

Retrieval-Augmented Generation with Multi-Role LLMs for Evidence-Aware Search

Fenne Oversteeg

School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
fenne1971@uc.edu

Angelino Alvaro

Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
angelinowork@colostate.edu

Abstract

Retrieval-augmented generation has become a central architecture for grounding large language models in external evidence, yet most deployed systems still treat retrieval, reranking, and generation as a linear pipeline rather than a contested socio-technical process. This paper examines retrieval-augmented generation with multi-role large language models for evidence-aware search, focusing on how specialized model roles can improve search quality while introducing new governance, robustness, fairness, and sustainability challenges. We synthesize research on dense retrieval, fusion-in-decoder architectures, retrieval-augmented pretraining, late interaction, benchmarking, self-reflection, tool use, planning, long-context handling, irrelevant context susceptibility, parametric memory limits, calibration, hallucination, and foundation model risks. We then analyze a multi-role architecture in which separate language model roles perform query interpretation, evidence retrieval, adversarial verification, synthesis, and audit. The paper argues that evidence-aware search should be understood as infrastructure: it requires versioned evidence stores, provenance metadata, role separation, observability, escalation paths, and policy controls that extend beyond model accuracy. Multi-role designs can improve precision, reduce hallucination, and support contestability, but they also increase latency, cost, energy use, and coordination failure modes. We conclude that sustainable deployment depends on matching role complexity to domain risk, investing in auditability and data governance, and treating retrieval corpora and ranking policies as public-interest infrastructure rather than opaque private assets.

Keywords

retrieval-augmented generation; multi-role large language models; evidence-aware search; information retrieval; artificial intelligence governance; robustness; fairness; sustainability.

1. Introduction

Retrieval-augmented generation has emerged as a dominant strategy for connecting large language models to external knowledge, especially in domains where factual freshness, source attribution, and domain specificity matter more than unconstrained linguistic fluency [1]. In contrast to purely parametric systems that store knowledge implicitly within model weights, retrieval-augmented systems separate knowledge storage from reasoning and generation, allowing updates to the evidence base without retraining the entire model. This separation is attractive for evidence-aware search because it creates a plausible pathway toward verifiable citations, traceable reasoning, and institutional control over knowledge sources. However, the apparent simplicity of retrieve-then-generate pipelines obscures the

complex decisions that shape what evidence is available, which passages are ranked highly, how conflicting sources are reconciled, and who bears responsibility when generated claims are wrong [2]. The growing literature on retrieval-augmented generation therefore needs to be read not merely as a technical optimization problem but as a systems and governance problem [8].

Evidence-aware search intensifies this concern because the purpose of retrieval is not only to improve answer quality but also to make the evidential basis of an answer inspectable. In legal research, clinical decision support, public policy analysis, investigative journalism, and enterprise compliance, a correct answer with an unattributable or fragile evidence chain is often unacceptable. A system may retrieve a relevant statute, a contradictory case, a withdrawn clinical guideline, or a partisan briefing, and the final response may be shaped as much by ranking and synthesis choices as by the underlying documents. Multi-role large language models offer one response to this challenge by dividing labor among specialized agents: one role interprets the query, another expands or reformulates it, another retrieves candidate evidence, another critiques the evidence for relevance and reliability, another synthesizes an answer, and still another audits the result for unsupported claims. This division of labor can create useful checks and balances, but it also multiplies failure points, increases computational overhead, and raises difficult questions about accountability when roles disagree.

This paper develops a system-level account of retrieval-augmented generation with multi-role large language models for evidence-aware search. It emphasizes architecture, governance, infrastructure, deployment, sustainability, robustness, fairness, and policy implications rather than narrow model benchmarking. The central argument is that multi-role evidence-aware search should be treated as critical information infrastructure. Its value depends not only on model capability but also on the quality of retrieval corpora, the transparency of ranking policies, the robustness of role coordination, and the presence of audit trails that allow institutions and affected publics to contest outputs. The paper proceeds by reviewing the technical foundations of retrieval-augmented generation, describing a multi-role architecture for evidence-aware search, analyzing governance and infrastructure requirements, examining operational trade-offs, and discussing robustness, fairness, policy, and accountability.

