COST OR COVERAGE: AGENT ORCHESTRATION UNDER ACCESS SCOPES

Anonymous authors
Paper under double-blind review

ABSTRACT

As LLM-based multi-agent systems grow more complex, orchestration increasingly requires coordinating specialized agents while controlling resource use. When agents have overlapping access to information sources and different engagement costs, this becomes a joint inference-and-allocation problem: the orchestrator must infer which sources contain the required evidence and select a low-cost agent set that covers them. Whether current LLMs can coordinate these decisions remains unclear. To quantify the orchestrators' coordination capabilities, we compile a benchmark containing 178 two-source queries, a fixed team of eight access-scoped agents, two retrieval-grounded engagement-cost schemes, and oracle retrieval that isolates orchestration from retriever error. Across eight model configurations, first-attempt behavior separates into narrow commitment, which limits cost but frequently misses required sources, and broad hedging, which achieves higher coverage through substantial over-engagement. Revealing the required sources largely resolves the former but not the latter. In the revision round, outcome feedback facilitates correction, while cost feedback can induce more restrained routing but does not reliably reduce cost without sacrificing task success. Overall, models respond well to individual source or cost signals, but struggle to balance them when both coverage and cost must be optimized jointly.

1 INTRODUCTION

As LLM-based multi-agent systems become increasingly common in complex applications, effective orchestration is becoming an important part of how these systems operate. A key challenge is completing tasks efficiently when multiple agents can contribute in different ways. Most prior work on LLM orchestration focuses on workflow design, communication structure, or agent participation (Zhuge et al., 2024; Zhang et al., 2025b;a; Wang et al., 2025b). Recent work also considers resource-aware allocation across heterogeneous agents (Amayuelas et al., 2025; Su et al., 2026; Cui et al., 2026). However, a harder and previously unexplored problem emerges when the information required by a query is not known in advance: the orchestrator must infer which sources are needed and then choose a low-cost set of agents whose combined access covers them. A model may perform well in either source identification or agent selection, but when both must be solved together, the problem becomes a lot more complex. We refer to this problem as "cost-or-coverage". Cost-or-coverage is a joint source-inference and agent-allocation problem.

In this work, we establish a benchmark, `CoC-Bench`, to study cost-or-coverage in a controlled setting. We construct 178 queries from EnterpriseRAG-Bench (Sun et al., 2026b), an enterprise knowledge corpus. To answer each query, the model is required to gather evidence from two distinct source units, and orchestrate a fixed team of eight agents with overlapping access to these sources. Agents incur different engagement costs under two retrieval-grounded pricing schemes: one based on the total breadth of an agent's searchable scope and the other on the breadth of the keyword-matched candidate set within that scope. We use oracle retrieval so that failures reflect orchestration decisions rather than retriever error. We further introduce diagnostic conditions that reveal the required sources, vary price information, reassign source access across agents, and provide second-round feedback.

Across eight model configurations, first-attempt routing shows two distinct behaviors. Some models make narrow assignments and often miss required sources. Others route broadly and achieve higher

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.

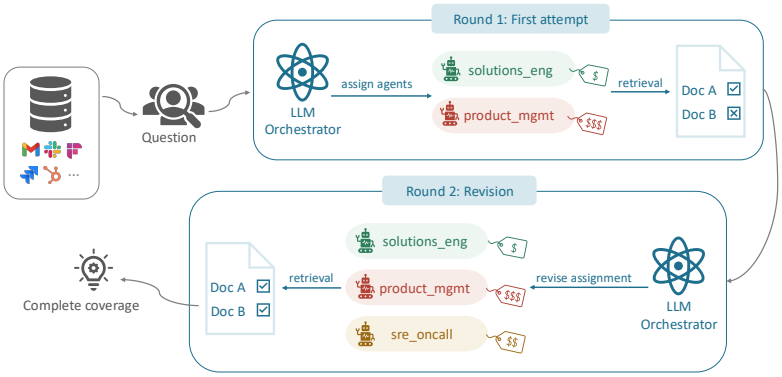

Figure 1: **Overview of access-scoped agent orchestration.** An LLM orchestrator selects agents with heterogeneous access scopes and engagement costs, receives retrieval outcomes, and can revise its assignment in a second round.

coverage, but engage substantially more of the team than necessary. Revealing the required sources largely fixes the first failure mode. Several narrow routers reach almost complete coverage at or near the cheapest covering cost, while broad routers continue to over-engage even when the required sources are known.

We further find that poor first-attempt routing is not simply a failure to respond to cost. Several selective routers change their assignments when relative prices change once source uncertainty is removed. The difficulty appears when source inference and cost control must be handled at the same time. Simple restraint prompts do not solve the problem.

We observe a similar pattern in second-round revision. Retrieval feedback makes models much better at correcting failed routes, and cost feedback can make routing more restrained. However, cost feedback does not reliably reduce cost without sacrificing task success. It also does not consistently guide models toward the cheapest additions needed to complete the task. Overall, we reveal a novel "cost-or-coverage" problem, an area for improvement for future LLM orchestrators.

2 Related Works

2.1 Adaptive and Resource-Aware Retrieval

Adaptive RAG treats retrieval as a resource-allocation problem, deciding whether and how much to retrieve for each query (Jeong et al., 2024; Tang et al., 2025; Marina et al., 2025). Agentic and search-augmented systems extend this to multi-step search by controlling when to retrieve, how deeply to search, and when to stop (Sun et al., 2026a; Hashemi et al., 2026; Mishra & Naik, 2026). Other work explicitly models evidence acquisition under externally specified access costs (Wu et al., 2026). These approaches typically represent retrieval cost through aggregate quantities such as steps, depth, tokens, latency, or scalar budgets. We instead study two structural, outcome-independent proxies for agent engagement cost: the breadth of the searchable document space and the breadth of the query-induced candidate set.

The first proxy concerns *access breadth*. Federated and multi-source retrieval systems reduce unnecessary search by selecting relevant resources, product domains, or repositories rather than querying all available collections (Wang et al., 2025a; Shojaee et al., 2025; Guerraoui et al., 2025). Unlike access-friction costs associated with restricted or purchasable evidence (Wu et al., 2026), our setting assumes documents are already accessible and uses the size of an agent's searchable collection as a coarse proxy for retrieval burden. This yields our *access-scope cost*.

The second proxy concerns the *query-induced candidate set*. Pre-retrieval query performance prediction relates query and collection characteristics, including query specificity, to retrieval difficulty before outcomes are observed (He & Ounis, 2006; Arabzadeh et al., 2020). Modern adaptive RAG similarly conditions retrieval on query-specific signals (Marina et al., 2025; Kim et al., 2026), while analyses of agentic RAG connect retrieval quality to both answer quality and search trajectory length

(Tian et al., 2025). Motivated by this view, we use the number of request-compatible documents within an agent's corpus as a second proxy for retrieval burden. In our customer-centric setting, this is the number of documents associated with the target customer, yielding our *candidate-breadth cost*.

2.2 Agent Orchestration

Agent orchestration has long involved allocating work and resources across distributed workers, with the Contract Net Protocol providing an early formulation of task allocation through decentralized negotiation (Smith, 1980). Recent LLM-based systems broaden this problem to the design and control of agent workflows. GPTSwarm represents multi-agent systems as optimizable computational graphs (Zhuge et al., 2024), AFlow searches over agentic workflow structures (Zhang et al., 2025b), G-Designer learns task-adaptive communication topologies (Zhang et al., 2025a), and AgentDropout dynamically removes unnecessary agents and interactions (Wang et al., 2025b). Other work adapts orchestration during execution (Dang et al., 2025) or learns orchestration decisions from execution traces (Zhang, 2026).

Recent work also directly evaluates whether LLM orchestrators can allocate tasks to heterogeneous agents under cost and capability constraints, including comparison against optimal algorithmic assignments (Amayuelas et al., 2025). A growing line of work makes resource constraints explicit in orchestration. Difficulty-aware orchestration jointly adjusts workflow depth, agentic operators, and heterogeneous model assignment across queries (Su et al., 2026), while Uno-Orchestra targets parsimonious agent use through selective delegation (Cui et al., 2026). Closely related cost-aware RAG work lets an LLM decide when to retrieve, which access-cost tier to query, and when to stop under an explicit evidence budget (Wu et al., 2026). Our setting isolates a complementary problem: the agent team and workflow structure are fixed, but agents have overlapping access to different information sources, and the sources required by a query are initially unknown. The orchestrator must therefore select a low-cost subset whose combined access covers the required evidence. Rather than optimizing the team topology or training an external orchestration policy, we evaluate whether an LLM orchestrator itself can jointly infer source requirements and control engagement cost.

3 The CoC-Bench Benchmark

We study retrieval orchestration over a fixed team of agents with heterogeneous access scopes. We first define the corpus, questions, agent team, and engagement-cost schemes used throughout the experiments.

3.1 The corpus and the questions

We use the synthetic enterprise knowledge corpus EnterpriseRAG-Bench (Sun et al., 2026b), containing 498,721 documents. Seven source types are partitioned into ten disjoint **source units**: `slack_co`, `slack_cust`, `slack_engops`, `slack_engx`, `gmail`, `google_drive`, `jira`, `linear`, `hubspot`, and `fireflies`. Each document belongs to exactly one unit, and unit sizes vary substantially.

Following the benchmark generation procedure, we construct 178 questions spanning 61 customer accounts and all 31 available source-unit pairs. Each question has exactly two gold documents from distinct source units, with no complete non-gold substitute among documents, Appendix B provides the full audit. Thus, the orchestration task is to engage agents whose access scopes cover the required evidence. All questions belong to the benchmark's *Project Related* category; our evaluation therefore isolates source routing rather than multi-hop reasoning or answer synthesis. Additional generation details are given in Appendix A.

