

Capability Is Not Complementarity: A Routing Feasibility Study in Financial Question Answering

Rishi N. Simhadri

The University of Texas at Dallas
Richardson, TX, USA

Abstract

Selective routing is useful only when an expensive branch solves cases that a cheaper branch misses. We call this prerequisite expert complementarity and ask whether a capable iterative agent actually provides it. In a prospectively frozen financial question answering study, we compare Static with bounded ReAct across multiple capability tiers. Within each tier, the arms share the model, answer contract, retrieval interface, scoring, and resource limits. A stronger model removes ReAct’s average deficit and brings it to parity with Static, showing that relative performance depends on model choice. Yet Static and ReAct still succeed on nearly the same items. Across 150 shared items per tier, ReAct adds only one or two unique exact successes; oracle quality headroom point estimates are at most 1.7 points, while ReAct takes 2.1 to 2.6 times as long. A stability check preserves this pattern. Capability and complementarity are therefore distinct: improving an expensive branch can make it competitive without making it useful for routing. Evaluation contracts should be harmonized and paired complementarity audited before training a selector. This result concerns bounded ReAct over controlled benchmark evidence, not iterative agents in general.

1 Introduction

Financial QA requires evidence selection, arithmetic, and careful handling of entities, periods, signs, and units. Iterative language model agents appear well suited to these demands: they can search, inspect evidence, and revise an answer. Each extra step, however, consumes latency and tokens and creates another opportunity for retrieval, arithmetic, or formatting error. The deployment question is therefore not simply whether an agent is capable, but whether it contributes answers that a cheaper workflow misses.

We study this necessary precondition for selective agency: *expert complementarity*. A router

can improve quality relative to cost only when its expensive branch supplies enough unique correct answers. This requirement is logically prior to learning a difficulty model or tuning a threshold. It also differs from asking which branch has higher average accuracy: two systems can tie while making nearly identical errors, leaving little routing headroom.

Our main evidence is a prospectively frozen capability study on 150 development items unused by earlier studies or freezes, from FinQA, TATQA, and ConvFinQA. We compare Static, which uses one model call, with bounded ReAct that retrieves before answering. Both use a shared strict action contract, retrieval interface, scorer, and resource ceilings with GPT-4o-mini, GPT-5.6 Luna, and GPT-5.6 Terra. Static significantly outperforms ReAct for the first two tiers, whereas Terra reaches parity; the positive interaction across tiers shows an observed, model-dependent tier interaction. Nevertheless, ReAct supplies only two, one, and two unique exact successes. Oracle headroom point estimates remain at most 1.7 points under the native quality metric and at most 1.3 under exact match; a post hoc check with a different seed on the same manifest preserves this pattern of low headroom.

An earlier selector pilot across five datasets and a separate harmonized comparison motivate and support this analysis. Both are unpublished, non-concurrent experiments: they overlap with the main study in question and workflow components, but use disjoint evaluated items. The pilot missed its quality and cost rule under unequal answer and context contracts; the matched study corrected those asymmetries and again found little ReAct complementarity. We preserve both as supporting evidence rather than treating the flawed pilot as the centerpiece.

Our contribution is thus a feasibility audit for routing: model capability can change the ordering of Static and ReAct, but it does not automatically

create unique expert value. Evaluation contracts must first be harmonized, then complementarity measured, before selector optimization. This is a scoped result for bounded ReAct that retrieves before answering over controlled benchmark evidence, not a universal claim against agents.

2 Related Work

FinQA, TAT-QA, and ConvFinQA evaluate numerical reasoning over financial reports and conversations (Chen et al., 2021; Zhu et al., 2021; Chen et al., 2022); FinanceBench and FinDER emphasize open book questions and professional queries grounded in filings (Islam et al., 2023; Choi et al., 2025a). Our setting uses controlled evidence and holds document access fixed, so it evaluates orchestration rather than full retrieval.

