is validated and hashed before branching. P2 draws 800 questions from the GraphRAG-Bench medical question pool

(Xiang et al., 2025) by the same seeded procedure, with its own hybrid index over the medical chunk corpus; all seven arms are new trajectories there (the B arm ran first and alone, so that the forecast of §6 could be frozen before any intervention arm). P3 hash-selects 20 baseline-trigger questions per benchmark from the P1 panels (15 discovery, 45 replication) and redraws each of the six intervention arms twice more with the cache bypassed after the branch, giving three rollouts per (question, arm); the B rollout is the P1 one and is not redrawn. Trigger definition, one-shot semantics and post-trigger tool exposure are those of Appendix A.2; the trigger is the first ordinary-loop final attempt (plain content or `final` tool call) with zero successful reads, and harness-forced finals at the loop or token limit are not triggers.

Pre-registration and analysis. Each manifest froze, before collection: the two primaries (RF−NF by exact paired McNemar over all questions; (R0−N0)−(RF−NF) by a stratum-weighted question-level Wald test), Holm over the two within each study, the seven secondaries with pointwise 10,000-draw question-bootstrap intervals (stratified by benchmark and panel, seed 43), equal benchmark weights, and a no-exclusion rule (no row dropped on hand-back, compliance, read or outcome). The Medical forecast target (R0−B ITT), its estimator and its comparators were frozen in the same manifest.

Table S22: Answer-obligation studies: every contrast of Table 2 with its nominal 95% question-bootstrap interval (10,000 draws, stratified by benchmark and panel), fixes/breaks and, for the two primaries, the Holm-corrected p within each study. Secondary and post-peek intervals are pointwise.

Contrast	role	effect (pp)	95% CI	fixes/breaks	Holm p
P1 Wikipedia (n = 1,800)					
RF−NF	primary	+1.22	[+0.17, +2.28]	(58/36)	0.030
(R0−N0)−(RF−NF)	primary	+2.06	[+0.56, +3.56]	—	0.015
RF−DF	secondary	+0.94	[−0.06, +2.00]	(53/36)	—
RF−SF	secondary	+0.67	[−0.33, +1.67]	(50/38)	—
R0−N0	secondary	+3.28	[+2.11, +4.50]	(90/31)	—
R0−B	secondary	+5.28	[+4.11, +6.50]	(109/14)	—
RF−B	secondary	+7.00	[+5.78, +8.22]	(133/7)	—
NF−N0	secondary	+3.78	[+2.67, +4.89]	(87/19)	—
RF−R0	secondary	+1.72	[+0.67, +2.83]	(66/35)	—
NF−B	post-peek	+5.78	[+4.67, +6.94]	(110/6)	—
SF−NF	post-peek	+0.56	[−0.50, +1.61]	(51/41)	—
P2 Medical (n = 800)					
RF−NF	primary	−0.38	[−2.25, +1.50]	(30/33)	0.801
(R0−N0)−(RF−NF)	primary	+2.12	[−1.12, +5.38]	—	0.397
RF−DF	secondary	−1.50	[−3.25, +0.12]	(18/30)	—
RF−SF	secondary	+0.12	[−1.62, +1.88]	(25/24)	—
R0−N0	secondary	+1.75	[−1.00, +4.50]	(71/57)	—
R0−B	secondary	−3.00	[−5.38, −0.62]	(36/60)	—
RF−B	secondary	+3.12	[+1.25, +5.00]	(42/17)	—
NF−N0	secondary	+8.25	[+5.75, +10.75]	(88/22)	—
RF−R0	secondary	+6.12	[+3.75, +8.50]	(73/24)	—
NF−B	post-peek	+3.50	[+1.38, +5.62]	(50/22)	—
SF−NF	post-peek	−0.50	[−2.25, +1.25]	(24/28)	—

Judgment-completion amendment. The parity judge (`gpt-5-mini`, low effort, 300 output tokens) returned a malformed verdict on 28–36 of 1,800 answers per P1 arm and 77–96 of 800 per P2 arm, which the original parser scored as wrong. The amendment was frozen after a schema-only check (2026-09-07, 02:09 UTC) and the completed-judgment analysis protocols registered at 02:40–02:47, before the authorized interim look (03:37), which already used completed verdicts; every malformed verdict was re-judged with a 1,500-token cap under the same prompt; 15–19 labels per P1 arm and 28–42 per P2 arm changed from wrong to right and none in the other direction. Table 2 reports the completed judgments. Under the original parsing the P1 primaries are +1.22pp and +1.89pp (Holm $p = 0.030$ for both) and every P1 secondary moves by at most 0.2pp. The Medical contrasts move more, because 77–96 malformed verdicts per arm were re-scored there: RF−NF goes from +0.62 to −0.38pp (a sign change; both non-significant), the interaction from +0.62 to +2.12pp,