2. Background and Related Work

The technical foundation of retrieval-augmented generation lies in the combination of neural retrieval and conditional generation. Early work demonstrated that a generator can condition on retrieved documents to improve knowledge-intensive tasks, effectively allowing a language model to consult an external memory at inference time [1]. Dense passage retrieval advanced this direction by learning embeddings that map questions and passages into a shared semantic space, making retrieval more robust to lexical variation than sparse keyword matching [2]. Fusion-in-decoder architectures further showed that generative models can integrate multiple retrieved passages rather than relying on a single best document, which is essential for questions that require synthesis across sources [3]. Retrieval-augmented pretraining extended the idea by incorporating retrieval during pretraining itself, suggesting that models can learn to use external evidence more effectively when retrieval is part of their developmental environment [5]. Large-scale retrieval from trillion-token corpora demonstrated that retrieval can improve language modeling at scale, though it also raised concerns about data quality, licensing, and the environmental cost of massive indexes [4].

Ranking and evaluation form a second foundation. Late interaction models such as ColBERT improved retrieval effectiveness by preserving token-level interactions between queries and documents while remaining efficient enough for large-scale search [6]. Heterogeneous benchmarks such as BEIR revealed that retrieval models often fail to generalize across domains, a finding with direct consequences for evidence-aware search in specialized fields [7]. Surveys of retrieval-augmented generation have consolidated these developments and highlighted open problems in retrieval quality, attribution, evaluation, and integration with reasoning [8]. At the same time, self-reflective and agentic approaches have pushed language models beyond single-pass generation. Self-RAG trains models to decide when to retrieve, critique retrieved passages, and revise outputs, thereby making evidence use more explicit [9]. ReAct combines reasoning traces with actions such as search or tool use, providing a template for multi-step evidence gathering [10]. Reflexion uses verbal reinforcement to let agents learn from prior mistakes, which is relevant for iterative search and verification [11].

Reasoning and planning research also informs multi-role architectures. Chain-of-thought prompting showed that eliciting intermediate reasoning steps can improve performance on complex tasks, although it does not guarantee factual correctness [12]. Plan-and-solve prompting separated planning from execution, improving zero-shot reasoning by first constructing a plan and then carrying it out [13]. These approaches matter for evidence-aware search because complex queries often require decomposition into sub-questions, selection of evidence sources, resolution of conflicts, and synthesis under uncertainty. However, the same literature warns that reasoning traces can be misleading, that retrieved context can distract models, and that multi-step pipelines can compound errors. A multi-role design therefore cannot be justified solely by the promise of better reasoning; it must be evaluated against its added cost, opacity, and governance burden.

3. Multi-Role LLM Architecture for Evidence-Aware Search

A multi-role architecture for evidence-aware search begins with the recognition that search is not a single cognitive act. It involves interpreting an often ambiguous query, deciding what kinds of evidence would be authoritative, retrieving candidate documents, judging relevance and credibility, identifying contradictions, synthesizing an answer, and communicating uncertainty. In a multi-role system, these tasks are distributed across distinct language model roles or agents, each with its own prompt, tools, permissions, and evaluation criteria. A query interpreter may clarify scope, jurisdiction, time period, and intent. A retrieval planner may choose among indexes, databases, and web sources. A reranker may reorder candidate passages according to relevance, recency, authority, and diversity. A verifier may search for counterevidence or check whether claims are supported by cited passages. A synthesizer may produce the final response, while an auditor may review the trace for unsupported inferences or missing citations. A recent multi-role collaborative reranking workflow illustrates how several model roles can critique and reorder candidate evidence before final synthesis, showing both the potential and the coordination complexity of such designs [14].

The principal advantage of role separation is epistemic discipline. When one model both retrieves and generates, it may favor evidence that supports a fluent answer rather than evidence that challenges it. A separate verifier can be prompted to seek disconfirming evidence, identify missing sources, or flag contradictions. A separate reranker can be optimized for authority and diversity rather than merely semantic similarity. A separate auditor can examine whether the final answer overstates certainty or relies on a source that is outdated, anonymous, or outside the relevant jurisdiction. In high-stakes domains, these

checks can transform retrieval-augmented generation from a plausible answer machine into a contestable evidence process. Role separation also supports modular governance: different roles can be assigned different access rights, logging requirements, and human escalation triggers. For example, a verifier may have permission to read sensitive internal documents but not to disclose them, while a synthesizer may be restricted to public sources.