3.2 The agent team and engagement costs

A fixed team of eight agents is granted overlapping subsets of the ten source units (Figure 2a). Each source unit is accessible to 2–6 agents, so most questions admit multiple covering assignments with different engagement costs.

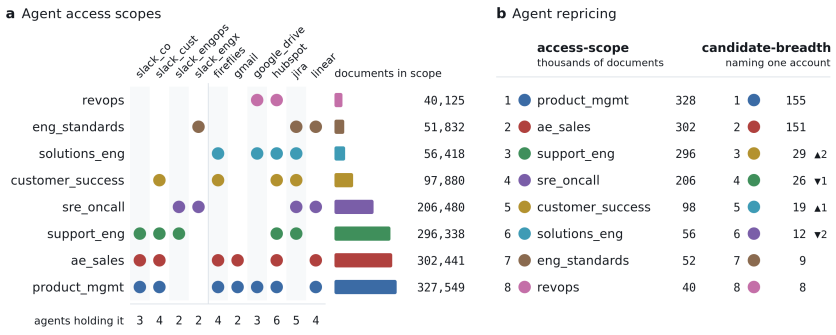

Figure 2: **Agent access scopes and pricing. (a)** Rows show the eight agents and columns the ten source units; filled cells indicate access, and bars show the number of documents in each agent's scope. **(b)** Agent cost rankings under access-scope and candidate-breadth pricing.

Engaging an agent incurs a one-time cost per episode, and the episode bill is the sum of the costs of the distinct agents engaged. We evaluate two pricing schemes on the same fixed team (Figure 2b). *Access-scope cost* is proportional to the total number of documents accessible to an agent. *candidate-breadth cost* is proportional to the mean number of documents in that scope matching a keyword (customer account name), computed over the same fixed set of 40 accounts for every agent. The former measures the total searchable space, whereas the latter measures the customer-specific candidate space.

Both schemes are outcome-independent structural proxies rather than prices fitted to model behavior. A separate retrieval experiment shows that broader scopes increase retrieval burden and that customer-specific candidate sets grow with corpus breadth; Appendix D reports the full results.

3.3 Task setup

Each question proceeds in two turns. The orchestrator receives the source descriptions, the fixed agent roster and access scopes, and the question, then selects agents and assigns each a short retrieval brief; the full prompt is given in Appendix C. We use oracle retrieval to isolate routing from retriever quality: an engaged agent receives a gold document whenever its scope contains the corresponding required source unit, and otherwise receives a missing-information signal. A route *covers* the question when the engaged agents collectively access both gold documents, so coverage failures reflect routing rather than retrieval error.

We evaluate three first-attempt information conditions. In the primary `agent_pricing` condition, agent-level engagement costs are given explicitly. In `unit_pricing`, only source-unit costs are provided and agent costs must be aggregated, testing the effect of price representation. In `gold_route`, agent costs and the required source units are both given, removing source inference. We additionally test a single-agent instruction that encourages selecting one agent when its scope appears sufficient and alternative team designs that reassign source units across agents; these controls are analyzed in §4.4 and §4.5.

3.4 Models and sampling

We evaluate eight model configurations spanning gpt-5.6-sol (OpenAI, 2026), Qwen3/Qwen3.5 (Yang et al., 2025; Qwen Team, 2026), and Llama families (Grattafiori et al., 2024; Meta AI, 2025); full model and sampling settings are given in Appendix E. The first-attempt grid crosses 8 models, 2 pricing schemes, the single-agent instruction on/off, 3 information conditions, and 3 draws over 178 questions, yielding 51,264 routing decisions. Results are averaged over draws. The revision experiment uses a subset of these configurations (§5).

	agent_pricing					gold_route				
Model	COVER↑	IDENT↑	#AGENTS	COST-R↓	EFF-OPT↓	COVER↑	IDENT↑	#AGENT	COST-R↓	EFF-OPT↓
Access-scope cost										
gpt-5.6-sol	28%	16%	1.24	1.36	5.93	100%	99%	1.48	1.00	1.01
Qwen3.5-35B	31%	17%	1.69	1.83	6.10	100%	96%	1.53	1.00	1.00
Qwen3.5-122B	38%	20%	1.86	2.35	5.69	100%	100%	1.48	1.00	1.00
Qwen3-32B (think)	36%	22%	2.04	2.46	6.02	99%	76%	1.32	1.29	1.34
Qwen3-8B (think)	47%	24%	2.16	3.37	6.00	98%	62%	1.24	1.55	1.59
Qwen3-32B (no think)	68%	39%	3.39	4.19	5.08	92%	63%	2.25	2.83	3.04
Llama-4-Scout	90%	60%	5.53	7.28	6.98	99%	89%	4.39	5.82	5.83
Llama-3.1-8B	78%	59%	5.16	7.23	7.29	84%	73%	3.88	4.66	4.82
Candidate-breadth cost										
gpt-5.6-sol	33%	19%	1.34	1.53	8.71	99%	99%	1.50	1.01	1.02
Qwen3.5-35B	33%	16%	1.75	1.79	8.35	100%	100%	1.52	1.00	1.00
Qwen3.5-122B	37%	18%	1.89	2.10	7.61	100%	100%	1.51	1.01	1.01
Qwen3-32B (think)	33%	21%	2.13	2.44	9.12	99%	74%	1.35	1.73	1.83
Qwen3-8B (think)	44%	25%	2.37	2.76	7.82	98%	58%	1.28	2.49	2.57
Qwen3-32B (no think)	63%	30%	3.38	4.03	6.40	93%	62%	2.35	3.36	3.90
Llama-4-Scout	90%	58%	5.64	12.00	12.07	99%	85%	3.98	7.97	8.05
Llama-3.1-8B	86%	79%	6.22	11.57	10.91	86%	82%	4.07	5.39	5.87

Table 1: First-attempt routing under `agent_pricing` and `gold_route` conditions, shown separately for the two pricing schemes.

3.5 METRICS

For a selected agent set S, let $c(S)$ denote its engagement cost. For each question and pricing scheme, c^* is the cost of the cheapest agent set covering both required source units, and c_{full} is the cost of engaging the full team. We report:

- COVER: percentage of questions whose engaged agents cover both required source units.
- IDENT: percentage of questions for which S contains the cheapest covering set, allowing extra agents.
- #AGENTS: number of distinct agents engaged.
- COST-R: engagement cost relative to the cheapest covering cost, $c(S)/c^*$, conditioned on coverage.
- EFF-OPT: effective cost relative to c^*, $\tilde{c}(S)/c^*$, where $\tilde{c}(S) = c(S)$ if S covers the question and $\tilde{c}(S) = \max\{c(S), c_{\text{full}}\}$ otherwise.

IDENT separates identification of a cheapest route from restraint after that route has been found, while EFF-OPT prevents uncovered questions from appearing artificially cheap.

Uncertainty. Matched settings are paired by question and draw; independently sampled settings, such as the single-agent instruction and its baseline, are paired only by question. We bootstrap over questions while preserving conditions and draws. Draw-level variability is $\mathrm{SD}(\Delta_1, \Delta_2, \Delta_3)$ for matched draws and $\sqrt{s_A^2 + s_B^2}$ otherwise. We emphasize an effect only when its magnitude exceeds twice both the bootstrap standard error and draw-level standard deviation. Metrics use parsable routing calls; 0.62% are excluded, with failure modes in Appendix F.

4 FIRST-ATTEMPT ROUTING

We first examine the orchestrator's initial routing decision. Our primary condition is `agent_pricing`, where agent-level engagement costs are given explicitly. We use `unit_pricing` and `gold_route` as diagnostic conditions. Unless otherwise stated, results use the single-agent instruction off and are averaged over three draws.

4.1 FIRST-ATTEMPT ROUTES SPAN A COST–COVERAGE SPECTRUM

First-attempt routes fall along a clear cost–coverage spectrum. Under access-scope pricing, the more selective configurations cover only 28–38% of questions while spending 1.36–2.35× the cheapest covering cost on the episodes they do cover. At the other end, the highest-coverage configurations reach 78–90% coverage but spend about 7× the floor on successful routes, with the remaining

models lying between these regimes (Table 1). Because COST-R is conditioned on coverage, sparse routes can appear inexpensive precisely because their misses are omitted.

The same pattern holds under candidate-breadth pricing. Selective configurations cover only 33–44% of questions, whereas the highest-coverage configurations reach 86–90% coverage at more than $11\times$ the cheapest covering cost. Once misses are penalized through EFF-OPT, every model remains well above the optimum of 1.0 under both pricing schemes. Thus, no evaluated configuration combines high first-attempt source coverage with low engagement cost.

This spectrum reflects two distinct routing styles under source uncertainty. Some models *commit* to a small set of agents and frequently miss the required sources, while others *hedge* by engaging much more of the team. In this lineup, gpt-5.6-sol and the Qwen models in thinking mode cluster toward the narrow-commitment end, whereas Qwen3-32B without thinking and the two Llama models route more broadly. We treat this grouping descriptively rather than as a causal effect of reasoning mode.

4.2 GOLD ROUTES SEPARATE SOURCE INFERENCE FROM OVER-ENGAGEMENT

Revealing the required source units sharply separates the two routing styles (Figure 3). For the narrow routers, most first-attempt failures disappear once source inference is removed. Under access-scope pricing, gpt-5.6-sol and the two Qwen3.5 models rise from 28–38% coverage to 100% while reaching the cost floor. The thinking Qwen3 models similarly reach 98–99% coverage while engaging fewer agents and reducing COST-R. Their low first-attempt coverage therefore largely reflects uncertainty about where the required evidence is located.

