

What Agent Benchmarks Don't Tell You: Recoverability and Credit Assignment in Agentic Retrieval Evaluation

Sai Krishna Rallabandi, Parag Pravin Dakle, Preethi Raghavan

Fidelity Investments

{saikrishna.rallabandi, paragpravin.dakle, preethi.raghavan}@fmr.com

Abstract

Agent quality on long-horizon tasks is usually reported as a single number: the aggregate gain of a tool-using agent over a non-agentic baseline on a benchmark. We argue that this number is not a property of the agent. It is jointly determined by a property of the benchmark and by which sub-component of the agent happens to be load-bearing, and reporting it alone makes agent quality look benchmark-dependent and irreproducible. We propose a four-stage evaluation protocol that separates the two. Stage 1 characterizes the benchmark *before* any outcome is observed, using an LLM-free diagnostic we call **recoverability**: whether evidence missed on the first retrieval is still reachable from what the system has already seen, and if so whether it is reachable by *reformulation* or only by *structural navigation*. Stage 2 assigns credit with an ablation ladder — one-shot RAG, a choice-free fixed-iterative loop, and a full agent — that localizes any measured gain to iteration or to open-ended agency. Stages 3 and 4 add the validity checks and cost accounting that change verdicts in practice. Applying the protocol to about 6,000 logged agent episodes over eight corpora and three reader families, we find that the headline agent gain ranges from -0.21 to $+0.30$ (each corpus in its native metric) purely as a function of the recoverability regime, and that on reformulation-recoverable corpora the choice-free loop recovers 65–96% of the gain normally credited to agency under the two stronger readers (and most of it under the open reader). Agency earns its cost only on the longest-horizon, navigation-recoverable corpus, where iteration saturates by round 2 and leaves a residual ($+0.15$ under the open reader at full N) that no round budget closes under any reader.

1 Introduction

Evaluations of tool-using agents on complex, long-horizon tasks are typically summarized by one quantity: the aggregate improvement of the agent

over a non-agentic baseline on some benchmark. That quantity is unstable in a way that is rarely reported. In this study, on the *same* agent harness, the *same* three reader families, and a retrieval stack held fixed within each corpus, the measured agent-minus-baseline delta ranges from -0.21 to $+0.30$ across corpora. Each corpus is scored in its own native metric, so these two endpoints are not on one scale; the finding is the change of sign, not the width of the span. A reader who sees only the favorable row concludes that agency works; a reader who sees only the unfavorable row concludes that it does not. Neither conclusion is about the agent.

We treat this as an evaluation-methodology problem rather than a modeling problem, and identify two distinct confounds that a single headline number silently absorbs.

Confound 1: benchmark selection determines the verdict. Whether extra retrieval can help at all is a property of the task, fixed before any system is run. If the first retrieval misses gold evidence but that evidence is still *addressable* from what the system has already seen, more retrieval has a target; if it is not addressable, no amount of agency helps. The proxies currently used for this — “multi-hop,” “hard,” or the raw retrieval coverage gap — do not separate the two cases. We exhibit subsets with comparable coverage gaps (0.65 – 0.68 versus ~ 1.00) and opposite agent value.

Confound 2: the gain is credited to the wrong component. An agent changes at least two things at once relative to a one-shot pipeline: it retrieves more than once, and it chooses what to do next. A positive delta is conventionally attributed to *agency*, but a fixed, choice-free loop that only re-queries is a strictly weaker system that may capture the same gain. Throughout this paper such a loop is a measuring instrument for credit assignment, not a system we propose. Without that intermediate arm, an evaluation cannot tell iteration from agency, and it systematically over-credits the expensive compo-

nent.

Contribution. We propose and instantiate a four-stage protocol for evaluating agents on long-horizon retrieval tasks (§3):

1. **A-priori benchmark characterization.** Compute an LLM-free *recoverability* profile for each benchmark and pre-defined subset before observing any system outcome, and stratify reporting by regime rather than by difficulty label (§5).
2. **Credit assignment by ablation ladder.** Insert a deterministic fixed-iterative arm between the one-shot baseline and the full agent, and report the fraction of the agent’s advantage that it already captures (§6).
3. **Validity checks that can flip a verdict.** Pre-registered equivalence testing for “no difference” claims, multi-reader replication, metric-artifact audits, and cross-family judge regrading (§7).
4. **Cost-aware reporting.** Quality against retrieval calls, tokens, and latency on a Pareto plane, since the agent’s advantage and its expense grow together (§8).

Applying the protocol to ~6,000 logged agent episodes over eight corpora and three reader families (Claude Sonnet 4.6, GPT-5.4, and an open Qwen3.6-35B-A3B) yields three measurements that a conventional single-number report would have hidden. First, the a-priori recoverability profile predicts the sign and rough magnitude of the measured agent advantage, whereas the raw coverage gap mispredicts it out of sample. Second, on reformulation-recoverable corpora the choice-free loop captures 65–96% of the agent’s advantage under the two stronger readers, so most of what is normally reported as evidence for agency is evidence for iteration. Third, the protocol does *not* collapse into a blanket negative result about agents: on the longest-horizon corpus in the study, where missing evidence must be navigated through document structure rather than rediscovered by a better query, iteration saturates by round 2 and full agency retains a residual (+0.15 under the open reader at full N) under every reader. An evaluation protocol that can only produce negative results about agency would not be a useful protocol; this one localizes where agency actually pays.

2 What a single headline number hides

Consider the standard reporting unit: “system A beats system B by Δ on benchmark X .” For agentic retrieval this hides a decomposition. Write M for the set of gold passages missed by one-shot retrieval on a question; the agent can convert a miss into an answer only if some available action makes an element of M addressable. So Δ factors, informally, into (i) how often the first retrieval misses, (ii) what fraction of those misses any subsequent action can reach, and (iii) which action class reaches them. Quantity (i) is the familiar coverage gap and is routinely reported. Quantities (ii) and (iii) are almost never reported, and they are what make Δ change sign across benchmarks.

Terms used throughout.

- **Coverage gap** $= 1 - \text{gold-coverage}@k$: how much gold evidence one-shot retrieval misses.
- **Reformulation recoverability** $R_{\text{ref}} = |M_{\text{ref}}|/|M|$, where $M_{\text{ref}} \subseteq M$ is the missed gold that a follow-up *query* recovers.
- **Navigation recoverability** R_{nav} : fraction recovered by *structural* actions (fetch-by-id, list neighboring passages).
- **Absent-from-query (AFQ)**: missed evidence addressable by neither channel.
- **Horizon**: number of *dependent* retrieval steps a question requires — each step’s query formulable only after the previous step’s result.

These quantities compose: the gap counts misses, and $R_{\text{ref}}/R_{\text{nav}}$ partition them into what is reachable and how. Two benchmarks can miss the same amount of gold while differing completely in whether the missing passages are nameable from the question, exposed by an intermediate result, or reachable only by walking document structure. The gap is therefore a *precondition* for retrieval improvement, not an explanation of it, and cannot choose between one-shot RAG, fixed iteration, and a full agent.

Horizon length is the third axis, and it is where long-horizon evaluation specifically goes wrong. A densely cross-referenced regulatory question can look long-horizon while being AFQ at step 1, so its effective horizon for a retrieval agent is one. Both horizon and recovery channel must be reported: agency pays where the horizon is long *and* the recovery channel is navigation, not merely where the horizon is long.