The disadvantages are equally significant. Multi-role systems introduce coordination overhead, because roles must communicate through structured messages, shared memory, or intermediate artifacts. If the query interpreter mischaracterizes the user’s intent, downstream roles may retrieve the wrong evidence. If the reranker overweights a single authoritative source, the verifier may not see contradictory evidence. If the auditor lacks access to the full retrieval trace, it may approve an answer that is locally coherent but globally unsupported. Each additional role increases latency and cost, and the system may become difficult to debug because errors emerge from interactions rather than from a single component. There is also a risk of false reassurance: a multi-role pipeline may appear more rigorous than it is, especially if all roles are based on similar models with shared biases. Without independent evidence sources, adversarial testing, and transparent logging, role separation can become a form of ritual rather than genuine verification.

Architecturally, multi-role evidence-aware search requires a shared evidence bus, a provenance layer, and a role coordinator. The evidence bus stores retrieved passages, metadata, embeddings, and intermediate judgments so that roles can inspect and challenge one another’s work. The provenance layer records where each passage came from, when it was retrieved, under what licensing or access conditions, and how it was transformed. The role coordinator manages turn-taking, conflict resolution, and escalation. In some designs, the coordinator may be a simple workflow engine; in others, it may itself be a language model. The choice matters for governance, because a language model coordinator may make opaque decisions about which role speaks next or which evidence prevails. A more auditable design uses explicit policies for conflict resolution, such as requiring human review when authoritative sources disagree or when confidence falls below a threshold. This reflects a broader principle: the more consequential the search domain, the more the architecture should favor inspectable rules over emergent model behavior.

4. Evidence-Aware Search, Reranking, and Governance Infrastructure

Evidence-aware search depends on the quality and structure of the underlying evidence infrastructure. A language model cannot retrieve what is not indexed, and it cannot verify what it cannot access. Institutions therefore need to treat retrieval corpora as governed assets with documented scope, update cycles, inclusion criteria, exclusion criteria, and retention policies. A legal search system may need to distinguish statutes, regulations, case law, secondary sources, and persuasive authority. A medical system may need to distinguish clinical guidelines, systematic reviews, randomized trials, observational studies, and preprints. A public policy system may need to distinguish official records, stakeholder submissions, media reports, and social media commentary. These distinctions are not merely metadata; they shape how evidence is ranked, how conflicts are resolved, and how uncertainty is communicated. If the corpus is incomplete or biased, multi-role verification may simply reproduce that bias with more elaborate justification.

Long-context language models are often proposed as a solution to evidence integration, but research shows that models do not use long contexts uniformly. Performance can degrade when relevant information appears in the middle of a long input, a phenomenon that has

important implications for retrieval-augmented generation [15]. Models can also be distracted by irrelevant context, leading them to produce answers that are influenced by passages that should have been ignored [16]. These findings suggest that simply retrieving more passages is not enough. Reranking, filtering, and evidence selection remain essential. A multi-role architecture can address this by assigning a role to compress and prioritize evidence before synthesis, but compression itself introduces loss and potential bias. The system must therefore preserve links to original sources so that users and auditors can inspect what was removed or summarized.

The distinction between parametric and non-parametric memory further complicates evidence-aware search. Language models encode knowledge in their weights, but that knowledge can be outdated, incomplete, or difficult to trace [17]. Retrieval provides non-parametric memory that can be updated and cited, but it also introduces dependence on retrieval quality and index coverage. A multi-role system can route queries between parametric knowledge and external evidence, but it must decide when model memory is sufficient and when retrieval is required. This decision is not purely technical. In domains with strict evidentiary standards, the default should favor retrieval and citation, even if the model is confident. Calibration research suggests that language models can sometimes estimate their own uncertainty, but self-assessment is imperfect and should not be the sole basis for high-stakes decisions [18]. A governance-oriented design therefore uses confidence as a trigger for escalation, not as a substitute for evidence.