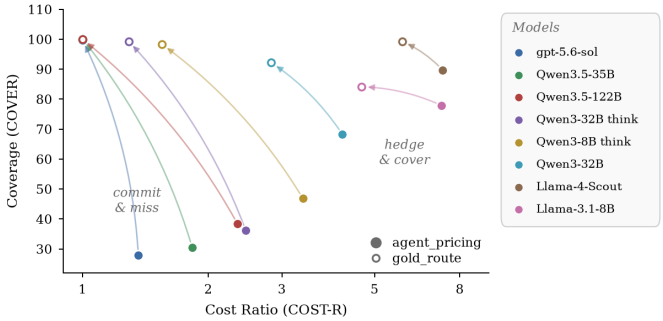

Figure 3: **First-attempt routing with and without source uncertainty.** Results under access-scope pricing. Filled markers show `agent_pricing`, where the orchestrator must infer the required source units; open markers show `gold_route`, where those units are given. Arrows connect the two conditions for each model.

Broad routers behave differently. Llama-4-Scout and Llama-3.1-8B remain substantially above the cheapest covering cost even after the required sources are given: under access-scope pricing they engage 4.39 and 3.88 agents on average, at $5.82\times$ and $4.66\times$ the floor, respectively. The same pattern holds under candidate-breadth pricing (Table 1). Their IDENT scores are nevertheless high, reaching 85–89% for Scout and 73–82% for Llama-3.1-8B across the two schemes. These models therefore often identify a cheapest covering set but continue engaging agents beyond it.

The two ends of the spectrum thus fail differently. Narrow routers *commit and miss*: once the required sources are known, they become both highly covering and inexpensive. Broad routers *identify and hedge*: even after source uncertainty is removed, they continue paying for substantial redundant engagement.

4.3 PRICE RESPONSES DIFFER ACROSS ROUTING STYLES

Price information affects different parts of the routing spectrum in different ways. In the first attempt, moving from `unit_pricing` to explicit `agent_pricing` changes little at either extreme: the sparsest models remain sparse, while the broadest models continue to engage much of the team (Figure 4a–b). The clearest shift occurs among the three Qwen3 configurations in the middle. When agent-level costs are stated directly, these models narrow their assignments and lose substantial coverage under both pricing schemes. Much of the lost coverage appears as UNDER failures, where

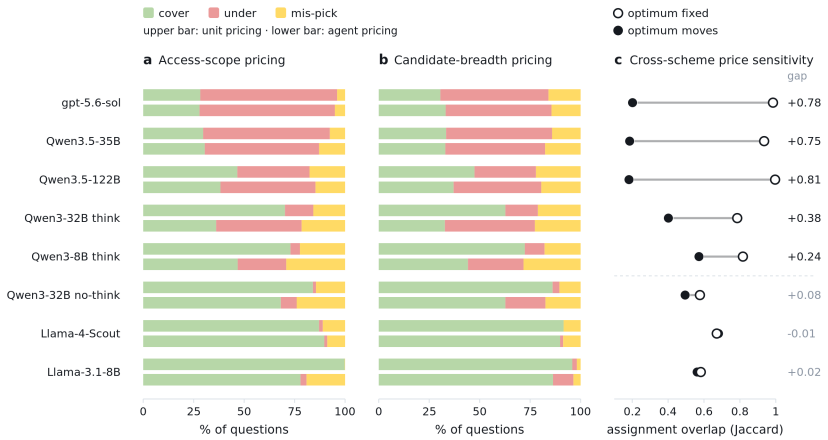

Figure 4: **Price sensitivity does not imply cost-effective routing. (a–b)** First-attempt coverage performance under `unit_pricing` and `agent_pricing`. **cover** denotes assignments that cover both required source units; **under** denotes misses whose engagement cost is below the cheapest possible covering assignment; and **mis-pick** denotes the remaining misses, which spend at least that amount but select the wrong agents. **(c)** Cross-scheme assignment overlap under `gold_route`. Open markers indicate source pairs whose cheapest covering assignment is unchanged by repricing, while filled markers indicate pairs whose optimum changes. The values at right report the difference in overlap between the two groups.

the selected assignment costs less than the cheapest possible covering route. Explicit agent prices therefore induce stronger restraint in these models, but often by shrinking the assignment below what coverage requires.

This does not mean that the sparse models ignore prices. To test sensitivity to relative prices directly, we remove source uncertainty with `gold_route` and compare assignments across the two pricing schemes (Figure 4c). For source pairs whose cheapest covering assignment is unchanged by repricing, a model that tracks the optimum should preserve its assignment, giving cross-scheme Jaccard overlap near 1.0. For pairs whose optimum changes, the two cheapest covering assignments themselves have mean Jaccard overlap of approximately 0.18; a model that follows the repriced optimum should therefore move toward this lower overlap.

The low-cost end of the spectrum follows this pattern most strongly. For gpt-5.6-sol and the two Qwen3.5 models, overlap remains near one when the optimum is fixed but falls sharply when it moves; Qwen3-32B and Qwen3-8B in thinking mode show the same separation more weakly. In contrast, Qwen3-32B without thinking and the two Llama models show little or no separation between the two groups. The first-attempt cost–coverage spectrum therefore does not arise from a uniform failure to read prices: several sparse routers are highly sensitive to repricing once the required sources are known, while broad routers are comparatively insensitive. What remains difficult is converting that sensitivity into a cheap covering assignment when source inference and cost control must be solved together.

4.4 A SIMPLE RESTRAINT PROMPT DOES NOT SOLVE THE COST–COVERAGE PROBLEM

Adding an instruction to use one agent when possible narrows routing, but does not reliably produce cheap, covering assignments. Broad routers shed some redundant agents, sometimes at the cost of coverage, while already restrained models often lose coverage with little cost benefit. Under `gold_route`, the instruction behaves more cleanly because source sufficiency is known, although fewer agents still need not imply lower engagement cost. A simple prompt-level restraint is therefore insufficient to resolve the cost–coverage problem. Full results are reported in Tables 7.

4.5 ROUTING STYLES REMAIN STABLE UNDER TEAM RE-GRANTING

We construct two control teams by reassigning source grants while keeping the questions, source units, agent count, and pricing scheme fixed. Qwen3.5-122B remains a narrow router and Llama-

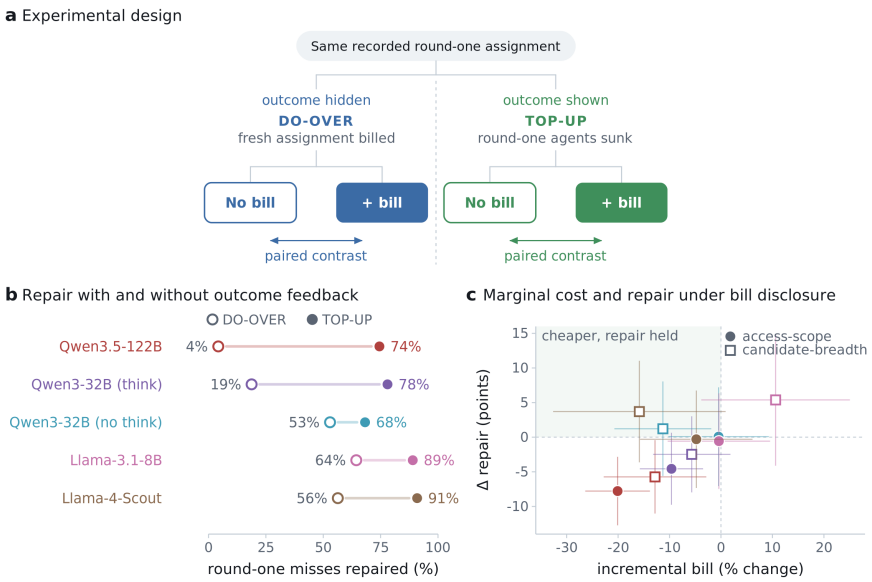

Figure 5: **Revision separates correction from cost control.** **(a)** Revision design and paired bill-disclosure contrasts. **(b)** Repair of round-one misses in no-bill blind do-over and outcome-shown top-up revisions under access-scope pricing. **(c)** Paired changes in top-up incremental bill and repair from bill disclosure; error bars show twice the bootstrap standard error on both axes, and the shaded region indicates lower cost with non-decreasing repair.

4-Scout remains a broad router across both controls, including under `gold_route`. Thus, the observed routing styles are not specific to the original source-to-agent assignment. Full results are reported in Tables 8.

5 Revision separates correction from cost control

A second routing decision asks whether models can coordinate cost control with coverage correction rather than respond to either signal in isolation. We rerun five representative models from the same recorded round-one assignments under two revision branches (Figure 5). In the *do-over* branch, the round-one retrieval outcome is withheld and the model writes a fresh assignment. In the *top-up* branch, the outcome is shown and previously engaged agents are treated as sunk, so only newly added agents incur additional cost. Within each branch, we compare a no-bill baseline with an otherwise matched revision that discloses the round-one bill. Bill effects are therefore paired within branch; comparisons between do-over and top-up are descriptive because outcome visibility and the billing frame change together.

5.1 Outcome feedback facilitates correction

A blind second attempt often fails to recover round-one misses, whereas outcome-shown top-ups repair them much more often. Under access-scope pricing without bill disclosure (Figure 5b), repair rises from 4% to 74% for Qwen3.5-122B and from 19% to 78% for Qwen3-32B with thinking. The same pattern holds for the broader routers, whose repair rates increase by 15–34 percentage points.