RF−DF from $+0.25$ to -1.50pp, RF−SF from $+1.38$ to $+0.12$pp, R0−B from -2.38 to -3.00pp and RF−B from $+4.62$ to $+3.12$pp. No Medical primary is significant under either parsing; the frozen JSON keeps both views (Table S23).

Table S23: **Medical contrasts under both judgment views.** Pooled ITT estimates (pp) with 95% intervals from the frozen analysis JSON: the original 300-token judge parsing, which scored malformed verdicts as wrong, and the completed judgments reported in the main text. No primary passes Holm under either view.

Contrast	Original parsing	Completed judgments
RF−NF (primary)	$+0.625$ $[-1.25, +2.50]$	-0.375 $[-2.25, +1.50]$
interaction (primary)	$+0.625$ $[-2.50, +3.63]$	$+2.125$ $[-1.12, +5.38]$
R0−B	-2.375 $[-4.75, +0.00]$	-3.000 $[-5.38, -0.62]$
RF−B	$+4.625$ $[+2.62, +6.62]$	$+3.125$ $[+1.25, +5.00]$

Interim look and post-peek additions. One interim descriptive view of P1 was authorized at 688 of 1,800 questions (264 triggers), documented with a hash-stamped protocol. After it, two descriptive contrasts were added (NF−B and SF−NF; Table 2, last block) and the hand-back cross-tables below were specified. No primary, secondary or test was changed.

Table S24: **Where the NF−N0 gain comes from.** Paired outcome of the noop demand without (N0) and with (NF) the answer obligation, split by whether the N0 final answer is flagged as a hand-back by the frozen first-200-character detector. All questions of each study; post-peek descriptive.

	P1 Wikipedia			P2 Medical		
N0→NF outcome	N0 hand-back	not	total	N0 hand-back	not	total
fix (wrong→right)	31	56	87	65	23	88
break (right→wrong)	1	18	19	3	19	22
both right	8	1068	1076	21	482	503
both wrong	51	567	618	23	164	187
total	91	1709	1800	112	688	800

Where the noop demand's gain comes from. Table S24 splits the paired N0→NF outcomes by whether the N0 final answer is flagged by the frozen first-200-character hand-back detector. On P1, 31 of the 87 fixes and 51 of the 618 both-wrong pairs are flagged hand-backs; on Medical, 65 of the 88 fixes. The obligation therefore converts hand-backs into answers that the judge accepts, and the AI-rater audit below checks that those answers are correct rather than merely well-formed.

Single-rollout noise (P3). On the 60 redrawn questions the trigger-sample RF−NF contrast is -5.0, $+5.0$ and $+5.0$pp on the three draws (pointwise 95% intervals about ± 10pp each) and $+1.67$pp $[-6.11, +10.00]$ for the within-question three-draw mean. This is a trigger-sample estimand with equal benchmark allocation, not the P1 ITT estimand; it documents draw-to-draw variation on 60 trigger questions and provides no noise bound for the 1,800-question ITT. All other pre-specified P3 contrasts are in the frozen analysis JSON.

Cost. The three studies plus the judgment amendment spent $64.26 of a $100 programme ceiling at full list price (settled-token estimate $40.72), all under per-stage ceilings and a pre-call reservation guard; P1 agent and judge spend per arm is recorded in the frozen analysis ($3.2–$8.1 per intervention arm).

