Performance Engineering of Retrieval Architectures for Enterprise Supply-Chain Question Answering

Alex Carroll and Elenna Dugundji
MIT Center for Transportation and Logistics
Cambridge, Massachusetts, United States
Corresponding author: Alex Carroll, alex629@mit.edu

***Abstract*—Enterprise question answering often selects evidence from heterogeneous documents, relational tables, event records, and supply-chain graphs before generating an answer. This source-selection stage creates a systems trade-off: local retrieval is inexpensive, while task-aware agentic routing can invoke additional reasoning, tools, and verification. We present a common harness and an 80-query, six-source workload for comparing TF-IDF lexical retrieval, deterministic GraphRAG-light, fixed lexical–graph fusion, retrieval-result caching, and a live agentic source router. We measure exact source-set accuracy, set F1, multi-source recall, abstention, latency, call counts, stability, and failure modes. The agentic router achieved 82.50% exact source-set accuracy and 86.84% set F1, compared with 38.75% and 61.00% for TF-IDF. GraphRAG-light did not improve the lexical baseline; fixed fusion increased exact accuracy to 41.25%. Agentic execution had 6.67 s mean and 9.88 s P95 latency, whereas local retrieval operated at millisecond scale. Fully warm cache hits preserved hybrid predictions and reduced repeated-result lookup to approximately 0.015 ms. The results quantify an accuracy–latency envelope without claiming functional equivalence between retrieval-only, cached, and live end-to-end execution.**

***Index Terms*—retrieval-augmented generation, graph retrieval, agentic systems, performance engineering, source routing, supply-chain question answering**

I. Introduction

Retrieval-augmented generation (RAG) grounds language-model output in external evidence [1]. Enterprise workloads, however, may require a contract clause, structured rate, event record, supplier profile, spend history, graph relationship, or several of these sources simultaneously. The system must select authoritative sources and decide whether to expand or abstain before answer generation.

Tool-using language models can interleave reasoning and external actions [2] or learn when to invoke tools [3]. This task-aware control may improve source selection but adds model calls, network latency, tool executions, and verifier loops. Graph-augmented retrieval can connect evidence that chunk-local similarity misses [4], but graph matching and expansion also add computation and can propagate irrelevant neighbors. Caching removes repeated deterministic work but does not replace fresh reasoning or generation. The relevant systems question is how accuracy, latency, tail latency, stability, and call burden change across these architectures under one controlled workload.

This paper contributes: (1) a common harness for lexical, graph, lexical–graph hybrid, cache-enabled, and agentic source retrieval; (2) a quantitative comparison using strict source-set accuracy, set F1, multi-source recall, abstention, tail latency, call counts, and stability; and (3) evidence that local retrieval is orders of magnitude faster but less accurate than live agentic routing, lightweight graph expansion does not automatically improve retrieval, fixed fusion gives a modest local gain, and caching removes repeated lookup cost without replacing end-to-end agentic reasoning.

The study is an AI-systems benchmark, not a production-assistant evaluation. Its primary dependent variable is source selection under strict set equality; Partial-overlap credit is not used as the headline accuracy metric.

II. Related Work and Positioning

RAG and retrieval architectures. RAG grounds generation in retrieved external evidence [1]. Many evaluations emphasize final answer quality, but enterprise QA also depends on selecting the correct authority before generation. In procurement, the relevant authority may be a contract clause, rate table, supplier record, event log, graph relation, spend database, or a combination of sources. This paper therefore evaluates source-set selection directly, using strict set equality rather than answer plausibility or partial evidence overlap.

Graph-augmented retrieval. GraphRAG methods use entities, relationships, and graph neighborhoods to connect evidence that chunk-level similarity may miss [4]. Our GraphRAG-light baseline is intentionally narrower than full GraphRAG: it uses deterministic typed-node matching and one-hop expansion over the available procurement graph, without LLM-generated graph summaries or learned graph retrieval. This isolates a specific systems question: whether lightweight graph structure improves source routing enough to justify its added retrieval cost.

Agentic AI systems and benchmarking. Agentic systems use language models to choose tools, interpret observations, and continue retrieval when initial evidence is incomplete. ReAct interleaves reasoning and action [2], and Toolformer studies learned tool use [3]. These systems motivate measuring not only accuracy, but also latency, call counts, stability, abstention, and failure modes. Selective prediction and unanswerable QA work show why abstention must be reported explicitly [5], [6], while behavioral testing motivates targeted error categories beyond a single score [7].