Because the two branches differ in both outcome visibility and billing frame, this contrast is descriptive rather than a causal estimate of outcome feedback alone. Still, once the revision setting reveals what the first assignment failed to retrieve, models can often redirect successfully. This is consistent with the source-inference bottleneck in Section 4.2: sparse routers frequently miss the required source initially but can recover once the missing evidence is made explicit.

5.2 COST FEEDBACK DOES NOT YIELD COST-EFFECTIVE CORRECTION

Cost feedback changes the complete revised assignment, but not uniformly. In the blind do-over branch, disclosing the round-one bill lowers the fresh second-turn assignment cost in all ten model–pricing cells, often together with lower repair or more broken coverage (Table 11). In the top-up branch, several models likewise produce smaller or cheaper complete second-turn assignments when the bill is shown, although the effect is weaker and not universal (Table 10). Thus, bill disclosure can induce thrift in the written revision, but this does not establish that the model chooses the right agents to add.

The top-up branch lets us test this marginal decision directly because round-one agents are already sunk. For a revised set S_2 following round-one set S_1, only the newly added agents, $S_2 \setminus S_1$, contribute additional cost. We therefore ask whether the model closes the remaining source gap with a low-cost addition. Figure 5c compares the change in incremental cost against the change in repair under bill disclosure. A few configurations fall nominally in the lower-cost, preserved-repair region, but their horizontal uncertainty intervals cross zero; under our uncertainty criterion, none shows a detectable reduction in incremental cost while preserving repair. Qwen3.5-122B is the only model with a detectable incremental-cost reduction under both pricing schemes, and its repair rate decreases at the same time.

Successful top-ups are also far from the cheapest available additions. Across models and pricing schemes, INCR-RATIO ranges from 2.58 to 7.68, while INCR-IDENT ranges from 38% to 79% (Tables 12–13). Cost feedback can therefore make the revised assignment more restrained, but it does not reliably help the orchestrator select the cheapest agents needed to repair the remaining source gap.

6 LIMITATIONS

One limitation of our study is the use of a single benchmark and retrieval setting. This was a deliberate choice to keep the orchestration problem controlled. However, we do not establish whether the same patterns hold across broader task domains or system settings. Extending the evaluation to other forms of agent collaboration is an important direction for future work. A second limitation is that our questions are automatically generated and initially verified using Qwen3-32B. Document-withholding tests and the corpus-wide uniqueness audit reduce the risk that the designated evidence is unnecessary or replaceable, but the questions may still retain model-specific phrasing or selection artifacts. A further limitation is that our engagement costs are structural proxies rather than full deployment costs. This lets us study cost independently of model outcomes, but real systems may also depend on latency, token usage, repeated calls, tool failures, and recovery cost. In such settings, engaging additional agents may improve reliability, and the expected total cost may depend on more than the initial assignment. Studying orchestration under such dynamic execution settings would provide a more complete picture of cost-effective agent selection.

7 CONCLUSION

We identify a novel problem, "cost-or-coverage" in the cost-aware orchestration over access-scoped agents, where an LLM must infer which information sources are needed and select a low-cost set of agents that covers them. We construct the `CoC-Bench`benchmark, facilitating a controlled evaluation with overlapping agent scopes, retrieval-grounded engagement costs, and oracle retrieval, where we separate source inference from cost control. Across eight model configurations, first-attempt routing falls into two common patterns: narrow routers often miss required sources, while broad routers achieve higher coverage through substantial over-engagement. Revealing the required sources largely fixes the former but not the latter. Revision experiments show a similar separation: outcome feedback helps models correct failed routes, while cost feedback does not reliably reduce cost without sacrificing task success. Overall, current LLM orchestrators can respond to source and cost information separately, but still struggle to balance them in a single decision. Improving this coordination is an important step toward more efficient multi-agent systems.

AI DISCLOSURE

In this work, we used Qwen3-32B to generate synthetic questions for the evaluation dataset, as described in §3.1 and Appendix A. Additionally, we used generative AI tools to improve readability of the manuscript. We have reviewed all AI-assisted work. We validated the LLM-generated dataset as described in Appendix B. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Alfonso Amayuelas, Jingbo Yang, Saaket Agashe, Ashwin Nagarajan, Antonis Antoniades, Xin Eric Wang, and William Wang. Self-resource allocation in multi-agent llm systems, 2025. URL https://arxiv.org/abs/2504.02051.

Negar Arabzadeh, Fattane Zarrinkalam, Jelena Jovanovic, Feras Al-Obeidat, and Ebrahim Bagheri. Neural embedding-based specificity metrics for pre-retrieval query performance prediction. *Information Processing Management*, 57(4):102248, 2020. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2020.102248. URL https://www.sciencedirect.com/science/article/pii/S030645731931266X.

Zhiqing Cui, Haotong Xie, Jiahao Yuan, Cheng Yang, Hanqing Wang, Yuxin Wu, Yifan Wu, Siru Zhong, Tao Yu, Yifu Guo, Siyu Zhang, Xinlei Yu, Qibing Ren, and Usman Naseem. Uno-orchestra: Parsimonious agent routing via selective delegation. *arXiv preprint arXiv:2605.05007*, 2026. URL https://arxiv.org/html/2605.05007v1.

Yufan Dang, Chen Qian, Xueheng Luo, Jingru Fan, Zihao Xie, Ruijie Shi, Weize Chen, Cheng Yang, Xiaoyin Che, Ye Tian, Xuantang Xiong, Lei Han, Zhiyuan Liu, and Maosong Sun. Multi-agent collaboration via evolving orchestration. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 165025–165059. Curran Associates, Inc., 2025. doi: 10.52202/085713-5502. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f1320d2e2842169c6fc89dcbd80e94d0-Paper-Conference.pdf.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan,

Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo

Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Rachid Guerraoui, Anne-Marie Kermarrec, Diana Petrescu, Rafael Pires, Mathis Randl, and Martijn de Vos. Efficient federated search for retrieval-augmented generation. In *Proceedings of the 5th Workshop on Machine Learning and Systems*, EuroMLSys '25, pp. 74–81, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715389. doi: 10.1145/3721146.3721942. URL https://doi.org/10.1145/3721146.3721942.

Helia Hashemi, Victor Rühle, and Saravan Rajmohan. Cost-aware retrieval-augmentation reasoning models with adaptive retrieval depth. In *Proceedings of the ACM Web Conference 2026*, WWW '26, pp. 2431–2440, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400723070. doi: 10.1145/3774904.3792693. URL https://doi.org/10.1145/3774904.3792693.

Ben He and Iadh Ounis. Query performance prediction. *Information Systems*, 31(7):585–594, 2006. ISSN 0306-4379. doi: https://doi.org/10.1016/j.is.2005.11.003. URL https://www.sciencedirect.com/science/article/pii/S0306437905000955. (1) SPIRE 2004 (2) Multimedia Databases.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.389. URL https://aclanthology.org/2024.naacl-long.389/.

Joeun Kim, Seunghyouk Yoon, Xuan-Bach Le, Youngeun Nam, Doyoung Kim, Hwanjun Song, and Jae-Gil Lee. QuDAR: Query-wise dual-perspective adaptive retrieval. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38662–38679, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1791. URL https://aclanthology.org/2026.acl-long.1791/.

Maria Marina, Nikolay Ivanov, Sergey Pletenev, Mikhail Salnikov, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Alexander Panchenko, and Viktor Moskvoretskii. LLM-independent adaptive RAG: Let the question speak for itself. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8697–8709, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.439. URL https://aclanthology.org/2025.emnlp-main.439/.

Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, April 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Sanjay Mishra and Ganesh R. Naik. Cost-aware query routing in rag: Empirical analysis of retrieval depth tradeoffs. *AI*, 7(7), 2026. ISSN 2673-2688. doi: 10.3390/ai7070250. URL https://www.mdpi.com/2673-2688/7/7/250.

OpenAI. GPT-5.6 System Card, July 2026. URL https://deploymentsafety.openai.com/gpt-5-6.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Parshin Shojaee, Sai Sree Harsha, Dan Luo, Akash Maharaj, Tong Yu, and Yunyao Li. Federated retrieval augmented generation for multi-product question answering. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, Steven Schockaert, Kareem Darwish, and Apoorv Agarwal (eds.), *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pp. 387–397, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-industry.33/.

Smith. The contract net protocol: High-level communication and control in a distributed problem solver. *IEEE Transactions on Computers*, C-29(12):1104–1113, 1980. doi: 10.1109/TC.1980.1675516.

Jinwei Su, Qizhen Lan, Yinghui Xia, Lifan Sun, Weiyou Tian, Tianyu Shi, and Lewei He. Difficulty-aware agentic orchestration for query-specific multi-agent workflows. In *Proceedings of the ACM Web Conference 2026*, WWW '26, pp. 2060–2070, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400723070. doi: 10.1145/3774904.3792240. URL https://doi.org/10.1145/3774904.3792240.

Jingbo Sun, Wenyue Chong, Songjun Tu, Qichao Zhang, Yaocheng Zhang, Jiajun Chai, Xiaohan Wang, Wei Lin, Guojun Yin, and Dongbin Zhao. AutoSearch: Adaptive search depth for efficient agentic RAG via reinforcement learning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 28059–28079, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1399. URL https://aclanthology.org/2026.findings-acl.1399/.

Yuhong Sun, Joachim Rahmfeld, Chris Weaver, Roshan Desai, Wenxi Huang, and Mark H. Butler. Enterpriserag-bench: A rag benchmark for company internal knowledge, 2026b. URL https://arxiv.org/abs/2605.05253.