ReAct interleaves reasoning with external actions (Yao et al., 2023), while financial agent work explores reflection, agentic retrieval, and multi-stage research workflows (Fatemi and Hu, 2024; Choi et al., 2025b; Sun et al., 2025). Prior studies also find diminishing or negative returns from additional reasoning (Chen et al., 2024b; Vamvourellis and Mehta, 2025). Selective prediction and model routing optimize tradeoffs between risk and coverage, or quality and cost (Geifman and El-Yaniv, 2019; Chen et al., 2024a; Ong et al., 2025). Routing work typically optimizes which model or workflow to call. We instead test whether the proposed expensive expert has enough unique value for any selector to be useful, and show how answer and context contracts affect that measurement. This expert feasibility audit is logically prior to learning a router.

3 Routing Feasibility

Let q be a question and $\mathcal{E}$ its controlled evidence. We compare a static workflow S with an iterative workflow R . A selector would choose R when $g(z(q, \mathcal{E})) \geq \tau$ and S otherwise, using no answer, program, scale, derivation, or other feature derived from the target.

Complementarity audit. Before fitting g , we measure the paired binary outcomes. Let n_{10} count items correct only under Static and n_{01} items correct only under ReAct. The union is the fraction answered by either branch, and its gain over the better single branch is $\min(n_{10}, n_{01})/n$. This is an optimistic ceiling: a learned router can only recover part of it and incurs its own errors and cost.

A small rare discordant cell therefore blocks useful routing even when the arms’ mean scores are competitive.

Static and iterative branches. S deterministically constructs entity, period, metric, and content word queries, ranks evidence chunks by lexical overlap, truncates their union to a fixed dossier budget, and makes one answer call grounded in evidence. R uses a bounded Search/Lookup/Finish interface over the same corpus. Its first model action must be Search; it may perform at most three evidence operations and seven model calls. Both paths record evidence exposure, parsing, calls, tokens, latency, and provider cost. These are complete retrieval policies, not pure reasoners over identical dossiers.

Supporting selector. The pilot selector used question and retrieval features independent of targets in an L2 logistic model, with a documented fallback when benefit examples were too sparse. We report its realized failure only as motivation. No selector is fit to the previously unused capability-study outcomes; the oracle is a diagnostic upper bound.

4 Experimental Protocol

Evidence and unused split. The main study uses 50 examples each from FinQA, TAT-QA, and ConvFinQA. Systems see only table, text, or dialogue evidence supplied by the benchmarks; gold answers and fields derived from targets are unavailable before Finish. The manifest excludes all identifiers used by earlier studies or stopped freezes and all protected final identifiers; ConvFinQA is also disjoint by dialogue. Final content was never loaded, executed, or scored. The same 150 items and frozen order are used for all tiers and both arms.

Capability ladder and action transport. The mandatory tiers are GPT-4o-mini (control), GPT-5.6 Luna, and GPT-5.6 Terra, all through the same OpenRouter backend restricted to the OpenAI provider, with fallbacks and provider data collection disabled. GPT-4o-mini omits an unsupported reasoning field; Luna and Terra use `reasoning_effort=none`. Each turn makes exactly one strict tool call containing action, argument, and answer; malformed output or telemetry stops the study. Static permits Finish; ReAct must begin with Search, observe evidence,

and use at least two calls. Calls have one attempt, an output ceiling of 384 tokens, prompt ceilings of 4,096 words and 8,192 bytes, and an evidence ceiling of 3,000 words.

The frozen schedule balances all six model orders across 25 items each and gives each tier 75 items with Static first and 75 with ReAct first. Two earlier freezes stopped under fixed transport rules; no partial outcomes were analyzed, and touched items were excluded. Full provenance, configurations, audit code, and sanitized metric pairs are in the public, provider-free artifact. The main run used 1,462 calls and US\$0.948; cumulative spend before replication was US\$1.007 under a US\$20 cap.

Analysis plan fixed before scoring. Quality is the unweighted macro average of FinQA exact match, official TAT-QA F1 with scale inferred from the prediction, and ConvFinQA exact match. For each tier we report a paired percentile bootstrap with 10,000 samples stratified by dataset, the exact complementarity matrix and McNemar test, oracle headroom, and efficiency. Interactions bootstrap differences between tier effects for ReAct minus Static. Scoring began only after all three tiers completed.