Reranking is a particularly consequential governance layer because it determines which evidence receives attention. A reranker optimized only for semantic relevance may promote popular but unreliable sources, or it may overlook minority perspectives that are relevant to fairness. A reranker optimized for authority may entrench institutional biases and marginalize community knowledge. A multi-role approach can make ranking criteria explicit by assigning separate roles to relevance, authority, recency, diversity, and contestability. However, these criteria often conflict. A recent document may be less authoritative than an older statute; a diverse source may be less reliable; an authoritative source may be inaccessible or paywalled. The system needs a transparent policy for how such trade-offs are resolved. In public-interest search, this policy should be subject to stakeholder review, because ranking is not a neutral technical operation but a normative act that shapes whose knowledge counts.

5. Deployment, Sustainability, and Operational Trade-offs

Deploying multi-role evidence-aware search involves significant operational trade-offs. Each additional language model role consumes compute, memory, and energy, and each additional retrieval step increases latency. In interactive search, users may tolerate some delay for higher accuracy, but the threshold varies by domain. A consumer product search may prioritize speed, while a regulatory compliance review may prioritize completeness and auditability. Multi-role systems can be scaled through smaller specialized models, caching, asynchronous verification, and tiered escalation. For example, a lightweight first-stage system may handle routine queries, while a full multi-role pipeline is reserved for ambiguous, high-risk, or contested cases. This tiered approach can improve sustainability by avoiding unnecessary computation, but it requires accurate routing. If the router sends difficult queries to the lightweight path, errors may go undetected. If it sends too many queries to the full pipeline, cost and energy use increase.

Sustainability is not only an environmental concern but also an institutional one. Large-scale retrieval indexes require continuous crawling, embedding, storage, and refresh cycles. Multi-

role inference requires repeated model calls, and verification may require adversarial retrieval that doubles or triples the evidence footprint. Over time, these costs must be borne by someone, and the distribution of costs affects who can deploy evidence-aware search. Well-resourced organizations may build proprietary evidence pipelines with premium sources, while public institutions and smaller actors may rely on less curated data. This could widen information asymmetries. Open standards for provenance, interoperable indexes, and shared evaluation benchmarks can help counter this tendency, but they also require investment and coordination. The governance question is whether evidence-aware search will become a public utility with transparent rules or a fragmented market of opaque assistants with unequal access.

Hallucination remains a central risk even in retrieval-augmented systems. Retrieval can reduce hallucination by grounding generation in external text, but it does not eliminate it. Models may misread sources, combine claims incorrectly, cite passages that do not support the conclusion, or invent citations when retrieved evidence is insufficient [19]. Multi-role verification can mitigate these risks, but it can also introduce new forms of error. A verifier may be fooled by a plausible but false source; an auditor may approve an answer because the citations are formatted correctly rather than because they are substantively adequate. This is why evaluation must go beyond answer accuracy to include citation precision, evidence sufficiency, contradiction handling, and uncertainty communication. In high-stakes settings, the system should be designed to abstain or escalate when evidence is weak, rather than forcing a fluent answer.

Broader risks associated with foundation models also apply to multi-role evidence-aware search. These models can concentrate power, reproduce social biases, and create dependencies that are difficult to audit or reverse [20]. Multi-role architectures can distribute some of this power by separating roles and enabling external oversight, but they can also obscure responsibility. If the final answer is produced by a synthesizer but shaped by a reranker, verifier, and coordinator, who is accountable for an error? The answer cannot be found in the model alone. It must be located in the socio-technical system: the organization that selected the corpus, the engineers who designed the prompts, the auditors who approved the pipeline, and the decision-makers who relied on the output. Governance frameworks should therefore require clear responsibility allocation, version control for prompts and models, and documentation of known limitations.

6. Robustness, Fairness, and Policy Implications

Robustness in multi-role evidence-aware search means more than resistance to adversarial inputs. It includes resilience to distribution shift, source manipulation, index corruption, and infrastructure failure. An evidence-aware system may be attacked through the retrieval corpus, for example by injecting documents designed to rank highly for certain queries. If the reranker and verifier rely on similar signals, the attack may pass through multiple roles. Defenses include source diversity, provenance verification, anomaly detection, and human review for sensitive topics. Robustness also requires graceful degradation. If the primary index is unavailable, the system should not silently fall back to parametric memory without disclosure. It should inform the user that evidence retrieval has failed and adjust its confidence accordingly.