K.1 Second controller: Qwen2.5-7B factorial

The pre-registered manifest (frozen before any continuation) reuses the frozen NO-GATE trajectories and LLM cache of the second-family core (Appendix H.2) as the pinned prefix and branches N0, R0, NF and RF at the first zero-read final; DF and SF were not run. Of the 600 core questions, 37 whose B run ended in a provider error and one whose prefix could not be replayed from the

cache were excluded before any outcome was read. The agent ran on a local vLLM server (bf16, 24,576-token limit, the limit the core run's error messages report) at no cost; the judge is the parity `gpt-5-mini` with the 1,500-token completed-judgment protocol ($0.13). The manifest pre-registered the reading rule: if NF−N0 and the interaction are both near zero, the obligation effect is read as not established beyond gpt-5-mini. Three provenance notes. The core's noop arm had already shown one flagged hand-back among its 136 trigger answers; this was observed, not frozen as a prediction. The core's 114 triggers count questions valid in all three core arms, the factorial's 136 every question with a valid B trajectory (Figure 3 reports the inherited grid). B answers and labels are the core's (300-token judge) while the four new arms use the 1,500-token protocol; this cancels in RF−NF and the interaction but not in R0−B and RF−B. Table S25 reports the pre-specified complete-case analysis (29 questions with a context-limit error in at least one arm are dropped: 25 under R0, 20 under RF, one under NF; the limit belongs to this serving configuration, not to the model) and the error-as-incorrect sensitivity; the frozen JSON also carries per-benchmark contrasts.

Table S25: **Second controller: the answer-obligation factorial on Qwen2.5-7B-Instruct.** Paired differences (pp) with nominal 95% question-bootstrap intervals on the 562 valid questions of the second-family core (136 triggers). Complete-case rows (not ITT) drop the 29 questions with a context-limit error in any arm (errors per arm: N0 0, R0 25, NF 1, RF 20); the sensitivity view scores an error as incorrect. Primaries are Holm-corrected over two tests within each analysis; secondaries are pointwise. Holm p-values in the role column refer to the complete-case analysis; the sensitivity values are 1.000 (RF−NF) and 0.980 (interaction).

(a) Paired contrasts

		Complete case ($n = 533$, 107 triggers)		Error as incorrect ($n = 562$)	
Contrast	role	effect [95% CI]	f/b	effect [95% CI]	f/b
RF−NF	P, Holm 0.53	$+0.94$ $[-0.38, +2.25]$	9/4	$+0.18$ $[-1.42, +1.60]$	10/9
(R0 − N0) − (RF − NF)	P, Holm 0.53	$+0.56$ $[-0.75, +1.88]$	—	$+0.53$ $[-1.07, +2.14]$	—
R0−N0	S	$+1.50$ $[+0.19, +3.00]$	11/3	$+0.71$ $[-0.71, +2.14]$	11/7
R0−B	S	$+0.75$ $[-0.19, +1.88]$	6/2	$+0.00$ $[-1.25, +1.25]$	6/6
RF−B	S	$+0.00$ $[-1.13, +0.94]$	4/4	-0.71 $[-1.96, +0.53]$	5/9
NF−N0	S	-0.19 $[-0.94, +0.56]$	2/3	-0.18 $[-0.89, +0.53]$	2/3
RF−R0	S	-0.75 $[-2.06, +0.56]$	5/9	-0.71 $[-2.14, +0.71]$	7/11

(b) Arm statistics, complete case

Arm	accuracy (all)	accuracy (trigger rows)	flagged hand-backs (trigger answers)	successful-read rate
B	45.6%	49.5%	0	79.9%
N0	44.8%	45.8%	1	79.9%
R0	46.3%	53.3%	0	99.4%
NF	44.7%	44.9%	1	79.9%
RF	45.6%	49.5%	1	99.6%

K.2 AI-RATER AUDIT OF JUDGE VALIDITY

The gain from the answer obligation could be a judge artifact: the automated judge marks hand-backs wrong and might mark any well-formed answer right. We therefore audited 842 blinded items — 663 from a stratified sample of the ladder and format arms and 179 further items covering every N0-hand-back→NF-fix pair in both studies (13 duplicates merged) — each showing the question, the accepted final answer and the reference answer with its evidence, but no arm, study or judge label. Two LLM raters from different vendors (Anthropic Claude, OpenAI Codex) labelled correctness (CORRECT/INCORRECT/UNCERTAIN), hand-back (HAND_BACK/ANSWER) and evidence sufficiency under a written protocol. Both covered all 842 items with no invalid rows; they agreed unanimously on correctness for 787 items and on hand-back for 828. The 65 disagreements were adjudicated by the author under four written rules, drafted by one rater after both raters' labels were in and approved by the author before being applied (multi-item references count as correct at half coverage with no contradiction; comparison questions answered by listing facts count if the correct entity is stated first; hand-back means deferral to the user, so an explicit abstention is an incorrect answer, not a hand-back; a stated headline governs when it conflicts with the response's own reasoning); 18 answers (19 item-source rows) whose question admits several valid answers were marked uncertain and excluded. These labels are author adjudication by rule, drafted by one of the raters, and not an independent human annotation.