3 A four-stage evaluation protocol

The protocol is a reporting discipline, not a new system. Each stage is a measurement that can be made from artifacts an evaluation already produces.

Stage 1 — characterize before you measure.

For each benchmark and each pre-defined subset, compute the coverage gap at the operating retrieval depth together with an estimate of the recovery channel, both fixed before any system outcome is inspected. The gap is a pure retrieval measurement requiring no model call. For the channel we use two LLM-free probes: (a) a one-step pseudo-relevance-feedback expansion (Rocchio, 1971) — expand the query with top-IDF terms from its own first-hop top-3 passages, re-retrieve, and ask whether gold is now covered — as a conservative lower bound on R_{ref} ; and (b) structural reachability, i.e. whether gold sits in an already-retrieved document but in a passage the first retrieval did not return, as an upper bound on R_{nav} 's contribution. Subsets are defined by question structure (hop type, hop depth, reference-group size), never by outcome. Stage 1's output is a position on the *recoverability plane* (Figure 1) — a prediction, not a result.

Stage 2 — assign credit with an ablation ladder.

Report three arms, not two, on every corpus where the middle arm is meaningful: **(R1)** one-shot top- k RAG; **(R2)** a *fixed-iterative* loop that reformulates and re-retrieves for a fixed number of rounds with no tool choice, no adaptive stopping, and no navigation policy — IRCoT/Self-Ask-style (Trivedi et al., 2023; Press et al., 2023), but used here as a deterministic ablation instrument rather than proposed as a system; **(R3)** the full tool-using agent. The reported statistic is the **gap-closing fraction** $\gamma = (R2 - R1)/(R3 - R1)$: the share of the agent's advantage that a choice-free system already delivers. A gain with $\gamma \approx 1$ is evidence for iteration; only the residual $1 - \gamma$ is evidence for agency. Sweeping R2's round budget additionally reports *where* the gain saturates, which is the operational horizon of the benchmark.

Stage 3 — make the null claims testable. Long-horizon evaluation produces many “no difference” cells, and a non-significant difference is not equivalence. We require: paired bootstrap CIs (Efron and Tibshirani, 1993) and permutation tests for difference claims with Holm–Bonferroni correction (Holm, 1979); two one-sided tests (Lakens, 2017) against margins fixed *before* any test-set delta is

viewed for equivalence claims; replication across ≥ 2 reader families at full N ; a metric-artifact audit (does the gain survive exact match, and are the winning answers simply longer?); and, where an LLM judge is used, a re-grade by an independent judge family over identical stored prompts. We report below one case where the artifact audit flipped a sign and one where the judge check changed a magnitude by $3.7\times$.

Stage 4 — price the horizon. Report quality against cost in retrieval/search calls on every corpus, and in input tokens and wall-clock latency wherever instrumented, plotting the arms as a Pareto frontier. This is not cosmetic: quality and cost both rise with recoverability, so a flat frontier is itself a finding.

The rest of the paper follows this order. §4 describes the experimental setup; §5–§8 then apply one stage each, in sequence; §9 condenses the result into a reporting checklist.

4 Experimental setup

Arms. The architectures are intentionally minimal, so that the within-corpus comparison is the only moving part. **R1/RAG-strong** retrieves top- k once and calls the reader; one-shot gold coverage saturates with depth, so we set $k=8$ as the operating point and retain $k=4$ as a breadth control. **R2/fixed-iterative** repeats a reformulate-and-retrieve prompt for a fixed number of rounds, with no tool choices and no adaptive stopping. **R3/agent-strong** may call `search`, `get_passage`, and `list_doc_passages`, choosing its own queries, reads, and stopping point. Within each corpus all arms share the reader, retriever setting, and native metric.

Readers. Three families covering the open/hosted split: Qwen3.6-35B-A3B, Claude Sonnet 4.6, and GPT-5.4. The open reader runs at full N on every arm; the two stronger families replicate the headline comparison at full N and the ablation ladder on a paired subset.

Corpora. Eight corpora spanning the recoverability plane: ObliQA (Gokhan et al., 2024) (regulatory QA over mutually referenced rules), HotpotQA (Yang et al., 2018) and MuSiQue (Trivedi et al., 2022) (Wikipedia multi-hop), ConditionalQA (Sun et al., 2022) (conditional QA over long UK government documents), plus SQuAD (Rajpurkar et al., 2016), ConvFinQA (Chen et al., 2022), TAT-QA (Zhu et al., 2021) and FinQA (Chen et al.,

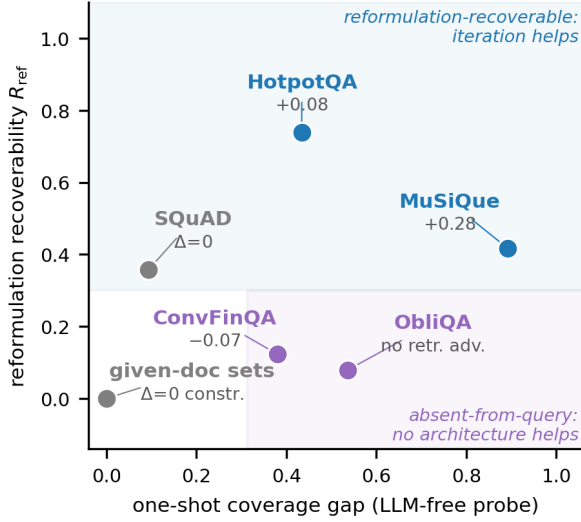


Figure 1: The recoverability plane produced by Stage 1. Horizontal: one-shot **coverage gap** ($1 - \text{gold-coverage}@k$); vertical: **reformulation recoverability** R_{ref} . Both coordinates are computed before any system outcome is observed; the grey annotations give the measured agent advantage for reference only. Position on this plane, not the horizontal coordinate alone, is what predicts the measured agent advantage.

2021) as controls; App. A lists the arms available for each, App. E discusses the two controls that look agent-favorable and are not, App. G gives benchmark context, and App. H shows qualitative agent-vs-RAG traces. Each corpus is scored in its native metric, so magnitudes are not comparable across rows — which is precisely why the protocol reports within-corpus contrasts and gap-closing *fractions* rather than pooled deltas. This one setup serves all four stages: Stage 1 uses only the corpora and the retriever, Stage 2 the three arms, Stage 3 the reader families and judges, and Stage 4 the logged search calls.

5 Stage 1: characterization predicts the verdict

Table 1 is the conventional headline table with one change: rows are grouped by the Stage-1 regime, assigned before any outcome was observed, and each delta carries a Stage-3 verdict. The grouping does the explanatory work.