Xiaqiang Tang, Qiang Gao, Jian Li, Nan Du, Qi Li, and Sihong Xie. MBA-RAG: a bandit approach for adaptive retrieval-augmented generation through question complexity. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3248–3254, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.218/.

Fangzheng Tian, Jinyuan Fang, Debasis Ganguly, Zaiqiao Meng, and Craig Macdonald. Am i on the right track? what can predicted query performance tell us about the search behaviour of agentic rag, 2025. URL https://arxiv.org/abs/2507.10411.

Shuai Wang, Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. Resllm: Large language models are strong resource selectors for federated search. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, pp. 1360–1364, New York, NY, USA, 2025a. Association for Computing Machinery. ISBN 9798400713316. doi: 10.1145/3701716.3715595. URL https://doi.org/10.1145/3701716.3715595.

Zhexuan Wang, Yutong Wang, Xuebo Liu, Liang Ding, Miao Zhang, Jie Liu, and Min Zhang. Agent-Dropout: Dynamic agent elimination for token-efficient and high-performance LLM-based multi-agent collaboration. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24013–24035, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1170. URL https://aclanthology.org/2025.acl-long.1170/.

Mingyan Wu, Han Yang, Omer Ben-Porat, and Yftah Ziser. When knowledge is not free: Cost-aware evidence selection in retrieval-augmented generation, 2026. URL https://arxiv.org/abs/2606.02245.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Chenchen Zhang. Reinforcement learning for llm-based multi-agent systems through orchestration traces, 2026. URL https://arxiv.org/abs/2605.02801.

Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, Tianlong Chen, and Dawei Cheng. G-designer: Architecting multi-agent communication topologies via graph neural networks. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76678–76692. PMLR, 13–19 Jul 2025a. URL https://proceedings.mlr.press/v267/zhang25cu.html.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, XiongHui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. Aflow: Automating agentic workflow generation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 34040–34077, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5492ecbce4439401798dcd2c90be94cd-Paper-Conference.pdf.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. GPTSwarm: Language agents as optimizable graphs. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62743–62767. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhuge24a.html.

A Question generation and selection

We follow the benchmark's question-generation procedure, adapted to construct two-source questions from customer-account documents. Each candidate is generated by Qwen3-32B from a bundle of three documents associated with the same customer account: one document from source unit X, one from source unit Y, and a third document from either X or Y used as a control. The prompt specifies the two required documents and asks the model to generate a question, atomic answer facts, and their document attributions.

We discard candidates that reduce to two independent lookups, use temporal wording that reveals a change in the underlying information, contain no answer facts, or fail to attribute at least one fact to each required document. We then verify document necessity by answering each question with the full bundle and with each document withheld in turn. This procedure yields 308 valid candidates, from which we select 178 using additional quality and balance constraints: questions must not explicitly name the required source platforms, each gold-document pair contributes at most one question, and selected questions must retain dependence on both required documents. The final set is balanced across customer accounts and routing conditions 2 as following:

We divide source-unit pairs into three routing classes. In class A, the cheapest covering assignment uses a single agent. In class B, one agent can cover both units, but a cheaper covering assignment uses multiple agents. In class C, no single agent covers both units. We further group pairs into wide, mid, and narrow bands according to the spread between their cheapest and most expensive minimal covering assignments.

	wide	mid	narrow	total
A	43	16	33	92
B	33	19	—	52
C	—	17	17	34
	76	52	50	**178**

Table 2: Question distribution by routing class and cost-spread band.

B GOLD-DOCUMENT UNIQUENESS

Our evaluation operates at the source level: a route succeeds when the selected agents' scopes cover the two required source units. Partial overlap from another document matters only if it provides a complete substitute from a non-required unit; overlap within a required unit leaves both the routing target and the minimum covering cost unchanged.

We audit all documents naming the question's customer account, a candidate set that contains both gold documents for all 178 questions and averages 55 documents per question. Because the 60%-token rule used for short generated answers produces accidental matches in long documents, this audit uses a conservative rule that requires every checkable token of a fact to appear as a whole word. For each question, we identify the highest-scoring non-gold document. None reproduces the complete answer.

We manually inspect all 13 cases in which the strongest non-gold document reproduces at least 60% of the answer facts (Table 3). Five contain genuine partial overlap, but only two are in non-required units. Both omit a pivotal fact unique to the gold evidence and therefore cannot be combined with other non-gold documents into a complete answer. Thus, no question in the 178-question set has a complete non-gold substitute or an alternative source-level route.

account	unit	facts	matched	assessment
In-unit competitors				
Photon Health	LINEAR	7	6	partial
Helio Health	SLACK-CUST	5	4	partial
Orbit Analytics	GDRIVE	3	2	lexical only
Helio Health	GDRIVE	5	3	partial
Off-unit competitors				
Aurora Health	SLACK-CUST	5	4	partial
Lumen Legaltech	FIREFLIES	5	4	partial
Northbridge Bank	JIRA	4	3	lexical only
Northbridge	SLACK-CUST	6	4	lexical only
Northstar Healthcare	FIREFLIES	3	2	lexical only
Brightline Insurance	JIRA	8	5	lexical only
Northstar Fintech	FIREFLIES	5	3	lexical only
HarborBank	SLACK-CO	5	3	lexical only
Northbridge	HUBSPOT	5	3	lexical only

Table 3: **Manual review of the 13 highest-overlap non-gold documents.** These are all questions whose strongest non-gold competitor matches at least 60% of the answer facts under exact whole-word matching. *Matched* is the number of facts for which every checkable token is present. A *partial* document contains genuine but incomplete evidence; *lexical only* indicates shared terminology without the requested content. None provides a complete substitute for the gold evidence.

C PROMPTS

Figure 6 shows the prompts in first-attempt turn (round 1) of main condition `agent-pricing`, notes indicating differences of other conditions. Figure 7 shows the prompts in revision turn (round 2).

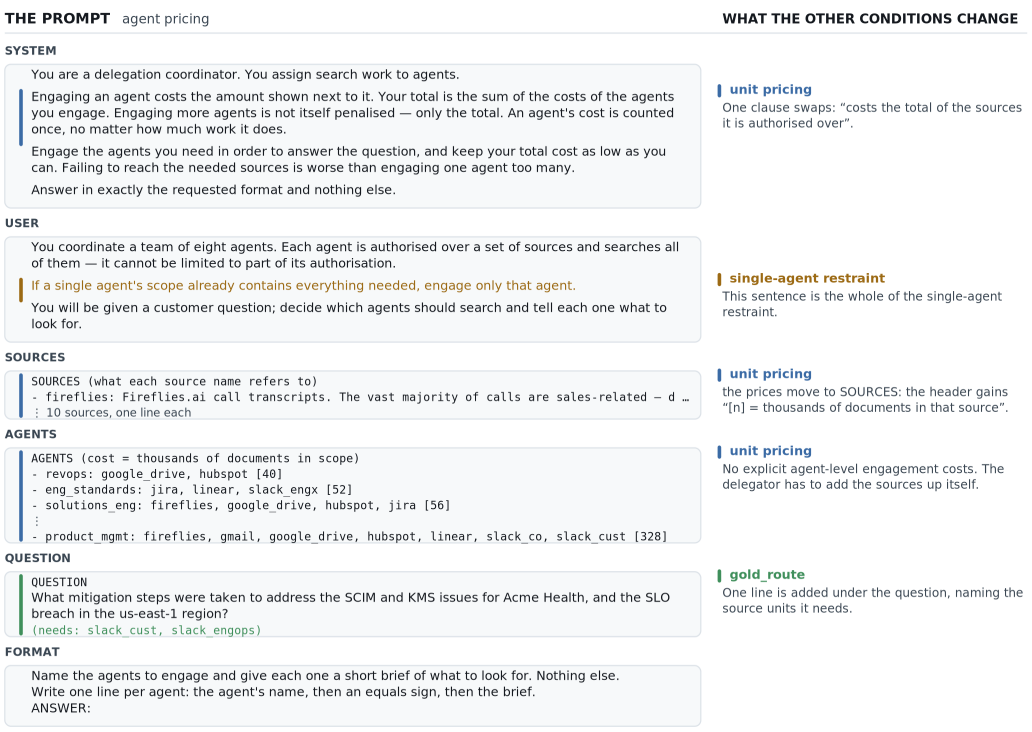

Figure 6: First attempt prompts.