Calls used a fixed request seed, but temperature and top p were omitted and provider defaults applied, so generation was not guaranteed to be deterministic. After seeing the main result, we froze one post hoc replication on the same manifest that changed only the seed. It completed all 900 episodes for US\$0.951 (US\$1.959 cumulative). We report it as a stability check conditional on the same sample, not an independent confirmation.

Supporting studies. The original Gemini study across five datasets compared five workflows and a selector under unequal contracts. A later harmonized study with GPT-4o-mini on 150 separate items unused by the pilot shared contracts and verified that ReAct retrieved before answering. These unpublished, non-concurrent experiments share the research question and implementation lineage but not evaluated identifiers with the main study. They are pilot and supporting evidence, respectively.

5 Development Results

The aggregate matrices hide no dataset with substantial exact complementarity. In ConvFinQA/FinQA/TAT-QA order, per-dataset

Table 1: Main result on the same 150 previously unused items per tier. Δ is ReAct minus Static macro quality with paired 95% CI. B/S/R/N are exact match counts for both, Static only, ReAct only, and neither. EMH is exact match oracle headroom with its 95% upper limit; QH is native quality headroom with a paired 95% CI. Headroom is in percentage points.

Tier	Static	ReAct	Δ [95% CI]
Control	0.226	0.155	-0.071 [-0.120, -0.025]
Luna	0.188	0.131	-0.057 [-0.098, -0.020]
Terra	0.210	0.218	0.008 [-0.015, 0.033]

Tier	B/S/R/N	EMH [95% UL]	QH [95% CI]
Control	12/13/2/123	1.3 [4.7]	1.7 [0.2, 3.8]
Luna	7/9/1/133	0.7 [3.7]	1.0 [0.3, 1.9]
Terra	21/1/2/126	0.7 [3.7]	1.0 [0.0, 2.1]

B/S/R/N are 3/5/1/41, 0/2/1/47, and 9/6/0/35 for control; 3/2/0/45, 1/1/0/48, and 3/6/1/40 for Luna; and 5/1/0/44, 2/0/1/47, and 14/0/1/35 for Terra.

Capability changes the average ordering. Table 1 is the narrative centerpiece. Static’s macro advantage excludes zero for control and Luna. Terra instead reaches parity, with a small nonsignificant ReAct edge. The ReAct minus Static interaction is 0.079 for Terra minus control (95% CI [0.028, 0.133]) and 0.065 for Terra minus Luna ([0.018, 0.117]); Luna minus control is 0.014 ([−0.043, 0.071]). Thus the earlier ReAct deficit is not invariant to model tier.

Complementarity remains small. Neither branch is exactly correct on 123/150 control items, 133/150 Luna items, and 126/150 Terra items (82.0–88.7%). Thus low complementarity here coexists with a near-floor operating point: the systems usually fail together, rather than partitioning a large set of solvable cases. Under the actual quality metric, including F1, oracle headroom point estimates are 1.7, 1.0, and 1.0 points, with paired 95% upper limits of 3.8, 1.9, and 2.1. These intervals are tighter than exact match intervals because they use bounded continuous gains for each item rather than binomial uncertainty around only one or two rare events. Exact headroom point estimates are 1.3, 0.7, and 0.7 points; conditional on the direction of the empirically better arm, exact binomial 95% upper limits are 4.7, 3.7, and 3.7. Even 4.7 points is an oracle ceiling: a learned router can recover only part of it and adds selection errors and cost. Concretely, control has B/S/R/N = 12/13/2/123: Static gets 25/150, the union gets 27/150, and even perfect routing adds only 2/150 = 1.3 points. McNemar tests favor Static

for control ($p = 0.0074$) and Luna ($p = 0.0215$); Terra shows no directional difference ($p = 1.0$). No router is fit to these outcomes.

Efficiency. Static uses one call per item; ReAct averages 2.31, 2.27, and 2.16 calls for control, Luna, and Terra. In the same order, ReAct takes 2.6, 2.4, and 2.1 times as long and costs 1.29, 1.29, and 1.14 times as much. It is therefore neither cheaper nor faster in the main study.