On **given-document** corpora both arms read the same supplied report, so they tie by construction. On **absent-from-query** ObliQA the first retrieval misses $\sim 54\%$ of gold, yet across all three reader $\times$ retriever systems the agent reaches no additional gold: the two strong systems are TOST-equivalent within the a-priori margin, and the open-

Qwen system falls *behind* RAG. The claim that survives every reader and every judge concerns *evidence reached*, not answer quality: on GPT/dense the agent’s gold coverage (0.48) does not exceed RAG’s hit@8 (0.64), and on the open Qwen system the agent searches $2.4\times$ as often while reaching gold indistinguishably (0.766 vs. 0.758, agent vs. RAG). On **reformulation-recoverable** corpora every CI excludes zero, and the advantage grows with the coverage gap (Figure 1) (+0.08 on HotpotQA, +0.28 on MuSiQue). On **navigation-recoverable** ConditionalQA the agent also wins under every reader (+0.16 to +0.30) — but §6 shows that this gain has a different internal composition, which is invisible here and is exactly what Stage 2 exists to expose. The grouping survives a depth control: at $k=4$ the AFQ and given-document settings stay near zero while the recoverable ones remain positive and ordered (HotpotQA +0.14, MuSiQue +0.17).

5.1 Stage 1 within corpora, and its counterexample

Four corpora are too few, and too confounded with domain and metric, to establish that the Stage-1 coordinate is load-bearing. We therefore cut each corpus into **pre-defined subsets** that vary how hidden the evidence is while holding corpus, reader, and metric fixed. ObliQA splits by reference-group size (G1, G2–3, G4+); HotpotQA by hop type (*comparison*, *bridge-open*, *bridge-hidden*); MuSiQue by reasoning depth (2/3/4 hops); given-document ConvFinQA pins the zero-gap origin. For each subset we pair the a-priori, LLM-free coverage gap with the measured agent advantage (Table 2).

A weighted least-squares fit gives advantage = $-0.08 + 0.31 \cdot \text{gap}$ ($R^2 = 0.52$); we read the line as descriptive rather than as a calibrated effect size, because the y -axis pools native metrics onto a common $[0, 1]$ scale. The rank relationship is the meaningful one: Spearman $\rho = 0.62$ over all ten points and $\rho = 0.79$ over the seven non-ObliQA subsets. The same ordering holds *within* corpora, where the metric is fixed — HotpotQA from bridge-open (+0.05) to bridge-hidden (+0.21), MuSiQue across hop depth (+0.22 $\rightarrow$ +0.36).

The counterexample is the point of the exercise. ObliQA’s G2–3 and G4+ subsets have gaps of 0.65–0.68, comparable to the recoverable Wikipedia subsets, so a coverage-only characterization would predict a large agent gain. They gain nothing. A leave-one-subset-out model fit only on

corpus	Stage-1 regime	reader	RAG (R1)	agent (R3)	Δ	95% CI	Stage-3 verdict	γ	calls $\times$
ConvFinQA	given-doc	Qwen	0.654	0.654	+0.000	—	$\Delta=0$ constr.	—	—
		Claude	0.967	0.967	+0.000	—	$\Delta=0$ constr. (pilot)	—	—
ObliQA	AFQ	Claude/BM25	3.338	3.414	+0.076	$[-0.018, +0.168]$	TOST-equiv.	no R2	1.50
		GPT/dense	2.952	2.926	-0.026	$[-0.134, +0.082]$	TOST-equiv.	no R2	1.69
		Qwen/BM25	2.669	2.461	-0.208	$[-0.325, -0.092]$	RAG ahead	no R2	2.39
HotpotQA	reform.-rec.	Claude	0.702	0.778	+0.077	$[+0.045, +0.110]$	significant	65%	1.55
		GPT	0.629	0.793	+0.163	$[+0.128, +0.199]$	significant	92%	1.51
		Qwen	0.592	0.684	+0.093	$[+0.053, +0.133]$	significant	122%	2.33
MuSiQue	reform.-rec.	Claude	0.384	0.660	+0.276	$[+0.237, +0.316]$	significant	77%	2.63
		GPT	0.284	0.545	+0.261	$[+0.221, +0.303]$	significant	96%	2.55
		Qwen	0.184	0.336	+0.152	$[+0.112, +0.191]$	significant	58%	3.63
ConditionalQA	navigation-rec.	Claude	0.473	0.631	+0.159	$[+0.107, +0.210]$	significant	8%	1.54
		GPT	0.206	0.509	+0.303	$[+0.242, +0.364]$	significant	27%	2.11
		Qwen	0.262	0.465	+0.203	$[+0.147, +0.261]$	significant	41%	2.75
<i>Controls</i>									
SQuAD [‡]	directly reachable	Qwen	0.84	0.84	0.00	—	no agent arm	—	—
TAT-QA [†]	given-doc	Qwen	0.79	0.79	0.00	—	given-doc tie	—	—
FinQA [†]	given-doc	Qwen	0.48	0.48	0.00	—	given-doc tie	—	—

Table 1: The headline table under the protocol: rows are grouped by the *Stage-1 regime assigned before any outcome was observed*, and every cell carries a Stage-3 verdict rather than a bare delta. Native metric per corpus (ObliQA grade/5; HotpotQA/MuSiQue token-F1; ConditionalQA format-aware fair-score; ConvFinQA/FinQA exec-acc; TAT-QA score), so magnitudes are not comparable across rows. $N=500$ per cell except ConditionalQA (285), MuSiQue/Claude (498), ObliQA/Qwen (499), SQuAD (300), ConvFinQA/Claude pilot (30). Reading the Δ column alone gives a swing of -0.21 to $+0.30$ with no explanation; reading it against the regime column does. The two right-hand columns bring the Stage-2 gap-closing fraction γ (Table 4; 80-question paired subset) and the Stage-4 search-call multiplier calls $\times$ (Table 6) into the same view, so the conventional two-arm delta and what the four stages add can be read side by side; ObliQA carries no R2 arm. *Controls*: [‡] SQuAD has no full-agent arm — Δ is R2 vs. R1; [†] TAT-QA/FinQA are given-document reader-competence controls ($\Delta=0$ by construction). Per-corpus tests in App. B, Table 8.

corpus	subset	gap	Δ	n
ObliQA	G1	0.21	+0.02	139
	G2-3	0.65	-0.00	227
	G4+	0.68	-0.04	134
HotpotQA	bridge-open	0.27	+0.05	305
	comparison	0.39	+0.04	96
	bridge-hidden	1.00	+0.21	99
MuSiQue	2-hop	0.81	+0.22	251
	3-hop	0.97	+0.36	161
	4-hop	1.00	+0.28	86
ConvFinQA	given-doc	0.00	0.00	30

Table 2: Stage-1 coordinates vs. measured advantage over ten pre-defined subsets (Figure 2). The gap column is an LLM-free retrieval measurement fixed before any outcome was inspected; subsets are defined by question structure. ObliQA’s high-gap subsets are the diagnostic counterexample: as under-covered as the recoverable ones, with no advantage.

the seven recoverable subsets predicts $+0.18$ for these points against a measured ≈ 0 , while achieving LOSO $r=0.87$ inside the recoverable regime. Since the regimes were assigned a priori and no curve was fit to produce them, this is a genuine