Against the frozen judge, the consensus labels agree on 93.9% of the 912 item-source rows with a resolved source link (509 both right, 347 both wrong, 48 judge-wrong/raters-right, 8 judge-right/raters-wrong); applying the strict reading of the multi-item rule to the 16 items it decided gives 94.3%. Agreement is 93.5–100% on the Wikipedia strata and 87.6% on Medical, where the disagreement is asymmetric — the judge is stricter than the raters — and concentrated on the obligated arms (NF 13 and RF 9 of the 31 judge-wrong/raters-right medical rows, against N0 7 and R0 2). In the key stratum, NF answers on questions whose N0 answer was a hand-back are rated correct on 30 of 31 P1 items (one uncertain) and 64 of 65 Medical items. On the audited sample the obligation's effect is therefore made of genuine answers, and the medical disagreements run in the direction of the judge under-crediting the F arms; the base sample is stratified over discordant and concordant pairs without population reweighting and the added stratum was defined after observing outcomes, so it bounds no full-population bias. The 2WikiMultihop reference evidence in the pack is stored upstream as relation labels split into characters, so evidence-sufficiency labels on that benchmark are uninformative; correctness and hand-back labels are unaffected.

K.3 Post-review descriptive analyses

Three descriptive analyses were added after a 2026-09-14 internal review of the submission draft; none is a pre-registered contrast, and none changes a label, a frozen analysis or a primary family.

A demanded read need not add text. The `search` tool returns, per hit, the chunk identifier and a sentence snippet. Hybrid search takes the leading five sentence segments for chunks in its BM25 candidate set and, for dense-only candidates, the similarity-selected sentences in document order. Sentence boundaries are detected at punctuation followed by an upper-case letter, digit or bracket; the Wikipedia chunk corpus of the LinearRAG release is lowercased, so the splitter often finds no boundary and a "segment" is the whole chunk, whereas the medical corpus is mixed-case and multi-sentence. Table S26(a) counts, among trigger rows of the two one-shot read arms that had a successful post-trigger read, those whose first read body was already present verbatim in a search result issued before the trigger: 39.1% of R0 and 38.1% of RF rows on Wikipedia (238 and 236 rows; 34.7% and 34.5% of all 685 triggers), all from single-segment chunks, and no such full-body match in the medical arms. The Figure 1 question is such a row: the 5,099-character body of chunk 446 was in the pre-trigger search result of every arm. Some demanded reads therefore repeat text already exposed by search, and the ITT contrast is not a pure contrast in new text exposure; a non-match does not imply that the remaining reads were wholly new, and the table does not identify how much of any gain arises on either path. Panel (b) counts triggers that fired before any search call. The decoy of SHAM-READ/SF is chosen by hashing the requested chunk identifier to a start position in the corpus and taking the first chunk from there that is neither the requested chunk nor one already shown in a search result; it is not an independent uniform draw. The compared strings are exported (`results/_format/post_review_read_exposure_export.json.xz`; search hits split into blocks, a block shorter than the read body stored as hash and length only) so that the containment test re-runs without the full transcripts.

Table S26: **A demanded read need not add text.** (a) For each one-shot read arm, among trigger rows with a successful post-trigger read, those whose first read body was already present verbatim (contiguous, unnormalised) in a search result issued before the trigger; percentages use the rows with a read as denominator. (b) Trigger rows whose premature final preceded any search call; percentages use all triggers. Descriptive, computed after the 2026-09-14 review.

(a) First demanded read already exposed by a pre-trigger search result

Corpus	Arm	Triggers	Rows with a read	Body already in search
Wikipedia	R0	685	608	238 (39.1%)
Wikipedia	RF	685	619	236 (38.1%)
Medical	R0	458	362	0 (0.0%)
Medical	RF	458	381	0 (0.0%)

(b) Triggers that fired before any search call

Corpus	Triggers	No search before the trigger
Wikipedia	685	55 (8.0%)
Medical	458	40 (8.7%)

What the obligation changed under the read demand. Table S27 gives recovery and disruption for R0 and RF on both corpora and splits the paired R0→RF change by the pre-intervention B outcome and by the post-treatment hand-back flag. Of the 49 net medical improvements of RF over R0, 43 occur on questions B answered correctly; separately, the 18 rows flagged as hand-backs under R0 contribute six, and 9 of those 18 are B-correct rows carrying all six, so the two partitions overlap. They are descriptive, the detector has misses, and no mechanism or causal share is identified. Under F the disruption rates of the two corpora are closer (3.1% vs. 5.6%) than without it (6.2% vs. 19.9%); this is a post-hoc comparison of a different policy and does not revise the frozen R0 forecast of §6.