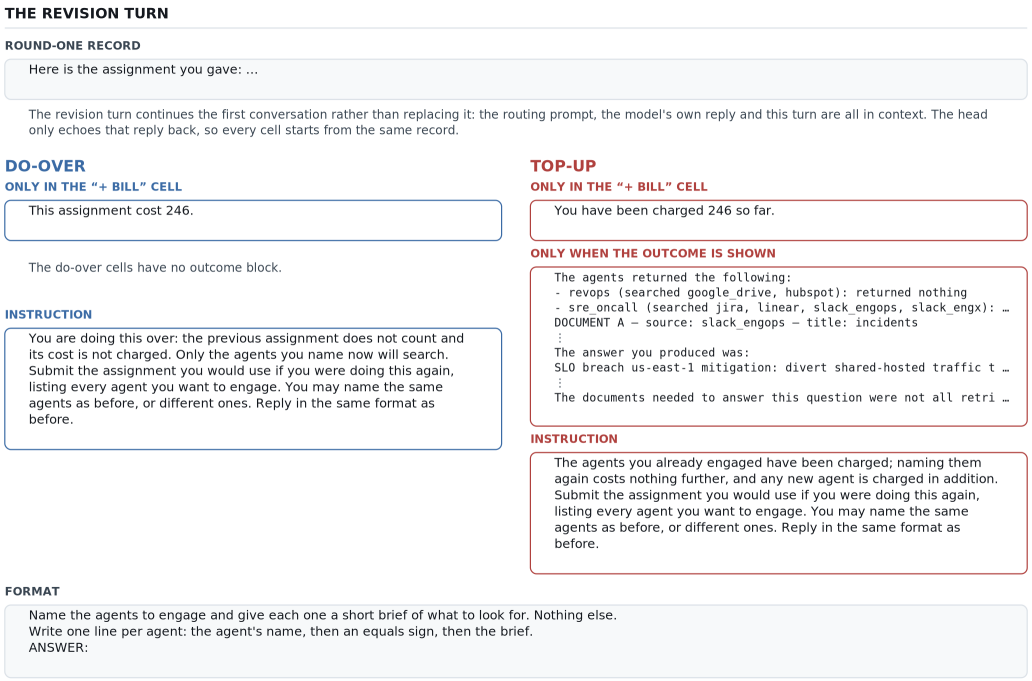

Figure 7: Revision round prompts.

D GROUNDING THE PRICE PROXIES

We ground the pricing proxies with a separate scoped-retrieval experiment that varies the size of the accessible search space while holding the retrieval model and target document fixed. A Qwen3-32B (think) agent searches nested unit-, agent-, and team-level scopes using a directory-restricted shell for up to ten turns. The experiment covers 1,272 episodes over 18 scopes containing 6,120 to 498,721 documents, under two clue types: a *verbatim* span known to occur in the target document and the benchmark gold *fact*.

Clue	Scope Level	RECALL@5	EXACTMATCH	TURNS	SHELLTIME(S)	PROMPTTOKENS
verbatim	unit	83%	65%	1.6	0.8	1,509
	agent	70%	44%	1.6	2.4	2,189
	team	57%	38%	1.9	4.8	3,315
fact	unit	56%	40%	1.8	1.2	1,790
	agent	48%	21%	1.9	2.7	2,668
	team	38%	19%	2.2	5.7	4,051

Table 4: **Retrieval becomes less accurate and more costly as the accessible scope expands.** RECALL@5 measures whether the target document appears among up to five returned paths, while EXACTMATCH requires the target to be the only returned path. TURNS counts shell commands, SHELLTIME reports their total wall-clock time, and PROMPTTOKENS reports cumulative model-input tokens over the episode.

Across both clue types, larger scopes are consistently associated with lower retrieval accuracy and higher retrieval cost. As the accessible search space expands from unit to agent to team scope, both target recovery and exact resolution decline, while shell time and prompt-token usage increase substantially. Together, these results support both pricing schemes as structural proxies for retrieval burden: broader access increases the overall search space, while candidate breadth captures the customer-specific portion of that space that remains relevant to the query.

E MODELS AND SAMPLING

Model	Weights	Sampling
gpt-5.6-sol	API	reasoning=high
Qwen3.5-122B-A10B	GPTQ Int4	$T = 0.6$
Qwen3.5-35B-A3B	GPTQ Int4	$T = 0.6$
Qwen3-32B (think)	AWQ	$T = 0.6$
Qwen3-8B (think)	bf16	$T = 0.6$
Qwen3-32B (no think)	AWQ	$T = 0.7$
Llama-4-Scout-17B-16E	bf16	$T = 0.6$
Llama-3.1-8B	bf16	$T = 0.6$

Table 5: **Model and sampling settings.** Open-weight models are served locally through vLLM; gpt-5.6-sol is accessed through the API.

Each condition is sampled independently three times. Seeds are treated as draw labels rather than deterministic replay identifiers. Generation limits were set above the length required for the routing outputs and were not used as an experimental variable.

F UNPARSED ROUTING CALLS

Of the 51,264 first-attempt routing calls, 318 (0.62%) produced no parsable agent assignment and are excluded from all routing metrics. Of these, 219 returned no assignment before generation ended, while 99 returned nonempty text that could not be resolved to agent names. Empty outputs were concentrated in Qwen3.5-35B-A3B, whereas malformed assignments occurred primarily in Qwen3-32B without thinking and Llama-4-Scout. We exclude these calls rather than score them as routing failures because no valid routing decision was produced.

G ADDITIONAL FIRST-ATTEMPT RESULTS

G.1 FULL FIRST-ATTEMPT ASSIGNMENT

model	condition	COVER	IDENT	AGENTS	COST-RATIO	EFF-OPT
Access-scope cost						
gpt-5.6-sol	`unit_pricing`	28%	18%	1.25	1.33	5.92
	`agent_pricing`	28%	16%	1.24	1.36	5.93
	`gold_route`	100%	99%	1.48	1.00	1.01
Qwen3.5-35B	`unit_pricing`	30%	18%	1.62	1.75	5.89
	`agent_pricing`	31%	17%	1.69	1.83	6.10
	`gold_route`	100%	96%	1.53	1.00	1.00
Qwen3.5-122B	`unit_pricing`	47%	20%	1.73	2.70	5.57
	`agent_pricing`	38%	20%	1.86	2.35	5.69
	`gold_route`	100%	100%	1.48	1.00	1.00
Qwen3-32B (think)	`unit_pricing`	70%	30%	2.61	4.27	5.20
	`agent_pricing`	36%	22%	2.04	2.46	6.02
	`gold_route`	99%	76%	1.32	1.29	1.34
Qwen3-8B (think)	`unit_pricing`	73%	29%	2.85	4.95	6.20
	`agent_pricing`	47%	24%	2.16	3.37	6.00
	`gold_route`	98%	62%	1.24	1.55	1.59
Qwen3-32B (no think)	`unit_pricing`	84%	39%	3.66	5.38	5.45
	`agent_pricing`	68%	39%	3.39	4.19	5.08
	`gold_route`	92%	63%	2.25	2.83	3.04
Llama-4-Scout	`unit_pricing`	87%	54%	5.28	7.10	6.83
	`agent_pricing`	90%	60%	5.53	7.28	6.98
	`gold_route`	99%	89%	4.39	5.82	5.83
Llama-3.1-8B	`unit_pricing`	100%	78%	6.99	8.74	8.73
	`agent_pricing`	78%	59%	5.16	7.23	7.29
	`gold_route`	84%	73%	3.88	4.66	4.82
Candidate-breadth cost						
gpt-5.6-sol	`unit_pricing`	31%	17%	1.35	1.49	9.02
	`agent_pricing`	33%	19%	1.34	1.53	8.71
	`gold_route`	99%	99%	1.50	1.01	1.02
Qwen3.5-35B	`unit_pricing`	33%	17%	1.71	1.81	8.14
	`agent_pricing`	33%	16%	1.75	1.79	8.35
	`gold_route`	100%	100%	1.52	1.00	1.00
Qwen3.5-122B	`unit_pricing`	47%	20%	1.93	3.45	8.28
	`agent_pricing`	37%	18%	1.89	2.10	7.61
	`gold_route`	100%	100%	1.51	1.01	1.01
Qwen3-32B (think)	`unit_pricing`	63%	28%	2.70	5.78	8.47
	`agent_pricing`	33%	21%	2.13	2.44	9.12
	`gold_route`	99%	74%	1.35	1.73	1.83
Qwen3-8B (think)	`unit_pricing`	72%	31%	2.96	6.74	8.43
	`agent_pricing`	44%	25%	2.37	2.76	7.82
	`gold_route`	98%	58%	1.28	2.49	2.57
Qwen3-32B (no think)	`unit_pricing`	86%	38%	3.90	8.32	8.72
	`agent_pricing`	63%	30%	3.38	4.03	6.40
	`gold_route`	93%	62%	2.35	3.36	3.90
Llama-4-Scout	`unit_pricing`	92%	51%	5.49	13.04	13.08
	`agent_pricing`	90%	58%	5.64	12.00	12.07
	`gold_route`	99%	85%	3.98	7.97	8.05
Llama-3.1-8B	`unit_pricing`	96%	80%	6.64	12.85	12.81
	`agent_pricing`	86%	79%	6.22	11.57	10.91
	`gold_route`	86%	82%	4.07	5.39	5.87

Table 6: Full first-attempt results.