III. Benchmark and Retrieval Architectures

A. Workload

The curated benchmark contains 80 natural-language queries over six enterprise procurement source classes: unstructured contracts; structured contract rates; supplier and procurement records; RFx, covering request-for-information, request-for-proposal, and request-for-quote events; a procurement relationship graph; and spend data. The workload contains 61 single-source, 15 multi-source, and four gold-abstention queries. Questions cover clauses, rates, profiles, spend, events, graph traversal, cross-source evidence, ambiguity, and unsupported requests. The existing query-class metadata contains 26 classes. The largest groups are cross-database questions (10), ambiguous questions (8), spend analytics (8), clause retrieval (7), and adversarial or unsupported requests (6). This mixture is deliberate: the benchmark includes easy single-source lookup, but it also tests authority boundaries where lexical similarity, graph adjacency, and task-aware reasoning can disagree.

Each query has a gold source set and abstention label. Adapters receive only the query identifier and text; gold labels never enter retrieval, fusion, graph expansion, cache keys, or predictions. Every configuration emits the same trace schema. Unsupported measurements remain unavailable rather than being converted to zero.

B. Harness and measurement model

All configurations run through one benchmark harness. The harness loads query text and immutable labels, invokes an adapter for the selected architecture, and writes one JSONL trace row per query and run. Architecture-specific code is confined to adapters: the TF-IDF, graph, hybrid, cache, and agentic systems all return predicted sources, retrieved identifiers where available, abstention decisions, stage latencies, call counts, and error fields through the same interface.

This design keeps the evaluation gold-blind. Gold source sets are used only after an adapter returns. Cache keys are derived from query text, configuration, fusion parameters, and artifact-manifest hashes; they exclude gold sources, query classes, and correctness labels. Missing measurements are represented as unavailable values, not zeros, because a zero would imply a measured absence of work rather than a trace limitation. This distinction matters most for token and cost accounting in the preserved agentic baseline and for final-answer fields in retrieval-only baselines.

C. Local retrieval baselines

TF-IDF lexical retrieval indexes a normalized corpus spanning all six source classes and returns the top five unigram/bigram TF-IDF chunks. Predicted sources are the unique source labels attached to those chunks. This is lexical retrieval, not semantic embedding RAG, and it does not generate an answer.

GraphRAG-light builds a typed graph over source, document, chunk, supplier, contract, event, category, material, and procurement-relationship records. Query terms deterministically match nodes, matched nodes expand by one hop, and attached chunks determine predicted sources. It uses no LLM for graph construction or query-time extraction and is not the full GraphRAG pipeline of Edge et al. [4].

Hybrid retrieval merges lexical and graph candidates by chunk identifier using the frozen score

$$s = 0.70 s_{\mathrm{tfidf}} + 0.25 s_{\mathrm{graph}} + 0.05 b_{\mathrm{agree}}, \tag{1}$$

with branch-normalized scores, agreement bonus b_{agree}, top-$k = 5$, and minimum score 0.05. Parameters were fixed before the full benchmark.

Cache-enabled hybrid stores deterministic hybrid results under a SHA-256 key covering the query, configuration, fusion parameters, and artifact manifest hashes. Cold runs clear the SQLite cache; warm runs use read-only lookup. Values contain predicted sources and retrieved identifiers, never gold labels, query classes, tokens, or answers.

D. Live agentic pipeline

The preserved **agentic source router** performs LLM-mediated planning and routing, invokes source-specific tools, conditionally applies verifier-gated follow-up retrieval, and generates a final answer. Its execution includes model, network, service, tool, verification, and generation latency. Local baselines stop after source retrieval, so the latency comparison represents a systems trade-off, not equal functionality.

IV. Metrics and Experimental Setup

Exact source-set accuracy requires equality between predicted and gold source sets. We also report set precision, recall, F1, multi-source recall, and binary abstention accuracy. Because selective behavior depends on class prevalence [5], [6], failure analysis separately reports missed and false abstentions. Deterministic error categories—strict over-routing, strict under-routing, disjoint prediction, partial overlap, and multi-source recall failure—characterize behavior beyond aggregate accuracy [7] without changing labels.