Post hoc seed stability. Changing only the request seed on the same manifest reproduced Static advantages for control and Luna; Terra’s small difference changed sign but remained near parity (Static/ReAct: 0.201/0.196). Exact B/S/R/N became 13/12/1/124, 5/13/1/131, and 20/1/0/129; the rare exact oracle benefit cells were one, one, and zero. Native quality headroom was 1.4, 0.8, and 0.7 points (95% upper limits 3.1, 1.8, and 1.4). Individual generations and Terra’s point ordering vary, but the low complementarity diagnosis does not in this single check on the same sample.

Supporting evidence. On a separate harmonized v2 sample, Static scored 0.342 versus ReAct’s 0.257 ($\Delta = -0.085$, 95% CI $[-0.147, -0.025]$), with 21 successes unique to Static versus five unique to ReAct and 3.3 points of oracle headroom. The earlier pilot selector likewise missed its documented quality and cost criterion. Contract asymmetries make the pilot noncentral, but neither supporting study contradicts the main complementarity diagnosis. The provider-free artifact, including the sanitized paired records behind every matrix, is publicly available at the tagged repository snapshot.¹

6 What the Tier Interaction Means

Capability is not complementarity. Terra answers more effectively under the iterative policy than the other two tiers: its ReAct score exceeds control ReAct by 0.063 ($[0.002, 0.124]$), while its Static score does not differ reliably from control. This produces the positive interaction and answers the principal robustness objection: ReAct’s earlier loss depended partly on model choice. But Terra’s ReAct successes almost entirely overlap its Static successes. The system tie therefore does not imply a useful mixture.

¹Camera-ready repository snapshot; see paper/realms2026/artifact.

The distinction is mechanical. A selector can recover at most the oracle union, whose gain over the better branch equals the rarer discordant cell. The point estimates are small, but the uncertainty bounds prevent treating them as hard ceilings. The conclusion instead survives the optimistic bound: even 4.7 exact match points is oracle performance that a learned selector only partly recovers, purchased here at 1.14 to $1.29\times$ the cost and 2.1 to $2.6\times$ the latency. This does not produce a favorable measured relation between quality and cost.

Why contract fairness matters. The pilot used unequal final answer and context contracts, and a post hoc author audit found discordant outputs sensitive to the scorer. We neither rescore those cases nor call the audit independent adjudication. The harmonized and capability studies instead share answer semantics, ceilings, scorer input, single attempt rules, and provider binding. They share a retrieval API and budget, not an identical retrieved dossier: Static averages about 2.85 retrieval operations, whereas ReAct adaptively uses 1.16 to 1.31. The estimates therefore compare deployed retrieval policies, including effects from query construction and evidence selection; they do not isolate iteration from retrieval behavior.

The practical sequence is: harmonize evaluation contracts, measure paired complementarity, and only then optimize a selector. At the three tiers, only two, one, and two outcomes are ReAct-only positives, too few to support learning a reliable selector here. Cost is conditional: an oracle would pay ReAct’s extra inference only on routed items, not on all items, but a learned router must identify those rare benefits without absorbing the weaker branch’s many unique errors. Retuning on the present outcomes and reporting the same sample as evaluation would invalidate that sequence.

7 Conclusion

Across three prospectively frozen model tiers, capability changes the relative accuracy of bounded ReAct that retrieves before answering: Static wins for control and Luna, while Terra reaches parity. Native quality oracle headroom point estimates are at most 1.7 points (95% upper limits at most 3.8); exact match estimates are at most 1.3 (upper limits below 4.8). A check with a different seed on the same sample retains the low headroom pattern. Even the optimistic bounds are oracle ceilings, while ReAct costs more calls, time, and dollars. Stronger ca-

pability can erase an iterative workflow’s average deficit without creating useful expert complementarity. Routing should therefore begin with a paired feasibility audit under fair contracts, not a selector.

Limitations

Evidence and inference scope. We evaluate orchestration over evidence supplied by the benchmarks, not full RAG, complete filings, OCR, or missing evidence. All samples are development evidence; final identifiers served only for overlap exclusion, and final content was never loaded or scored. No router is evaluated on the main outcomes; the oracle is diagnostic.