G.2 SINGLE-AGENT INSTRUCTION

model	condition	COVER	IDENT	AGENTS	COST-RATIO	EFF-OPT
Access-scope cost						
gpt-5.6-sol	`unit_pricing`	27%	16%	1.10	1.33	6.49
	`agent_pricing`	36%	14%	1.07	1.70	5.89
	`gold_route`	100%	91%	1.39	1.15	1.15
Qwen3.5-35B	`unit_pricing`	26%	18%	1.53	1.77	5.98
	`agent_pricing`	30%	18%	1.55	1.75	6.11
	`gold_route`	100%	88%	1.41	1.12	1.12
Qwen3.5-122B	`unit_pricing`	49%	18%	1.50	2.53	5.52
	`agent_pricing`	30%	18%	1.44	1.85	6.22
	`gold_route`	100%	98%	1.46	1.03	1.03
Qwen3-32B (think)	`unit_pricing`	59%	24%	1.87	3.22	5.40
	`agent_pricing`	34%	20%	1.58	2.05	6.26
	`gold_route`	100%	66%	1.20	1.41	1.41
Qwen3-8B (think)	`unit_pricing`	66%	25%	2.38	4.23	6.14
	`agent_pricing`	39%	22%	1.88	2.76	6.40
	`gold_route`	99%	64%	1.20	1.49	1.52
Qwen3-32B (no think)	`unit_pricing`	81%	37%	3.40	4.93	5.11
	`agent_pricing`	63%	37%	3.12	4.02	5.01
	`gold_route`	95%	62%	2.03	2.50	2.69
Llama-4-Scout	`unit_pricing`	73%	43%	4.23	6.35	6.76
	`agent_pricing`	79%	46%	4.42	6.31	6.41
	`gold_route`	99%	85%	4.04	5.09	5.08
Llama-3.1-8B	`unit_pricing`	96%	68%	6.25	8.20	8.19
	`agent_pricing`	78%	59%	5.18	7.50	7.52
	`gold_route`	83%	72%	3.81	4.78	4.92
Candidate-breadth cost						
gpt-5.6-sol	`unit_pricing`	29%	14%	1.06	1.81	9.75
	`agent_pricing`	35%	15%	1.09	2.48	9.38
	`gold_route`	100%	82%	1.34	1.67	1.67
Qwen3.5-35B	`unit_pricing`	30%	17%	1.66	1.83	8.60
	`agent_pricing`	26%	15%	1.59	1.57	9.38
	`gold_route`	100%	92%	1.44	1.21	1.21
Qwen3.5-122B	`unit_pricing`	47%	17%	1.71	3.25	8.48
	`agent_pricing`	25%	15%	1.51	1.67	9.74
	`gold_route`	100%	96%	1.47	1.05	1.05
Qwen3-32B (think)	`unit_pricing`	48%	21%	2.04	4.64	9.17
	`agent_pricing`	28%	16%	1.66	2.02	9.73
	`gold_route`	100%	68%	1.22	2.14	2.16
Qwen3-8B (think)	`unit_pricing`	64%	26%	2.52	5.52	8.22
	`agent_pricing`	39%	22%	2.00	2.46	8.59
	`gold_route`	98%	61%	1.21	2.22	2.35
Qwen3-32B (no think)	`unit_pricing`	80%	34%	3.63	7.54	8.25
	`agent_pricing`	61%	31%	3.14	3.68	6.24
	`gold_route`	92%	63%	2.04	2.81	3.24
Llama-4-Scout	`unit_pricing`	82%	38%	4.38	11.02	11.63
	`agent_pricing`	78%	44%	4.36	9.58	10.09
	`gold_route`	99%	82%	3.50	6.33	6.36
Llama-3.1-8B	`unit_pricing`	87%	73%	5.95	12.05	11.66
	`agent_pricing`	83%	75%	5.94	10.49	10.13
	`gold_route`	83%	79%	3.82	5.67	6.38

Table 7: Full first-attempt results **with the single-agent instruction**. The corresponding hint-off results are reported in Table 6.

G.3 RE-GRANTED TEAM CONTROLS

			team A			team B		
condition	metric	original	Δ	2 SE	2 SD	Δ	2 SE	2 SD
Qwen3.5-122B								
`unit_pricing`	IDENT	0.197	−0.008	0.057	0.092	+0.024	0.055	0.107
	COVER	0.466	+0.041	0.067	0.286	−0.017	0.066	0.269
	AGENTS	1.730	+0.156	0.073	0.449	+0.160	0.070	0.363
	POSITION	0.595	−0.040	0.058	0.227	+0.006	0.058	0.221
`agent_pricing`	IDENT	0.200	+0.024	0.057	0.070	**+0.075**	0.061	0.064
	COVER	0.384	+0.077	0.062	0.118	+0.055	0.064	0.118
	AGENTS	1.859	+0.021	0.067	0.365	+0.028	0.059	0.300
	POSITION	0.652	−0.073	0.057	0.093	−0.061	0.058	0.097
`gold_route`	IDENT	0.998	+0.000	0.005	0.009	−0.002	0.007	0.015
	COVER	1.000	+0.000	0.000	0.000	−0.002	0.004	0.007
	AGENTS	1.483	+0.073	0.102	0.000	**+0.132**	0.099	0.013
	POSITION	0.000	+0.000	0.000	0.000	+0.002	0.004	0.007
Llama-4-Scout								
`unit_pricing`	IDENT	0.545	+0.006	0.079	0.043	+0.064	0.070	0.059
	COVER	0.871	+0.063	0.051	0.080	**+0.068**	0.052	0.064
	AGENTS	5.277	−0.044	0.221	0.451	+0.249	0.205	0.407
	POSITION	0.671	−0.023	0.038	0.048	+0.044	0.037	0.047
`agent_pricing`	IDENT	0.605	+0.015	0.079	0.056	**+0.135**	0.071	0.066
	COVER	0.897	−0.027	0.057	0.045	**+0.056**	0.045	0.018
	AGENTS	5.534	**−0.484**	0.218	0.195	**+0.401**	0.174	0.190
	POSITION	0.678	**−0.082**	0.046	0.051	**+0.067**	0.030	0.021
`gold_route`	IDENT	0.890	**−0.071**	0.062	0.056	**−0.079**	0.061	0.040
	COVER	0.993	−0.006	0.015	0.015	+0.006	0.009	0.015
	AGENTS	4.386	**−0.429**	0.359	0.262	+0.172	0.355	0.265
	POSITION	0.531	**−0.114**	0.058	0.027	−0.028	0.053	0.021

Table 8: **Re-granted-team control for two models under access-scope pricing.** The original team is compared with two independently re-granted teams. Because re-granting changes the cost floor, cross-team cost is reported using POSITION $=$ (cost $-$ FLOOR)/(ALL-EIGHT $-$ FLOOR). Each contrast is paired by question. 2 SE is the question-level bootstrap uncertainty and 2 SD is the draw-to-draw uncertainty of the contrast. Bold values exceed both reporting thresholds.

G.4 PRICE-SENSITIVITY VALUES

Model	Optimum Unchanged	Optimum Changed	Gap	Null Gap
Qwen3.5-122B	0.996	0.182	**+0.814**	−0.004
gpt-5.6-sol	0.985	0.202	**+0.783**	−0.005
Qwen3.5-35B	0.935	0.186	**+0.750**	−0.027
Qwen3-32B (think)	0.785	0.402	**+0.383**	+0.122
Qwen3-8B (think)	0.817	0.572	**+0.245**	+0.036
Qwen3-32B (no think)	0.577	0.495	+0.082	+0.115
Llama-3.1-8B	0.582	0.562	+0.021	+0.038
Llama-4-Scout	0.671	0.681	−0.010	−0.029

Table 9: **Cross-scheme assignment overlap under `gold_route`.** Questions are split by whether repricing leaves the cheapest covering assignment unchanged or changes it. The *gap* is the difference in overlap between these two groups. The *null gap* applies the same split within a single pricing scheme across draws, providing an empirical baseline for differences due to question composition and sampling. Bold gaps exceed the empirical noise threshold, defined as the larger of 0.10 and the absolute null gap.

H ADDITIONAL REVISION RESULTS

H.1 FULL SECOND-TURN ASSIGNMENTS

model	condition	COVER	IDENT	AGENTS	COST-RATIO	EFF-OPT
Access-scope cost						
Qwen3.5-122B	round one	38%	20%	1.86	2.35	5.69
	do-over, no bill	32%	19%	1.65	1.93	6.12
	do-over, + bill	31%	20%	1.63	1.79	5.99
	top-up, no bill	83%	42%	2.57	2.68	3.31
	top-up, + bill	78%	40%	2.40	2.35	3.15
Qwen3-32B (think)	round one	36%	22%	2.04	2.46	6.02
	do-over, no bill	43%	27%	2.44	2.87	5.56
	do-over, + bill	33%	23%	2.09	2.22	5.70
	top-up, no bill	85%	52%	3.43	4.05	4.30
	top-up, + bill	82%	47%	3.14	3.48	4.02
Qwen3-32B (no think)	round one	68%	39%	3.39	4.19	5.08
	do-over, no bill	82%	45%	3.75	4.71	5.02
	do-over, + bill	72%	41%	3.58	4.50	5.04
	top-up, no bill	90%	51%	3.54	4.07	4.46
	top-up, + bill	90%	51%	3.53	4.12	4.41
Llama-3.1-8B	round one	78%	59%	5.16	7.23	7.29
	do-over, no bill	89%	73%	5.88	7.11	6.93
	do-over, + bill	85%	70%	5.59	6.87	6.67
	top-up, no bill	97%	84%	6.57	8.14	8.07
	top-up, + bill	97%	84%	6.57	8.32	8.22
Llama-4-Scout	round one	90%	60%	5.53	7.28	6.98
	do-over, no bill	92%	65%	5.70	7.56	7.36
	do-over, + bill	89%	60%	5.17	6.95	6.74
	top-up, no bill	99%	77%	5.30	6.85	6.82
	top-up, + bill	99%	75%	5.19	6.73	6.69
Customer-breadth cost						
Qwen3.5-122B	round one	37%	18%	1.89	2.10	7.61
	do-over, no bill	32%	16%	1.77	1.78	8.21
	do-over, + bill	31%	15%	1.69	1.68	8.18
	top-up, no bill	79%	40%	2.59	2.71	3.94
	top-up, + bill	76%	40%	2.52	2.41	3.85
Qwen3-32B (think)	round one	33%	21%	2.13	2.44	9.12
	do-over, no bill	45%	27%	2.58	2.73	7.62
	do-over, + bill	33%	20%	2.17	2.05	8.44
	top-up, no bill	81%	48%	3.37	5.02	6.28
	top-up, + bill	80%	46%	3.10	4.79	6.01
Qwen3-32B (no think)	round one	63%	30%	3.38	4.03	6.40
	do-over, no bill	75%	37%	3.81	6.19	7.32
	do-over, + bill	72%	36%	3.62	5.56	6.70
	top-up, no bill	88%	47%	3.55	4.77	5.69
	top-up, + bill	88%	47%	3.54	4.52	5.33
Llama-3.1-8B	round one	86%	79%	6.22	11.57	10.91
	do-over, no bill	85%	78%	5.99	8.49	7.85
	do-over, + bill	81%	75%	5.76	7.45	6.78
	top-up, no bill	97%	94%	7.06	12.48	12.30
	top-up, + bill	98%	94%	7.14	12.97	12.81
Llama-4-Scout	round one	90%	58%	5.64	12.00	12.07
	do-over, no bill	95%	58%	5.59	12.51	12.35
	do-over, + bill	90%	49%	4.85	11.25	11.14
	top-up, no bill	98%	72%	5.21	10.97	11.13
	top-up, + bill	99%	72%	5.10	10.73	10.84

Table 10: **Full revision results.** Every cell is scored on the complete set of agents that model selected on the second turn. Round one is the same recorded assignment all four branch from.