For a query with gold source set S and prediction $\hat{S}$, exact accuracy is $1[\hat{S} = S]$. Set precision and recall are computed from $|\hat{S} \cap S|$ with conservative empty-set handling: a gold abstention is evaluated by the abstention label, while a non-empty gold set with an empty prediction receives zero source recall. Multi-source recall is the fraction of gold sources recovered on queries where $|S| > 1$. The headline accuracy is strict set equality; no special overlap credit is assigned to ambiguous queries.

Systems metrics include mean, median, and P95 total latency, retrieval calls, traced LLM calls, cache hits, and repeated-run source-set stability. Each main configuration ran three times over the same 80 queries. Local baselines ran in a CPU-only Python 3.11 process on an 18-logical-core arm64 macOS host with no GPU or external model. The agentic baseline used live external model and source services. Cache characterization separated clear-cache initialization, misses, and three fully warm read-only trials. Agentic token, cost,

TABLE I
RETRIEVAL ARCHITECTURE CHARACTERISTICS. LOCAL BASELINES PERFORM SOURCE RETRIEVAL ONLY AND DO NOT GENERATE FINAL ANSWERS.

Configuration	Retrieval architecture	Graph	LLM	Verifier	Cache	Retrieval calls/query	LLM calls/query
TF-IDF	Lexical TF-IDF	No	No	No	No	1.00	0.00
GraphRAG-light	Typed graph expansion	Yes	No	No	No	1.00	0
Hybrid	TF-IDF + graph fusion	Yes	No	No	No	2.00	0
Hybrid warm cache	Cached hybrid result	Yes	No	No	Yes	0.00	0
Agentic router	Agentic source routing	Tool-selective	Yes	Yes	No	2.84	3.64

Agentic call counts are trace-derived lower bounds; local graph and hybrid traces store LLM calls as unavailable, but the architectures invoke no LLM.

and some intermediate call counts remain trace-derived lower bounds.

Table II summarizes the repeated-run quality, latency, and stability measurements.

V. RESULTS

A. Accuracy–latency trade-off

The live agentic router achieved 82.50% exact accuracy and 86.84% set F1. TF-IDF achieved 38.75% and 61.00%; GraphRAG-light achieved 36.25% and 59.76%; and hybrid retrieval achieved 41.25% and 62.04%. Thus, deterministic one-hop graph expansion did not outperform lexical retrieval on this workload. This does not imply that graph retrieval is generally ineffective; this lightweight graph, matching, and expansion strategy did not overcome its source boundary errors.

Hybrid improved exact accuracy by 2.5 percentage points over TF-IDF and 5.0 points over pure GraphRAG-light. Agentic routing remained 41.25 percentage points more accurate than hybrid. Its largest quality advantage was multi-source recall: 74.07% versus 34.44% for TF-IDF and hybrid.

Figure 1 shows the measured operating frontier. The agentic pipeline required approximately 6.67 s mean and 9.88 s P95 latency, with 2.84 retrieval calls and 3.64 traced LLM calls per query. Hybrid P95 was 3.18 ms; TF-IDF and GraphRAG-light P95 were 0.856 and 2.43 ms. The TF-IDF warm median is a latency lower bound for local lexical source retrieval, not a semantic-RAG quality target or full answer latency.

B. Cache characterization

Table III separates initialization, miss, and warm-hit behavior. The clear-cache first query required 761.96 ms to initialize retrieval artifacts. The remaining 79 misses averaged 2.85 ms. Three fully warm trials averaged 0.0147 ms with 0.0170 ms P95 latency, a 100% hit rate, and zero underlying retrieval calls.

Figure 2 places these conditions beside the live agentic reference. Warm hits preserved every hybrid prediction and therefore retained 41.25% exact accuracy. The approximately 0.015 ms result is a lower bound for repeated deterministic result lookup. It excludes fresh retrieval, reasoning, verification, and answer generation and is not end-to-end QA latency.

C. Query classes and stability

Class-level results show where the agentic advantage comes from. On the ten cross-database questions, the agentic router reached 90.0% exact accuracy and 95.7% set F1, compared with 10.0% exact accuracy and 48.3% set F1 for both TF-IDF and hybrid retrieval. On the eight ambiguous queries, TF-IDF and hybrid had 0% exact accuracy under strict set equality, while the agentic router reached 33.3% exact accuracy and 61.3% set F1. The agentic router also handled clause retrieval, event detail, graph traversal, and several supplier profile classes perfectly across the repeated trials. Its weaker classes were adversarial or unsupported questions, ambiguous source-boundary questions, and some supplier or spend cases where an otherwise plausible source set omitted an authority required by the gold label.