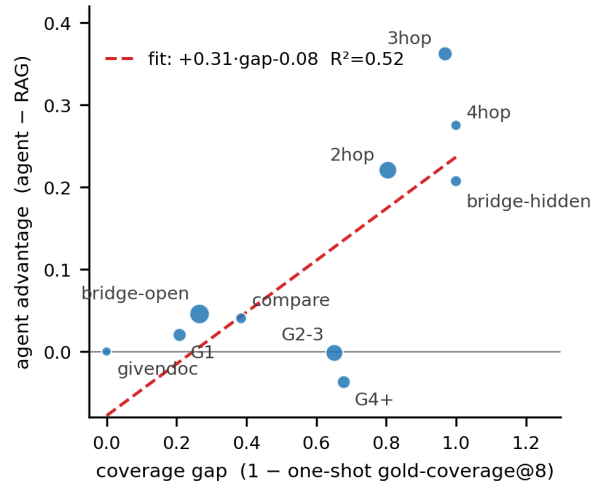


Figure 2: Agent advantage vs. a-priori coverage gap across the ten subsets of Table 2. Advantage rises with the gap when the missing evidence is recoverable; ObliQA’s high-gap subsets sit *below* the line because their misses are absent-from-query. The residual, not the line, is the informative part. App. F adds the ConditionalQA point.

out-of-sample failure of the coverage-only diagnostic. Stage 1 therefore needs the channel probe, not

corpus	R1 ($r=1$)	2-rd	3-rd	5-rd	R3 (agent)
MuSiQue	0.184	0.286	0.304	0.312	0.336
HotpotQA	0.592	0.695	0.707	0.702	0.684
ConditionalQA	0.262	0.317	0.315	0.311	0.465

Table 3: Round-budget sweep of the R2 ablation instrument (open Qwen reader, full N , $k=8$; each corpus’s native metric). On the reformulation-recoverable corpora the choice-free loop exceeds the agent on HotpotQA and closes most of the MuSiQue gap by round 3. On navigation-recoverable ConditionalQA it saturates at ≈ 0.31 from round 2 onward and never approaches the agent’s 0.465 even at five rounds — so the residual is not a budget problem.

just the gap: the PRF probe recovers only 1.5% of ObliQA’s uncovered gold against 19.8% of HotpotQA’s.

6 Stage 2: credit assignment

Stage 1 tells us *whether* extra retrieval can help; Stage 2 asks *which component* delivers it. We hold the number of retrieval opportunities roughly fixed and remove the choice: R2 runs a fixed number of reformulate-and-retrieve rounds at a budget comparable to the agent’s.

On reformulation-recoverable corpora, most of the reported agent gain is iteration. At full N under the open reader (Table 3), R2 closes +0.12 of MuSiQue’s +0.15 agent gap and *exceeds* the agent’s HotpotQA score. Under the two stronger readers, an 80-question paired-subset replication gives γ between 65% and 96% (Table 4). The gain is spent by rounds two to three; round 5 adds nothing. Under a conventional two-arm report, all of this would have been filed as evidence for open-ended agency.

On the navigation-recoverable corpus, agency is the residual. ConditionalQA is the longest-horizon corpus here in the sense that matters: recovering an answer requires reading a specific passage, listing its neighbors, and deciding when enough conditions are gathered. R2 is flat from round 2 (≈ 0.31) and never reaches the agent’s 0.465; γ is 41%/27%/8% under Qwen/GPT/Claude, against 65–96% on the wiki corpora under the two stronger readers (Table 4). We read γ as an operational estimate of R_{ref} ’s share of total recoverability, and the +0.15 residual as the one place here where agency, not iteration, does the work.

γ is stable across the axes an evaluation would

		MuSiQue	HotpotQA	CondQA
iter. gain $r3-R1$	$k=8$	+0.12	+0.12	+0.05
	$k=4$	+0.13	+0.13	+0.09
gap-closing fraction γ	Qwen	58%	122%	41%
	GPT	96%	92%	27%
	Claude	77%	65%	8%

Table 4: Robustness of the Stage-2 statistic. *Top:* the iteration gain is undiminished — slightly larger — at $k=4$, so it is reformulation rather than retrieval breadth. *Bottom:* $\gamma = (R2 - R1)/(R3 - R1)$ under three reader families (80-question paired subset; Claude, and ConditionalQA/GPT, against their full- N agent anchors; HotpotQA/Qwen exceeds 100% because R2 beats the agent). γ separates the two recoverable regimes under *every* reader.

vary: undiminished at $k=4$, replicated under all three reader families, and preserved under dense retrieval (FAISS over bge-base), where R2 still closes 95% of MuSiQue’s gap and exceeds the agent on HotpotQA while closing 16% of ConditionalQA’s. A negative control confirms that γ measures the corpus rather than rewarding extra calls: on single-hop SQuAD, where one-shot coverage is already 0.91, R2 yields $\Delta F1 = 0.00$ at $N=300$ even though its reformulations lift $\text{recall}@k$ from 0.91 to 0.94 — it recovers passages the reader did not need.

6.1 Behavioral evidence for the same split

Stage 2’s conclusion is a claim about mechanism, so we check it against what the agent actually did. We classify every logged episode from its ordered query log (Table 5) into **one-shot** (a single search), **parallel** (≥ 2 searches whose entities all already appear in the question), and **sequential** (a later query names a *new* entity discovered from a prior result). Sequential share is a behavioral proxy for realized horizon.

The **strong** agents adapt: on MuSiQue, Claude abandons one-shot retrieval almost entirely (7.6%) and chains sequentially 85% of the time (GPT 58%), while on ObliQA both stay near 60% one-shot. For each strong reader the sequential share orders the corpora exactly as the measured gain does (Claude: ObliQA 21% $\approx$ HotpotQA 25% $\ll$ MuSiQue 85%). This supports the Stage-2 reading twice over. Chaining, not repetition, is the mechanism — which is why one-shot RAG cannot emulate it. But chaining *is* a form of reformulation, which R2 does emulate — which is why γ is high on the wiki corpora. Structural navigation is the behavior R2 cannot emulate, and correspond-

corpus	agent	searches	1-shot	par.	seq.
ObliQA	Claude	1.50	61.8	17.0	21.2
	GPT	1.69	60.4	18.4	21.2
	Qwen	2.39	15.2	0.8	84.0
HotpotQA	Claude	1.55	50.6	24.8	24.6
	GPT	1.51	66.0	13.8	20.2
	Qwen	2.33	23.4	36.6	40.0
MuSiQue	Claude	2.63	7.6	7.4	85.0
	GPT	2.55	32.4	9.2	58.4
	Qwen	3.63	3.8	23.0	73.2

Table 5: Agent search strategy (% of episodes) under one entity-novelty classifier. Strong agents match their realized horizon to the corpus; the open reader does not, chaining 84% of the time on ObliQA where there is nothing to chain.

ingly `get_passage/list_doc_passages` calls appear in 223 of 500 ObliQA traces and in *zero* HotpotQA or MuSiQue episodes. The open reader is an instructive negative: it chains where there is nothing to chain (84% on ObliQA) and reaches no additional gold, which is why search-call count alone is not a quality signal and should not be reported as one.

7 Stage 3: checks that changed a verdict

We report the three Stage-3 checks that were not merely confirmatory.

Equivalence testing turned an ObliQA non-result into a testable claim. A bare $p > 0.05$ on the two strong ObliQA systems would have licensed no conclusion at all. Under pre-registered TOST margins (grade ± 0.25 , a quarter of a grade point and below the just-noticeable difference; rates ± 0.05), both are positively equivalent ($p=4\times 10^{-5}$ and 2×10^{-4}), while the open-Qwen system is *not* equivalent and favors RAG (Table 8).