Table S27: **Recovery, disruption and the RF−R0 change by stratum.** Top: per corpus and arm, recovery r = share of B-wrong trigger rows fixed, disruption d = share of B-correct trigger rows broken, and the ITT gain over B (pp of all questions). Bottom: paired R0→RF fixes and breaks on strata defined before the intervention (B outcome; the two rows partition the trigger rows) and, separately, on the post-treatment stratum of rows whose R0 answer the frozen hand-back detector flags; the flagged rows overlap the B strata (on the medical corpus 9 of the 18 are B-correct and carry all 6 net improvements). Descriptive, computed after the 2026-09-14 review; the R0 rows reproduce Table 4.

Corpus	Arm	r (fixes / B wrong)	d (breaks / B correct)	Arm−B (pp)
Wikipedia	R0	109/459 = 23.7%	14/226 = 6.2%	+5.28
Wikipedia	RF	133/459 = 29.0%	7/226 = 3.1%	+7.00
Medical	R0	36/157 = 22.9%	60/301 = 19.9%	−3.00
Medical	RF	42/157 = 26.8%	17/301 = 5.6%	+3.12
Corpus	**RF vs. R0 on**	**n**	**fixes / breaks**	**net (pp of all questions)**
Wikipedia	all trigger rows	685	66 / 35	+31 (+1.72)
Wikipedia	B correct	226	10 / 3	+7 (+0.39)
Wikipedia	B wrong	459	56 / 32	+24 (+1.33)
Wikipedia	R0 answer flagged as hand-back	42	9 / 0	+9 (+0.50)
Medical	all trigger rows	458	73 / 24	+49 (+6.12)
Medical	B correct	301	49 / 6	+43 (+5.38)
Medical	B wrong	157	24 / 18	+6 (+0.75)
Medical	R0 answer flagged as hand-back	18	7 / 1	+6 (+0.75)

L ADDITIONAL LADDER FINDINGS

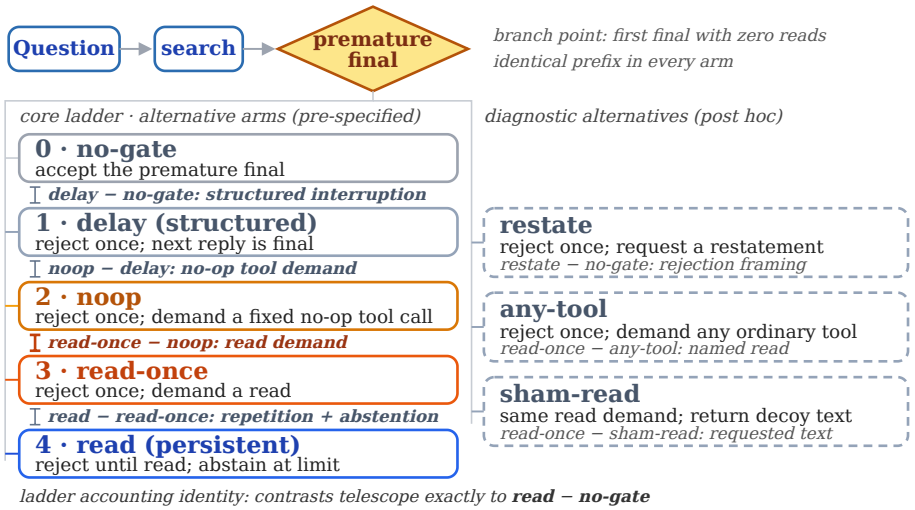

Figure S3: **The original intervention ladder.** All arms share the complete pre-trigger trajectory and are alternative continuations at the first zero-read final. Solid boxes form the pre-specified ladder; dashed boxes are post-hoc diagnostic controls. Contrast names are bundle labels, not identified mechanisms.