Repeated-run stability separates deterministic retrieval from model-mediated routing. TF-IDF, GraphRAG-light, hybrid retrieval, and warm-cache hybrid had 100% source-set stability because they are deterministic over a fixed corpus and configuration. The agentic router had 98.33% mean source-set stability: four of 80 queries changed predicted source sets across the three live runs. Its abstention stability was 99.17%, with two queries changing abstention decision. These changes did not create execution errors, but they show that model-mediated planning introduces a small amount of routing variance even when the benchmark, labels, and source registry are fixed.

VI. FAILURE ANALYSIS AND DISCUSSION

Table IV summarizes deterministic post hoc failure modes computed from modal predicted source sets. Modal predictions produced 49 exact mismatches for TF-IDF, 51 for GraphRAG-light, 47 for hybrid and warm-cache hybrid, and 14 for the agentic router. Local errors were dominated by source-boundary mistakes. TF-IDF had 20 strict over-routing errors and 14 multi-source recall failures. GraphRAG-light had 19 over-routing errors and failed full recall on all 15 multi-source queries. Hybrid reduced over-routing to 18 but retained 14 multi-source recall failures.

Agentic errors were fewer but included seven under-routing errors, four disjoint predictions, six multi-source recall failures, and abstention errors. Its source-set stability was 98.33%, compared with 100% for deterministic local configurations. Planning and verifier-gated expansion reduce broad source

TABLE II
MAIN ROUTING-QUALITY AND LATENCY RESULTS ACROSS THREE REPEATED RUNS.

Configuration	Exact (%)	Set F1 (%)	Multi-source recall (%)	Abstention (%)	Mean (ms)	Median (ms)	P95 (ms)	Stability (%)
TF-IDF	38.75	61.00	34.44	95.00	3.752	0.785	0.856	100.00
GraphRAG-light	36.25	59.76	32.78	93.75	1.947	1.574	2.426	100.00
Hybrid	41.25	62.04	34.44	95.00	5.640	2.428	3.182	100.00
Hybrid warm cache	41.25	62.04	34.44	95.00	0.0147	0.0139	0.0170	100.00
Agentic router	82.50	86.84	74.07	95.00	6,671.8	6,836.1	9,877.4	98.33

TF-IDF, GraphRAG-light, hybrid, and warm-cache configurations are local retrieval-only baselines. Warm-cache lookup is not functionally equivalent to live agentic execution. For local retrieval baselines, mean latency includes first-query artifact loading; median and P95 reflect warm-query behavior. Cache characterization separates first-query and non-first-query latency.