H.2 Effect of Bill Disclosure on Revision Behavior

Table 11 compares matched revisions with and without the round-one bill. In blind do-overs, bill disclosure lowers the fresh second-turn assignment cost in all model–pricing cells, often together with lower repair or higher break. In outcome-shown top-ups, cost reductions are less consistent and do not improve repair. Bill disclosure can therefore encourage more restrained revision, but does not reliably improve the cost–coverage trade-off.

		REPAIR (%)			BREAK (%)			COST		
Model	Pricing	no bill	+ bill	Δ [2SE, 2SD]	no bill	+ bill	Δ [2SE, 2SD]	no bill	+ bill	Δ [2SE, 2SD]
(a) Do-over										
Qwen3.5-122B	A	4.2	3.0	−1.2 [2.8, 2.0]	23.4	23.6	+0.2 [6.9, 18.2]	169	161	−8 [8, 25]
	C	1.8	2.1	+0.3 [1.8, 3.0]	17.8	18.9	+1.1 [6.1, 20.5]	29	27	−2 [1, 4]
Qwen3-32B (think)	A	18.8	9.3	−9.5 [4.8, 16.0]	14.2	25.2	**+11.0** [6.0, 4.6]	282	189	**−93** [15, 27]
	C	22.9	11.3	−11.5 [4.5, 12.2]	9.5	23.0	+13.5 [6.8, 16.4]	55	35	**−20** [4, 2]
Qwen3-32B (no think)	A	52.8	31.5	**−21.4** [8.7, 20.7]	4.6	9.0	**+4.4** [2.9, 3.9]	604	542	**−62** [18, 26]
	C	42.9	34.2	−8.7 [7.1, 12.2]	6.2	6.2	0.0 [3.2, 4.7]	149	129	**−20** [7, 20]
Llama-3.1-8B	A	65.1	56.4	−8.7 [9.6, 14.9]	3.6	6.2	+2.6 [2.7, 4.0]	916	853	**−63** [28, 58]
	C	43.9	48.6	+4.8 [11.9, 8.8]	8.9	14.0	**+5.2** [2.9, 1.9]	196	166	**−30** [13, 22]
Llama-4-Scout	A	56.9	54.9	−2.0 [9.7, 16.5]	3.6	7.5	**+4.0** [2.0, 3.8]	989	884	**−105** [21, 28]
	C	57.0	48.0	−9.0 [12.0, 16.2]	0.8	4.8	**+4.0** [1.8, 0.8]	322	278	**−44** [8, 3]
(b) Top-up										
Qwen3.5-122B	A	74.4	67.3	**−7.1** [5.4, 5.2]	4.3	3.8	−0.5 [3.2, 4.3]	410	353	**−58** [17, 23]
	C	68.7	62.9	−5.8 [5.9, 12.0]	4.5	2.2	−2.3 [3.4, 5.6]	89	81	**−8** [6, 7]
Qwen3-32B (think)	A	78.0	73.1	−4.9 [5.1, 9.8]	2.6	3.1	+0.5 [2.8, 1.9]	555	497	**−58** [22, 35]
	C	72.7	71.7	−1.1 [5.9, 5.2]	1.1	3.4	+2.4 [2.8, 2.5]	147	135	−12 [9, 13]
Qwen3-32B (no think)	A	68.0	67.5	−0.5 [7.2, 5.3]	0.0	0.0	0.0 [0.0, 0.0]	593	598	+5 [12, 12]
	C	68.2	69.4	+1.2 [6.8, 9.2]	0.0	0.3	+0.3 [0.6, 1.1]	156	149	**−8** [6, 3]
Llama-3.1-8B	A	89.4	87.9	−1.5 [7.1, 13.6]	0.2	0.9	+0.7 [1.1, 1.4]	1108	1118	+11 [23, 17]
	C	82.0	87.4	+5.4 [9.5, 1.9]	0.2	0.2	0.0 [0.6, 1.3]	334	343	+9 [8, 11]
Llama-4-Scout	A	91.0	90.7	−0.3 [7.0, 10.9]	0.0	0.2	+0.2 [0.4, 0.7]	965	943	**−22** [14, 21]
	C	88.9	92.6	+3.7 [7.3, 12.8]	0.4	0.6	+0.2 [0.4, 0.7]	308	299	−9 [5, 10]

Table 11: **Effect of round-one bill disclosure on revision.** Panel (a) reports blind do-overs and Panel (b) outcome-shown top-ups. A and C denote access-scope and candidate-breadth pricing. Effects are $(+\text{bill}) - (\text{no bill})$ and levels are means over three draws. Brackets report $[2\text{SE}, , 2\text{SD}]$ for the paired effect; bold effects exceed both uncertainty measures.

H.3 TOP-UP BILLING AND INCREMENTAL ROUTING

For a round-one assignment S_1 and revised assignment S_2, the actual top-up bill is

$$\text{INCR-COST} = c(S_1 \cup S_2) - c(S_1) = c(S_2 \setminus S_1).$$

For round-one misses, let I^* be the minimum additional cost needed to cover the missing sources. INCR-IDENT indicates whether the added agents contain the minimum-cost gap-closing set, INCR-RATIO $=$ INCR-COST$/I^*$ on repaired episodes, and $\text{REPAIR}_\cup$ measures coverage by $S_1 \cup S_2$.

Bill disclosure detectably reduces incremental cost for Qwen3.5-122B under both pricing schemes, but also reduces repair; no configuration shows a detectable cost reduction while preserving repair. Successful top-ups remain 2.58–7.68 times more expensive than the minimum required addition. Changes in assignment size are descriptive: on round-one-covered questions, removing a previously engaged agent changes the stated assignment but cannot reduce the actual bill because its cost is already sunk.

		INCR-COST					INCR-IDENT (%)			$\text{REPAIR}_\cup$ (%)				
model	pricing	no bill	+ bill	Δ	2SE	2SD	no bill	+ bill	Δ	no bill	+ bill	Δ	2SE	2SD
Qwen3.5-122B	A	396	317	**−80**	25	26	48	43	−4.3	78	70	**−7.8**	4.9	7.7
	C	96	84	**−12**	10	11	43	38	−5.6	73	67	**−5.7**	5.3	5.2
Qwen3-32B (think)	A	492	445	−47	30	55	50	45	−5.1	79	74	−4.6	5.2	11.3
	C	151	143	−9	11	26	47	43	−3.9	76	74	−2.5	5.5	2.6
Qwen3-32B (no think)	A	353	352	−2	35	65	40	39	−0.4	68	68	+0.1	7.1	6.0
	C	163	145	−18	15	25	44	42	−2.6	68	69	+1.2	6.8	9.2
Llama-3.1-8B	A	526	524	−2	52	144	79	77	−1.9	89	89	−0.6	6.9	15.7
	C	207	228	+22	30	54	69	79	+10.5	82	87	+5.4	9.5	1.9
Llama-4-Scout	A	439	418	−21	48	61	58	60	+1.8	91	91	−0.3	7.0	10.9
	C	115	97	−18	19	12	39	46	+7.1	89	93	+3.7	7.3	12.8

Table 12: **Top-up behavior under incremental billing.** A and C denote access-scope and candidate-breadth pricing. Effects are $(+\text{bill}) - (\text{no bill})$; bold effects exceed both uncertainty measures.

		INCR-RATIO		ΔAGENTS | R1 miss		ΔAGENTS | R1 cover	
model	pricing	no bill	+ bill	no bill	+ bill	no bill	+ bill
Qwen3.5-122B	A	3.91	3.25	+1.32	+1.15	−0.25	−0.44
	C	3.97	3.43	+1.23	+1.17	−0.22	−0.34
Qwen3-32B (think)	A	5.40	4.73	+1.89	+1.67	+0.48	+0.07
	C	7.58	7.68	+1.77	+1.49	+0.15	−0.09
Qwen3-32B (no think)	A	3.47	3.31	+0.88	+0.86	−0.19	−0.19
	C	6.36	5.66	+0.71	+0.69	−0.14	−0.14
Llama-3.1-8B	A	5.25	5.39	+2.52	+2.24	+1.08	+1.17
	C	6.46	6.64	+2.22	+2.12	+0.62	+0.74
Llama-4-Scout	A	3.09	2.58	+1.35	+1.22	−0.42	−0.52
	C	5.91	4.78	+1.24	+1.08	−0.62	−0.72

Table 13: **Efficiency and assignment change under top-up revision.** ΔAGENTS is the change in the complete assignment size from round one to round two, reported separately for round-one misses and covered questions. A and C denote access-scope and candidate-breadth pricing.