L.1 THE LADDER ACCOUNTING IDENTITY

With $A(\cdot)$ denoting paired accuracy,

$$A(\text{READ}) - A(\text{NO-GATE}) = \underbrace{A(\text{READ}) - A(\text{READ-ONCE})}_{\text{repeated enforcement + abstention}} + \underbrace{A(\text{READ-ONCE}) - A(\text{NOOP})}_{\text{read demand vs. information-free action}} + \underbrace{A(\text{NOOP}) - A(\text{DELAY})}_{\text{noop demand vs. delay}} + \underbrace{A(\text{DELAY}) - A(\text{NO-GATE})}_{\text{structured interruption}}.$$

The identity is exact for any paired outcomes. Each term differences two bundles sharing the branch point and the complete pre-trigger trace; rungs 1–3 also share one-shot enforcement, whereas the rung-4 term deliberately mixes repeated enforcement with abstention. The interruption term combines one more decision step with the implicit negative signal that rejection carries about the first answer, and the restate control separates the two only partly (Appendix D). Pooled over 1,800 questions the persistent gate's $+5.56$pp splits into $+3.78$ (delay), -2.06 (noop minus delay), $+3.33$ (read-once minus noop) and $+0.50$ (repetition); measured in retrieved tokens the interruption costs about 3% and the read 19–22%. Figure S4 draws the pooled terms.

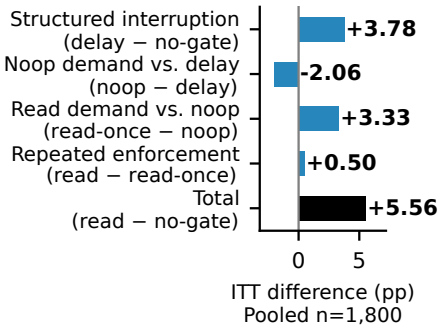

Figure S4: **Ladder identity, pooled.** Terms over the 1,800 paired questions of Table 1; they sum to the total up to rounding. Intervals in Appendix B.

L.2 An unspecified tool demand does not substitute for a named read

The post-hoc ANY-TOOL control rejects the final once and demands any ordinary tool. The agent complies on 99% of trigger rows, but its first post-message action is another search on 672 of 685 rows; it eventually reads on only 19% (panel) and 24% (replication) of them, against about 91% under READ-ONCE (Table S9). Its pooled accuracy lies $+1.17$pp above NOOP; the named read beats it on both panels ($+3.00$ and $+1.75$pp; $+2.17$ pooled, 88 fixes vs. 49 breaks).

L.3 Requested versus decoy content

SHAM-READ returns, under the requested chunk id, an unseen decoy chosen by a hash from the benchmark corpus (excluding the requested chunk and every chunk already shown in a search result); its tool-free and read rates after the message match READ-ONCE by construction, and trajectories diverge only after the returned text. The decoy read beats the information-free action by $+2.22$pp pooled (76 fixes vs. 36 breaks) and is nearly tied with the structured delay ($+0.17$pp by subtraction of tabulated contrasts). The requested chunk's increment over the decoy is $+2.50$pp on the panel (21 vs. 6) but $+0.42$pp on replication (27 vs. 22), $+1.11$pp pooled (Holm-adjusted $p = 0.057$ in the pooled side-control family). This establishes neither equivalence nor a causal share for evidence content, and the decoy is not guaranteed irrelevant.

L.4 A forced read is not always a useful read

Two proxies for the first induced read — gold-answer containment and question-entity overlap, checked against a blinded two-model audit with author adjudication (proxy-versus-adjudicated $\kappa = 0.58/0.54$; inter-rater $\kappa = 0.80$; Appendix E.2) — separate *a read occurred* from *useful evidence was inspected*. On the 600-question panel the persistent gate induced a first read on 189 of 197 trigger rows (the one-shot demand on 179); 50.3% of those first reads contained the gold answer (65.0% for voluntary reads on non-trigger rows) and 20.1% had low entity overlap without gold containment, against 8.7% (Appendix Tables S10–S12). Accuracy after a gold-containing induced read was 73.7–74.7%, against 25.5–26.1% after one that missed. Trigger rows and rows that read unprompted are differently selected populations, so the gap is descriptive, not a causal effect of relevance, and a chunk lacking the final answer can still supply intermediate-hop evidence.