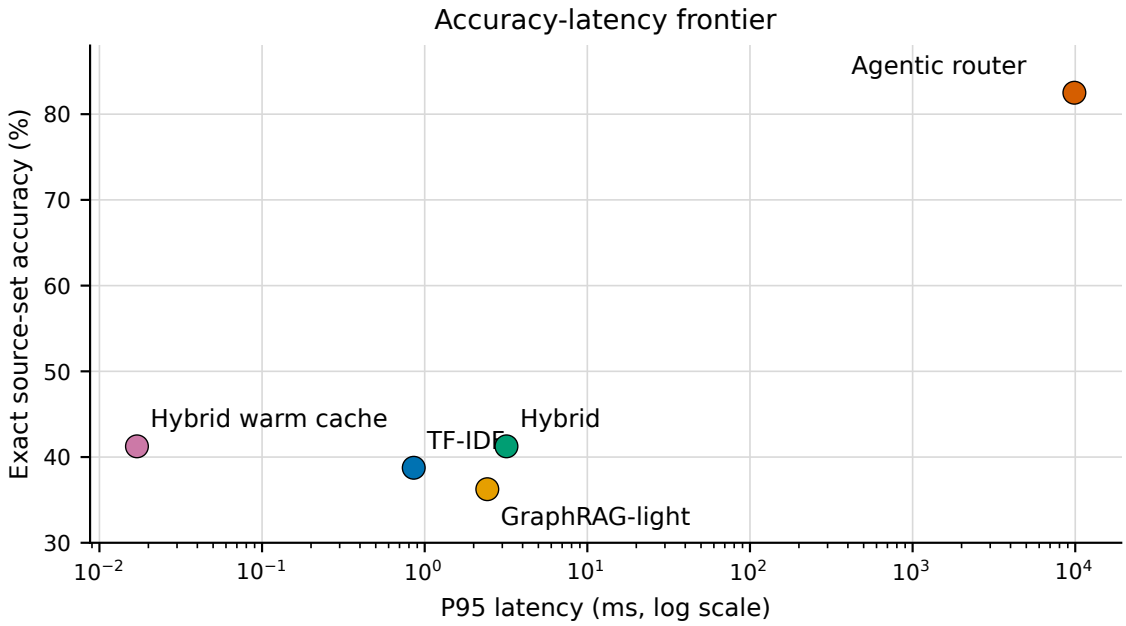


Fig. 1. Exact source-set accuracy versus P95 latency on a logarithmic axis. Local configurations are retrieval-only; agentic execution includes live routing, tools, verification, and generation.

TABLE III
CACHE CHARACTERIZATION FOR THE FIXED HYBRID CONFIGURATION.

Condition	Exact (%)	Set F1 (%)	Mean (ms)	Median (ms)	P95 (ms)	First query (ms)	Non-first mean (ms)	Hit rate (%)	Retrieval calls/query
Uncached hybrid	41.25	62.04	5.640	2.428	3.182	760.2	2.483	–	2.00
Cold cache run	41.25	62.04	12.3	2.811	3.547	762.0	2.852	0.00	2.00
Cache miss	41.25	62.04	2.852	2.802	3.447	–	2.852	0.00	2.00
Warm cache hit	41.25	62.04	0.0147	0.0139	0.0170	0.125	0.0147	100.00	0.00
Agentic router	82.50	86.84	6,671.8	6,836.1	9,877.4	6,930.1	6,670.7	–	2.84

Cache-miss latency excludes the first cold-load query. Warm hits return previously computed retrieval outputs and do not execute either retrieval branch.

confusion, but remaining errors concentrate in omitted evidence and authority boundaries.

The failure-mode pattern clarifies why the hybrid gain is limited. The graph branch changes some boundary decisions, but fixed fusion often preserves the same top lexical evidence and therefore does not recover missing sources in multi-source questions. GraphRAG-light failed to recover the complete gold source set for every multi-source query, showing that one-hop expansion over the available typed graph is not sufficient to identify all required authorities. The agentic router does not merely retrieve more sources; it changes the control policy by planning which tools to call, inspecting source-specific outputs, and conditionally performing follow-up retrieval. That policy reduces over-routing sharply, from 18–20 local over-routing errors to one agentic over-routing error, but it still misses some required sources.

Abstention accuracy requires care. Only four of 80 queries require abstention. TF-IDF, hybrid, and cached hybrid never abstained yet scored 95% because 76/80 queries are answerable; all four true abstentions were missed. GraphRAG-light

TABLE IV
POST HOC DETERMINISTIC SOURCE-SET FAILURE ANALYSIS. PERCENTAGES USE THE APPLICABLE QUERY SUBSET FOR EACH FAILURE MODE.

Configuration	Exact mismatch	Over	Under	Disjoint	Partial	Multi-recall fail.	False abst.	Missed abst.	Ambiguous mismatch
TF-IDF	49 (61.3%)	20 (26.3%)	5 (6.6%)	15 (19.7%)	30 (39.5%)	14 (93.3%)	0 (0.0%)	4 (100.0%)	8 (100.0%)
GraphRAG-light	51 (63.7%)	19 (25.0%)	9 (11.8%)	16 (21.1%)	30 (39.5%)	15 (100.0%)	1 (1.3%)	4 (100.0%)	5 (62.5%)
Hybrid	47 (58.8%)	18 (23.7%)	5 (6.6%)	15 (19.7%)	28 (36.8%)	14 (93.3%)	0 (0.0%)	4 (100.0%)	8 (100.0%)
Hybrid warm cache	47 (58.8%)	18 (23.7%)	5 (6.6%)	15 (19.7%)	28 (36.8%)	14 (93.3%)	0 (0.0%)	4 (100.0%)	8 (100.0%)
Agentic router	14 (17.5%)	1 (1.3%)	7 (9.2%)	4 (5.3%)	6 (7.9%)	6 (40.0%)	3 (3.9%)	1 (25.0%)	5 (62.5%)