The artifact audit caught a sign error on one corpus. Naive token-F1 on ConditionalQA favored RAG by 0.10; format-aware scoring reversed the sign. Having found one artifact we audited the rest: on the wiki corpora the agent’s exact-match gain equals its F1 gain ($\Delta\text{EM}/\Delta\text{F1} = 1.04$ and 1.06) and its answers are no longer than RAG’s, so those advantages are not verbosity effects (Table 9).

Cross-family judge re-grading changed a magnitude by $3.7\times$ without changing the conclusion (App. D). ObliQA is graded by an LLM judge (Zheng et al., 2023), with known self-preference and reliability risks (Liu et al., 2023; Wang et al., 2024). Re-grading both strong systems with an independent local open-model judge over identical

corpus	reader	$\Delta\text{qual.}$	$\text{calls}\times$	$\text{tok}\times$	$\text{lat}\times$	verdict
ObliQA	GPT	−0.005	1.69	4.4	3.1	✗ no
	Claude	+0.015	1.50	—	—	✗ no
	Qwen	−0.042	2.39	—	—	✗ no (tax)
HotpotQA	Claude	+0.077	1.55	—	—	route
	GPT	+0.163	1.51	—	—	route
	Qwen	+0.093	2.33	—	—	iterate
MuSiQue	Claude	+0.276	2.63	—	—	iterate
	GPT	+0.261	2.55	—	—	iterate
	Qwen	+0.152	3.63	—	—	iterate
Cond. QA	Claude	+0.159	1.54	—	—	✓ agent
	GPT	+0.303	2.11	—	—	✓ agent
	Qwen	+0.203	2.75	—	—	✓ agent

Table 6: Agent cost relative to one-shot RAG. $\Delta\text{qual.}$ is agent-minus-RAG (grade/5 for ObliQA); $\text{tok}\times/\text{lat}\times$ measured only on ObliQA/GPT. Verdicts follow a fixed rule applied per cell: **✗** when $\Delta\text{qual.}$ is inside the equivalence margin; “route” when a per-query signal recovers the gain more cheaply than blanket escalation; “iterate” when R2 matches the agent (§6); **✓** when the residual after R2 justifies the full agent.

stored prompts gives high item-level agreement (Pearson 0.85–0.86, within-one-grade 95–98%) but a judge-sensitive *magnitude*: on Claude/BM25 the agent’s edge widens from +0.076 to +0.282, and the open Qwen system mirrors this under the GPT judge (−0.062 $\rightarrow$ −0.208). The disagreement is not explained by refusals (unchanged at +0.281 on the 381 questions where both arms answer substantively) and only weakly tracks length ($r=0.17$). Two lessons follow: a single-judge magnitude on a long-horizon generative task should be reported as judge-conditional, and an evaluation should anchor its load-bearing claim on a quantity the judges cannot disagree about. The agent reaches no additional gold in any ObliQA system, so “the agent reaches no additional evidence” survives every judge, whereas “the agent’s answers are no better” does not.

8 Stage 4: pricing the horizon

Quality and cost both rise with recoverability, so the relevant object is a frontier, not a delta (Figure 3, Table 6). In retrieval/search calls — the only unit available on every corpus — the agent runs $1.5\text{--}3.6\times$ RAG’s single shot. Calls understate the true expense, since each agent step is a full reader call: on the one fully instrumented system (ObliQA/GPT) the agent costs $3.1\times$ the latency and $4.4\times$ the input tokens of one-shot RAG for *no* quality gain.

Cost-aware reporting also exposes when a cheap signal suffices. As a retrospective ceiling

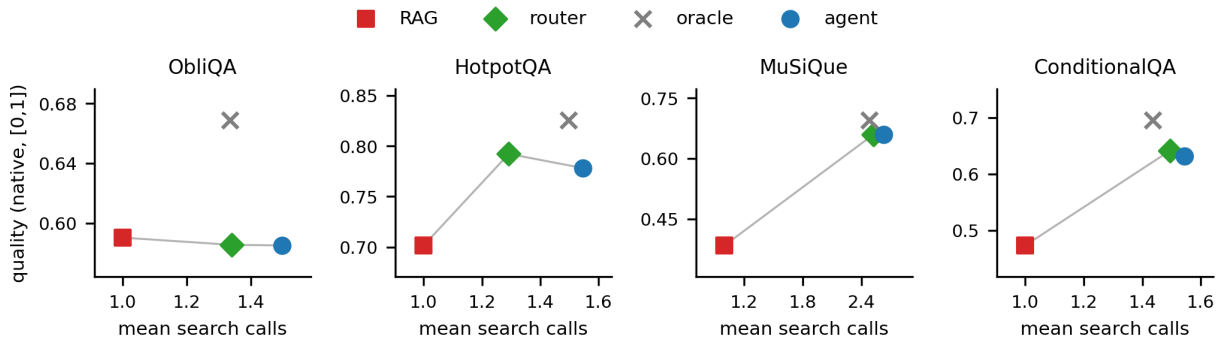


Figure 3: Quality vs. cost (mean search calls) for four policies under the Claude reader (ObliQA panel: GPT/dense); oracle and router are analysis instruments, not deployable systems. The frontier is flat on ObliQA (paying more buys nothing); the per-query router sits above the full agent at lower cost on mixed-recoverability HotpotQA; and only on the high-recoverability corpora does the frontier reward full agent cost.

estimate (App. C) we route a query to the agent iff its gold is not fully covered by one-shot BM25 top-8 — LLM-free but label-dependent, so an analysis instrument rather than a policy. It captures 66–79% of the oracle’s gain over RAG across the three readers while escalating 43% of HotpotQA queries. On ObliQA it *over-escalates* and, except under Claude (+14%), loses (−6% to −21%), reproducing the Stage-1 failure of coverage-only characterization at policy level (Table 10). Gating on the PRF probe removes that over-escalation almost entirely (0.8% escalated rather than 53.6%), but one lexical expansion recovers only 4.1% of MuSiQue’s uncovered gold and under-escalates on deep chains: the recoverability test that works is one that iterates, i.e. R2 itself.

9 What we recommend reporting

The protocol reduces to a checklist. (1) State each benchmark’s a-priori coverage gap and recovery channel and group results by regime; a suite with no AFQ member and no navigation-recoverable member cannot distinguish iteration from agency at all. (2) Report three arms, the gap-closing fraction γ , and a round-budget sweep for R2; a two-arm table cannot support a claim about agency. (3) Test null cells for equivalence against pre-registered margins, replicate across ≥ 2 reader families, audit for metric artifacts, and report LLM-judge magnitudes as judge-conditional. (4) Report quality against calls, and against tokens and latency where instrumented.

Applied here, the checklist yields a falsifiable summary of which architecture a task justifies: one-shot RAG when evidence is directly reachable; fixed iteration when misses are reformulation-recoverable, which covers most of what the multi-

hop literature currently attributes to agency; a full agent when misses require structural navigation; and no escalation when misses are absent-from-query, where the right intervention is to fix the task, metadata, or input context rather than the architecture.

10 Related work

Agentic and iterative retrieval. RAG couples a retriever with a reader (Lewis et al., 2020; Karpukhin et al., 2020). Agentic extensions let models ask follow-ups, call tools, or critique retrieval mid-generation (Yao et al., 2023; Schick et al., 2023; Press et al., 2023; Trivedi et al., 2023; Jiang et al., 2023; Asai et al., 2024). We treat these mechanisms as the object of measurement rather than proposing another; our R2 arm is deliberately a stripped-down IRCOT used as an instrument.