Model, policy, and scoring dependence. The three provider product tiers do not constitute a general capability scale, and Luna is not monotonically better on these tasks. We test one prompt family, one bounded ReAct policy that retrieves before answering and has reasoning disabled, one lexical evidence environment, and native metrics. Nonzero reasoning is the key untested boundary and requires a prospectively frozen later study, not a post hoc arm. The arms share contracts and ceilings, not identical dossiers, so effects include differences in retrieval policy rather than pure iteration. Provider behavior, prices, and latency can change. Although the request seed was fixed, temperature and top p used provider defaults and determinism was not guaranteed; one replication with a different seed does not characterize full generation variance.

Operating point. FinQA and ConvFinQA accuracy is near floor. TAT-QA is higher (main run F1 from 0.314 to 0.493), yet continuous headroom is at most 2.9 points (2.5 in the seed check). Complementarity may therefore emerge at realistic deployment accuracy levels that this study does not reach.

Public-benchmark exposure. All three benchmarks are public and may have appeared in model pretraining or other post-training data. Model providers do not expose sufficient provenance to test memorization. Supplying controlled evidence does not remove this risk, so absolute scores and tier differences should not be read as performance on previously unseen financial documents.

References

- Lingjiao Chen, Matei Zaharia, and James Zou. 2024a. [FrugalGPT: How to use large language models while reducing cost and improving performance](#). *Transactions on Machine Learning Research*.
- Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2024b. [Do NOT think that much for \$2 + 3 = ?\$ on the overthinking of o1-like LLMs](#). *Preprint*, arXiv:2412.21187.
- 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 the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711. Association for Computational Linguistics.
- Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. [ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6292. Association for Computational Linguistics.
- Chanyeol Choi, Jihoon Kwon, Jaeseon Ha, Hojun Choi, Chaewoon Kim, Yongjae Lee, Jy-yong Sohn, and Alejandro Lopez-Lira. 2025a. [FinDER: Financial dataset for question answering and evaluating retrieval-augmented generation](#). In *Proceedings of the 6th ACM International Conference on AI in Finance*, pages 638–646. Association for Computing Machinery.
- Chanyeol Choi, Jihoon Kwon, Alejandro Lopez-Lira, Chaewoon Kim, Minjae Kim, Juneha Hwang, Jaeseon Ha, Hojun Choi, Suyeol Yun, Yongjin Kim, and Yongjae Lee. 2025b. [FinAgentBench: A benchmark dataset for agentic retrieval in financial question answering](#). In *Proceedings of the 6th ACM International Conference on AI in Finance*, pages 632–637. Association for Computing Machinery.
- Sorouralsadat Fatemi and Yuheng Hu. 2024. [Enhancing financial question answering with a multi-agent reflection framework](#). In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 530–537. Association for Computing Machinery.
- Yonatan Geifman and Ran El-Yaniv. 2019. [SelectiveNet: A deep neural network with an integrated reject option](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2151–2159. PMLR.
- Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. [FinanceBench: A new benchmark for financial question answering](#). *Preprint*, arXiv:2311.11944.

- Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. 2025. [RouteLLM: Learning to route LLMs with preference data](#). In *International Conference on Learning Representations*.
- Rui Sun, Zuo Bai, Wentao Zhang, Yuxiang Zhang, Li Zhao, Shan Sun, and Zhengwen Qiu. 2025. [Fin-ResearchBench: A logic tree based agent-as-a-judge evaluation framework for financial research agents](#). In *Proceedings of the 6th ACM International Conference on AI in Finance*, pages 656–664. Association for Computing Machinery.
- Dimitris Vamvourellis and Dhagash Mehta. 2025. [Reasoning or overthinking: Evaluating large language models on financial sentiment analysis](#). In *Proceedings of the 6th ACM International Conference on AI in Finance*, pages 299–307. Association for Computing Machinery.
- 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 *International Conference on Learning Representations*.
- 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 the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287. Association for Computational Linguistics.