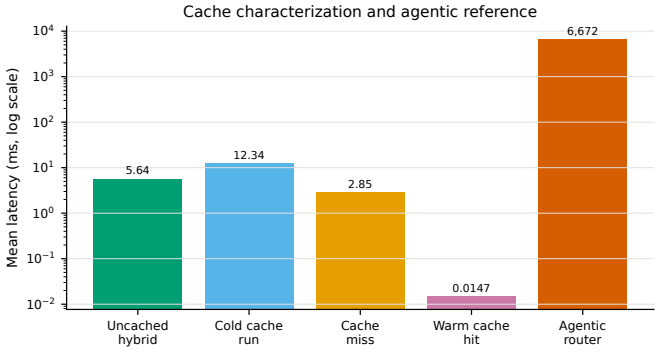


Fig. 2. Mean latency for hybrid retrieval and cache conditions, with live agentic execution as a reference. The vertical axis is logarithmic.

also missed all four and added one false abstention. The agentic modal decision missed one true abstention and falsely abstained on three answerable queries.

The results identify an operating frontier rather than a universally best architecture. Deterministic retrieval suits latency-sensitive workloads with limited source ambiguity. Under this workload, agentic routing is preferable when exact authority selection and multi-source coverage justify seconds of additional latency. Fixed lexical–graph fusion occupies a middle point but does not close the task-aware reasoning gap.

VII. LIMITATIONS

The benchmark contains 80 curated queries over six source classes; its source distribution and four-query abstention subset limit generalization. TF-IDF is lexical rather than semantic retrieval. GraphRAG-light is deterministic one-hop expansion, not full GraphRAG, so its negative result is specific to this design and workload. Local configurations retrieve source-labeled chunks but do not generate answers; agentic execution includes external services and generation, making latency asymmetric. Network and service load affect live measurements. Some legacy document identifiers and intermediate calls are untraced, so agentic token and cost accounting is incomplete. Cache latency measures exact query lookup, not semantic caching, invalidation, or end-to-end QA.

Several design choices also make the results conservative. The local baselines use simple deterministic retrieval rather than tuned learned rankers, neural embeddings, or LLM-generated graph queries. The hybrid weights were frozen before the full benchmark rather than optimized on the test labels. The graph baseline uses lightweight entity and relationship structure available in the underlying enterprise benchmark artifacts, not a separately curated knowledge graph. These choices make the baselines auditable and reproducible, but stronger local systems could move the local accuracy frontier. Finally, source-set accuracy measures authority selection, not factual answer quality; a system can select the right source set and still generate an incorrect answer, or select an incomplete source set and produce a plausible but unsupported response.

VIII. CONCLUSION

Across one gold-blind harness, local TF-IDF, GraphRAG-light, and hybrid retrieval operated at millisecond scale but were substantially less accurate than a live agentic source router. Pure lightweight graph expansion did not improve lexical retrieval; frozen lexical–graph fusion provided a modest exact-accuracy gain. Retrieval-result caching removed repeated deterministic work while preserving predictions but did not replace fresh retrieval or agentic reasoning. These measurements provide an accuracy, latency, stability, and failure-mode basis for selecting retrieval architectures according to workload requirements. The practical implication is not that one architecture dominates every deployment. Rather, the correct choice depends on the value of exact authority selection, the cost of seconds-scale live reasoning, the expected repetition rate of queries, and the risk of multi-source omissions or missed abstentions.

RELATIONSHIP TO COMPANION WORK

A companion manuscript by the same authors evaluates procurement-aware agentic source-routing policy bundles for heterogeneous enterprise RAG. This HPEC paper uses the same enterprise source universe and curated workload for a different purpose: benchmarking the accuracy–latency trade-offs of lexical, graph, hybrid, cache-enabled, and live agentic retrieval architectures. The contribution here is a performance-engineering comparison of retrieval architectures rather than a duplicate evaluation of the agentic routing policy itself.

ACKNOWLEDGMENTS

ChatGPT (OpenAI) was used solely for language editing, including checking spelling, grammar, punctuation, clarity, and readability. The authors reviewed and approved all suggested edits and remain fully responsible for the content, analysis, interpretations, and conclusions presented in this manuscript.

REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[2] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," *arXiv preprint arXiv:2210.03629*, 2022.

[3] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," *Advances in neural information processing systems*, vol. 36, pp. 68 539–68 551, 2023.

[4] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, and J. Larson, "From local to global: A graph rag approach to query-focused summarization," *arXiv preprint arXiv:2404.16130*, 2024.

[5] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," *Advances in neural information processing systems*, vol. 30, 2017.

[6] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 784–789.

[7] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of nlp models with checklist," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 4902–4912.