Difficulty- and routing-based accounts. Adaptive-RAG routes by predicted question complexity (Jeong et al., 2024) and GRADE separates retriever- from generator-side difficulty (Lee et al., 2025). Our diagnostic is narrower and evaluation-facing: we split retriever-side difficulty by *recovery channel* rather than by amount, which explains why difficulty-based escalation over-escalates on corpora like ObliQA.

Agent evaluation and benchmark validity. The closest empirical result is Ferrazzi et al. (2026), who find that strong non-agentic systems often match agentic RAG at lower cost — establishing *that* the headline gain can vanish, where we identify *when* (the Stage-1 regime) and *which component* was responsible (the Stage-2 γ). Stage 3 draws on the LLM-as-judge reliability literature (Zheng et al., 2023; Liu et al., 2023; Wang et al., 2024)

and standard equivalence and multiple-comparison practice (Lakens, 2017; Holm, 1979; Efron and Tibshirani, 1993).

11 Conclusion

A single agent-minus-baseline number on a long-horizon benchmark confounds a property of the benchmark with a claim about the agent’s architecture. The four-stage protocol separates them: a-priori recoverability characterization predicts the sign of the measured advantage where coverage gap alone mispredicts it out of sample, and the ablation ladder shows that 65–96% of what is normally credited to agency (under the two stronger readers) is delivered by a choice-free loop — while isolating the one navigation-recoverable regime where agency retains a residual that no round budget closes.

Limitations

The dataset taxonomy is intentionally heterogeneous, and each row supports a different claim; we do not claim a uniform design across rows. The fully evaluated three-arm comparisons are HotpotQA, MuSiQue, and ConditionalQA; ObliQA carries R1 and R3 only, because Stage 1 predicts and Stage 3 confirms that R3 adds nothing there, so an R2 arm would have no gap to close. The controls are narrower by design: SQuAD has no full-agent arm, and ConvFinQA’s given-document equivalence is true by construction, so we treat it as anchoring rather than as testing the retrieval contrast.

The navigation-recoverable regime rests on a single corpus. ConditionalQA is the only benchmark here whose misses require structural actions, so “agency pays on navigation-recoverable tasks” is supported by three reader families on one corpus rather than by several corpora, and we flag this as the weakest link in the protocol’s coverage. R_{nav} is also measured less directly than R_{ref} : we estimate it as the residual $1 - \gamma$ plus trace-level tool-use evidence, rather than with a standalone LLM-free probe of the kind Stage 1 provides for R_{ref} .

The agent results use a single harness, and a different harness could shift absolute numbers, although the direction tracks the Stage-1 regime across five independent corpora and replicates at full N under three reader families. ObliQA depends on an LLM judge; §7 reports that dependence rather than resolving it, and a human-

agreement study remains future work. Finally, the parallel-vs-sequential classification is an entity-novelty heuristic over the query log rather than a hand-verified label, and token/latency instrumentation exists on only one system, so the cross-corpus cost unit is search calls.

Ethics Statement

All corpora are public and no proprietary data is used. Regulatory and financial QA carries a compliance-advice misuse risk; we treat all outputs as decision support rather than advice, and we note that the AFQ finding is itself safety-relevant — on that corpus the agent produces more answers than one-shot RAG (21% vs. 35% refusal) without reaching more evidence, which is a failure mode worth measuring rather than a capability. Readers and agents are hosted on open LLMs with pinned identifiers and dates. All corpus builders, characterization probes, scoring scripts, and raw generations and grades will be released, and ObliQA is additionally graded with a fully local open-model judge so that the headline scoring is reproducible without any metered API.

Use of AI assistants. LLM-based coding and writing assistants were used to scaffold experiment and scoring code, to edit prose and $\LaTeX$, and to check consistency between tables and text. The experimental design, data handling, analyses, reported numbers, and claims are the authors’ own, and the authors verified and take full responsibility for all content.

References

- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of ICLR*.
- Islam Aushev, Egor Kratkov, Evgenii Nikolaev, Andrei Glinskii, Vasilii Krikunov, Alexander Panchenko, Vasily Kononov, and Julia Belikova. 2025. RAGulator: Effective RAG for regulatory question answering. In *Proceedings of the 1st Regulatory NLP Workshop (RegNLP 2025)*, pages 114–120.
- Zhiyu Chen, Wenhui Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of EMNLP*.
- Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Con-

- vFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of EMNLP*.
- Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall/CRC.
- Pietro Ferrazzi, Milica Cvjetanin, Alessio Piraccini, and Davide Giannuzzi. 2026. Is agentic RAG worth it? An experimental comparison of RAG approaches. arXiv preprint arXiv:2601.07711.
- Tuba Gokhan, Kexin Wang, Iryna Gurevych, and Ted Briscoe. 2024. RIRAG: Regulatory information retrieval and answer generation. *arXiv preprint arXiv:2409.05677*.
- Tuba Gokhan, Kexin Wang, Iryna Gurevych, and Ted Briscoe. 2025. Shared task RIRAG-2025: Regulatory information retrieval and answer generation. In *Proceedings of the 1st Regulatory NLP Workshop (RegNLP 2025)*, pages 1–4.
- Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.
- Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of NAACL*.
- Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of EMNLP*.
- Vladimir Karpukhin, Barlas Öguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*.
- Daniël Lakens. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362.
- Jeongsoo Lee, Daeyong Kwon, and Kyohoon Jin. 2025. GRADE: Generating multi-hop QA and fine-grained difficulty matrix for RAG evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4405–4424.
- 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. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Yang Liu, Dan Iter, Yichong Xu, Shuhang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of EMNLP*.
- Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.
- Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of EMNLP*.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*.
- Joseph John Rocchio. 1971. Relevance feedback in information retrieval. In Gerard Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall, Englewood Cliffs, NJ.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Haitian Sun, William W. Cohen, and Ruslan Salakhutdinov. 2022. ConditionalQA: A complex reading comprehension dataset with conditional answers. In *Proceedings of ACL*.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of ACL*, pages 10014–10037.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of ACL*.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of EMNLP*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *Proceedings of ICLR*.
- 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. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of ACL-IJCNLP*.

A Dataset details

Table 7 places every dataset on the recoverability coordinates and records which protocol arms are measured for each.

B Significance tests and artifact checks

Per corpus we compute a paired bootstrap (Efron and Tibshirani, 1993) (95% CI, 10k resamples, seed 27), a paired permutation test (10k), a-priori TOST equivalence (Lakens, 2017) for tie claims (margins fixed before viewing any test-set delta: grade ± 0.25 , below the just-noticeable difference; rates ± 0.05), and McNemar (McNemar, 1947) on the binarized outcome; difference claims are Holm-Bonferroni corrected (Holm, 1979).

C Router validation

D Judge sensitivity on ObliQA

Table 11 reports the ObliQA grade gap under each judge family. Item-level agreement is high throughout, but the magnitude is judge-sensitive on two systems in opposite directions — a judge-dependent assessment of answer quality given the *same* evidence, not a retrieval difference. On the GPT/dense system, citation-in-corpus is 93=93 (TOST-equivalent, $p = 0.004$) and the one architectural difference is refusal (RAG 35% vs. agent 21%, $p < 0.001$): the agent answers more, not more faithfully, and that extra answering is not bought with extra evidence.

E Two corpora that look agent-favorable and are not

Table 12 gives representative items for both corpora. **Difficulty without retrieval dependency: ConvFinQA.** ConvFinQA is hard — 72% of turns are history-dependent follow-ups and 67% require a computed program — yet its evidence is a single *given* report, so the agent’s retrieval freedom has nothing to act on. The control is not vacuous, because the reader is competent once evidence is supplied: the open reader scores 65.4% [61.1, 69.4] at $N=500$ given the document, and the same reader scores 0.79 on TAT-QA and 0.48 on FinQA. To obtain a non-trivial Δ we also run ConvFinQA pooled

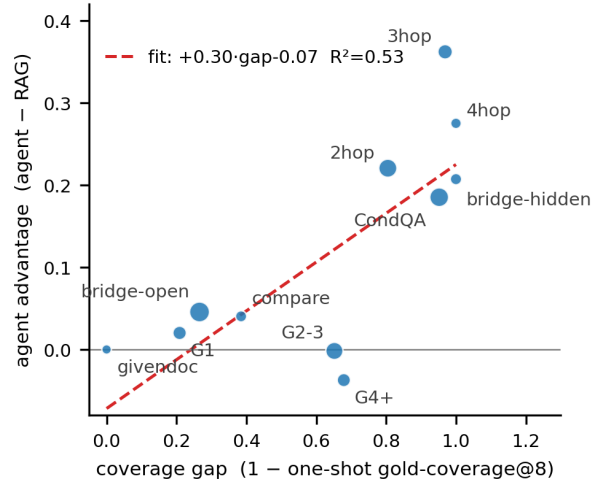


Figure 4: Agent advantage vs. a-priori coverage gap across all 11 analysis subsets, including ConditionalQA.

over all 1,806 reports ($N=300$), the purest AFQ case: the agent does *not* beat RAG ($\Delta = -0.07$, CI $[-0.13, -0.02]$), even though its reformulations do lift gold-report recall ($0.62 \rightarrow 0.67$) — the bottleneck is un-addressability, not search effort.

Two LLM-free attributions. (i) On ObliQA, decomposing each question into ~ 3 sub-queries and unioning retrievals at matched budget is indistinguishable from one shot (87% exact ties), and following the real cross-reference graph rescues zero navigation-exclusive gold. (ii) In a retriever $\times$ architecture 2×2 , the agent shows no difference on ObliQA and HotpotQA-not-hidden under *both* dense and BM25, and helps on HotpotQA-hidden under *both*; the retriever is a large orthogonal main effect (dense $\rightarrow$ BM25 moved the ObliQA grade $+0.52$).

F Cross-subset trend including the navigation case

Figure 4 repeats Figure 2 with the ConditionalQA point added. It sits off the reformulation trend by construction: its coverage gap predicts an advantage that reformulation cannot deliver (Table 3) and that only structural navigation does, which is the second dimension the plane of Figure 1 makes explicit.

G Additional benchmark context

HotpotQA (Yang et al., 2018) and MuSiQue (Trivedi et al., 2022) supply bridge/comparison and 2–4-hop questions; MuSiQue is constructed to resist retrieval shortcuts. ConditionalQA (Sun et al., 2022) targets conditional answers over long gov-

Dataset	Task/domain	R1	R2	R3	Outcome metric	Stage-1 probe	Role
ObliQA (FSRA)	Regulatory QA over mutually referenced rules	Yes	No	Yes	LLM grade/5	Coverage gap, PRF probe, trace-derived R_{nav}	Absent-from-query anchor
HotpotQA	Wikipedia multi-hop QA	Yes	Yes	Yes	Token F1 / EM	Coverage gap, PRF probe, hidden-subset split	Reformulation-recoverable
MuSiQue	Wikipedia compositional QA	Yes	Yes	Yes	Token F1 / EM	Coverage gap, PRF probe, hop-depth split	Reformulation-recoverable, long horizon
ConditionalQA	UK government conditional QA	Yes	Yes	Yes	Format-aware fair-score	Coverage gap, low R_{ref} , structural actions	Navigation-recoverable, long horizon
SQuAD	Single-hop Wikipedia QA	Yes	Yes	No	Token F1 / EM	One-shot coverage, fixed-loop recall gain	Control: directly reachable
ConvFinQA (given)	Conversational financial QA, report provided	Yes	No	Yes	Executed-answer accuracy	Zero gap by construction	Control: given-document
ConvFinQA (pool)	Financial QA under pooled retrieval	Yes	No	Yes	Executed-answer accuracy	Coverage gap, low R_{ref}	Control: absent-from-query
TAT-QA [†]	Financial table/text QA, document provided	Yes	No	No	Dataset score	Zero gap by construction	Control: reader competence
FinQA [†]	Financial numerical QA, document provided	Yes	No	No	Executed-answer accuracy	Zero gap by construction	Control: reader competence

Table 7: Dataset taxonomy and available measurements. Arms are deliberately not uniform: the protocol requires R2 only where Stage 1 predicts a recoverable gap, and requires R3 only where a retrieval contrast exists. [†]: given-document control.

ernment documents. As controls we add single-hop SQuAD (Rajpurkar et al., 2016) and the given-document financial sets TAT-QA (Zhu et al., 2021) and FinQA (Chen et al., 2021). For the financial and regulatory domain we use ConvFinQA (Chen et al., 2022), built on FinQA, and the RegNLP/ObliQA benchmark (Gokhan et al., 2024), whose RIRAG-2025 shared task (Gokhan et al., 2025) drew systems such as the BM25-prefiltered RAGulator (Aushev et al., 2025) — reinforcing that strong lexical retrieval is the regulatory norm our R1 arm adopts. We treat these not as leaderboards but as points on a recoverability plane.

H RAG-wins / agent-inert traces (ObliQA, the equivalence corpus)

To substantiate the claim that we report equivalence and divergence with equal weight (§7), we show the direction the headline tables do *not* foreground: cases where one-shot RAG beats the agent on the low-dependency regulatory corpus. On ObliQA (Claude/BM25, $N = 500$) the two arms are near-mirror images: **agent-wins 145, RAG-wins 112, ties 243**—a 49% exact-grade tie rate and a net

agent edge of only +0.076 grade. Where the gold is reachable in one query, the agent’s extra retrieval is at best inert and sometimes *harmful*: it over-navigates into a refusal or drifts to a related-but-wrong passage. Three representative cases where one-shot RAG scored strictly higher follow.

Q. What metrics or indicators are recommended or required to communicate the environmental and social impacts of a Petroleum Reporting Entity?

RAG (1-shot, top- k): answers from 12.3.1, correctly noting no specific metrics are prescribed. ✓

(A)

Agent: “Not found in provided context” — refused despite the gold passage being retrievable. ✗ **(F)**

Q. What is the required audit process for verifying the origin of Accepted Spot Commodities under ADGM regulation?

RAG (1-shot, top- k): states the firm must have delivery/storage arrangements covering approval, management, monitoring and audit (22.4.4). ✓ **(A)**

Agent: reformulated toward “third-party verification,” retrieved a related passage, and missed the core delivery-arrangement requirement. ✗ **(D)**

corpus	reader	claim	n	R1	R3	Δ	95% CI	verdict
ConvFinQA	Qwen	equiv. (constr.)	500	0.654	0.654	+0.000	—	$\Delta=0$ by construction
	Claude	equiv. (constr.)	30	0.967	0.967	+0.000	—	$\Delta=0$ by construction (pilot)
ObliQA	GPT / dense	equivalence	500	2.952	2.926	-0.026	$[-0.134, +0.082]$	TOST-equiv. ($p=4\times 10^{-5}$)
	Claude / BM25	equivalence	500	3.338	3.414	+0.076	$[-0.018, +0.168]$	TOST-equiv. ($p=2\times 10^{-4}$)
	Qwen / BM25	equivalence	499	2.669	2.461	-0.208	$[-0.325, -0.092]$	not equiv.—RAG ahead
HotpotQA	Claude	difference	500	0.702	0.778	+0.077	$[+0.045, +0.110]$	significant (Holm $p<10^{-3}$)
	GPT	difference	500	0.629	0.793	+0.163	$[+0.128, +0.199]$	significant (CI excl. 0)
	Qwen	difference	500	0.592	0.684	+0.093	$[+0.053, +0.133]$	significant (CI excl. 0)
MuSiQue	Claude	difference	498	0.384	0.660	+0.276	$[+0.237, +0.316]$	significant (Holm $p<10^{-3}$)
	GPT	difference	500	0.284	0.545	+0.261	$[+0.221, +0.303]$	significant (CI excl. 0)
	Qwen	difference	500	0.184	0.336	+0.152	$[+0.112, +0.191]$	significant (CI excl. 0)
ConditionalQA	Claude	difference	285	0.473	0.631	+0.159	$[+0.107, +0.210]$	significant (Holm $p<10^{-3}$)
	GPT	difference	285	0.206	0.509	+0.303	$[+0.242, +0.364]$	significant (CI excl. 0)
	Qwen	difference	285	0.262	0.465	+0.203	$[+0.147, +0.261]$	significant (CI excl. 0)

Table 8: Stage-3 tests per reader family. The three Claude-reader difference corpora reach the permutation-test resolution floor (raw $p < 10^{-4}$ at 10,000 resamples); we report the Holm-corrected bound $p < 10^{-3}$ rather than a point estimate. GPT and Qwen rows are full- N replications reported by paired-bootstrap CI and lie outside the Holm family.

corpus	reader	$\Delta F1$	ΔEM	$\frac{\Delta EM}{\Delta F1}$	len (R/A/g)
HotpotQA	Claude	+0.077	+0.080	1.04	2.2 / 2.3 / 2.4
MuSiQue	Claude	+0.276	+0.293	1.06	2.2 / 2.4 / 2.7

Table 9: Stage-3 metric-artifact audit. The agent’s exact-match gain equals its F1 gain and its answers are no longer than RAG’s, so the wiki advantages are not verbosity effects. “—”: not measured in this audit.

Q. Are there collaborative mechanisms for Relevant Persons to share best practices on risk assessments of Restricted Scope Companies?

RAG (1-shot, top- k): correctly answers that no general sharing mechanism exists; the regulation points to bilateral dialogue under the risk-based approach. ✓ **(B)**

Agent: refused (“Not found”) despite the relevant context. ✗ **(F)**

corpus	reader	all-R1	router	oracle	all-R3	escalated	gain captured	extra cost
HotpotQA	Claude	0.702	0.792	0.826	0.778	43%	73%	54%
	GPT	0.629	0.783	0.823	0.792	43%	79%	68%
	Qwen	0.592	0.701	0.759	0.684	43%	66%	52%
MuSiQue	Claude	0.384	0.659	0.695	0.660	89%	88%	94%
	GPT	0.284	0.549	0.593	0.545	89%	86%	94%
	Qwen	0.184	0.342	0.397	0.336	89%	74%	93%
ConditionalQA	Claude	0.473	0.641	0.696	0.631	90%	76%	91%
	GPT	0.206	0.516	0.566	0.509	91%	86%	91%
	Qwen	0.262	0.462	0.525	0.465	91%	76%	91%
ObliQA	GPT	0.590	0.586	0.669	0.585	54%	−6%	68%
	Claude	0.668	0.679	0.745	0.683	54%	14%	68%
	Qwen	0.534	0.519	0.606	0.492	54%	−21%	63%

Table 10: Coverage-only router validation, by reader family (quality on each corpus’s native metric rescaled to $[0, 1]$; ObliQA rows are grade/5 under the GPT judge; cost in mean retrieval/search calls; the coverage signal, and hence the escalation %, is reader-independent). The pattern replicates across all three families and reproduces the Stage-1 failure mode: useless-to-harmful on absent-from-query ObliQA (+14% to −21%).

system	Δ_{GPT}	Δ_{qwen}	agree
GPT / dense	−0.026 (eq.)	+0.035 (eq.)	$r=0.85$
Claude / BM25	+0.076 (eq.)	+0.282 (sig.)	$r=0.86$
Qwen / BM25	−0.208 (sig.)	−0.062 (eq.)	$r=0.87$

Table 11: ObliQA agent–RAG grade gap under each judge ($N=500$ per system). Within-one-grade agreement is 95–98% throughout.

corpus	a representative question (verbatim)	why it <i>looks</i> agent-favorable	what actually decides it ($\Delta = \text{Agent} - \text{RAG}$)
ObliQA absent-from-query	“... acceptable arrangements for ... control of <i>storage facilities</i> under COBS Rule 22.4.3? ”	A second <i>required</i> clause—the AML/CFT obligation for the same spot-commodity activity—sits in a different part of the rulebook, so it looks like the agent must chase a cross-reference.	Nothing in the question names “AML” or “money laundering,” so no reformulation can target that clause; the agent reaches no more gold than one shot. $\Delta \approx 0$.
ConvFinQA given document	“and how much does <i>that change</i> represent in relation to <i>this 2005</i> weighted average exercise price?”	The follow-up is uninterpretable without the dialogue history and needs a two-step program (subtract then divide)—textbook multi-step reasoning.	Both numbers (60.94, 25.14) are in the <i>one given table</i> ; both arms read the same document and run the same executor. The difficulty is reasoning, not retrieval. $\Delta = 0$ by construction.
HotpotQA hidden-but-findable	“What government position was held by the woman who portrayed Corliss Archer in the film <i>Kiss and Tell?</i> ”	The answer hinges on a person the question never names.	Hop 1 reveals the bridge entity <i>Shirley Temple</i> ; a query naming her then retrieves “Chief of Protocol.” Iteration turns the miss into the answer. +0.21 F1 on the hidden subset.
MuSiQue hidden-but-findable	“Who is the grandmother of Philippe, Duke of Orléans? ”	The grandmother is two hops away; the intermediate parent is unnamed.	Hop 1 → the father <i>Louis XIII</i> ; a query naming him → “Marie de’ Medici.” One-shot retrieval stops at the parent. +0.28 F1 .

Table 12: One representative question per corpus. The upper two *look* well suited to an iterative agent, yet the agent earns nothing, because the missing or hard part is not reachable by another query. The lower two look no different on the surface, but their missing evidence is hidden-but-findable. A retrieval-side coverage probe alone cannot tell the upper pair from the lower — which is why Stage 1 requires a channel probe.