Fairness in evidence-aware search is complex because retrieval and ranking can disadvantage groups even when the model itself is not explicitly discriminatory. A corpus may underrepresent certain languages, regions, or perspectives. A reranker may favor documents

written in dominant dialects or published by well-indexed institutions. A synthesizer may summarize marginalized viewpoints as less credible. Multi-role architectures can improve fairness by assigning roles to test for representational gaps and to seek diverse evidence. However, diversity is not a simple fix. Adding low-quality or extremist sources in the name of balance can harm users. The goal should be principled pluralism: including relevant perspectives while maintaining evidentiary standards and clearly distinguishing empirical claims from normative positions. This requires domain expertise, not just model prompting.

Policy implications follow from these concerns. Governments and institutions may need to establish standards for evidence-aware search in public services, including transparency about sources, ranking criteria, and conflict resolution. Procurement rules could require vendors to provide audit logs, model cards, and data statements. Liability rules may need to clarify responsibility when an AI-assisted search contributes to harm. Data protection rules may restrict the use of personal data in retrieval indexes and require deletion mechanisms. Competition policy may address market concentration in foundation models and search infrastructure. These are not peripheral issues; they shape whether evidence-aware search can be trusted in democratic and institutional contexts. A multi-role system that is technically sophisticated but institutionally opaque may fail the legitimacy test even if its benchmark scores are high.

7. Evaluation, Auditability, and Institutional Accountability

Evaluation of multi-role evidence-aware search must be multidimensional. Traditional retrieval metrics such as precision and recall are necessary but insufficient. The system must also be evaluated on citation correctness, evidence sufficiency, robustness to adversarial documents, fairness across queries and populations, latency, cost, and user comprehension. Human evaluation remains important because some qualities, such as whether an answer appropriately communicates uncertainty, are difficult to measure automatically. Evaluation should also be continuous rather than one-time. Retrieval corpora change, models are updated, and social contexts shift. A system that performed well at deployment may degrade as sources are removed, new regulations are passed, or language use evolves. Continuous monitoring and periodic re-audit are therefore essential.

Auditability requires that the system preserve a trace of how an answer was produced. This trace should include the original query, any reformulations, retrieved passages, ranking scores, role interactions, verification steps, and the final answer. It should also record which sources were considered but rejected and why. Such traces enable internal review, external audit, and user contestation. However, auditability raises privacy and security concerns. Detailed traces may contain sensitive information or reveal proprietary ranking policies. A balanced approach uses role-based access, redaction, and independent oversight. In public-interest applications, some level of transparency should be mandatory, especially when the system influences access to benefits, legal outcomes, or medical care.

Institutional accountability means that someone must be responsible for the system's behavior. Multi-role architectures can create a diffusion of responsibility, where each component seems to be doing its part but no one owns the overall outcome. To counter this, organizations should designate accountable roles for corpus governance, model selection, prompt design, evaluation, incident response, and user redress. They should also establish escalation paths to human experts when the system encounters contested evidence, ethical dilemmas, or high-stakes uncertainty. Accountability is not a technical feature that can be added at the end; it must be designed into the workflow. This includes deciding in advance what kinds of errors

are tolerable, what kinds require disclosure, and what kinds should trigger suspension of the service.

8. Conclusion

Retrieval-augmented generation with multi-role large language models offers a promising path toward evidence-aware search that is more transparent, verifiable, and contestable than purely parametric generation. By separating query interpretation, retrieval, reranking, verification, synthesis, and audit, such systems can create useful checks and balances and can make evidence use more explicit. However, this architectural choice is not a cure-all. It introduces new costs, coordination failures, and governance burdens. Multi-role systems can amplify biases in the evidence base, create false confidence through redundant model roles, and obscure responsibility when errors occur. The value of role separation depends on the quality of the underlying infrastructure, the independence of verification, the transparency of ranking policies, and the presence of human oversight.

The paper has argued that evidence-aware search should be understood as infrastructure rather than as a standalone application. Its design must account for corpus governance, provenance, versioning, observability, escalation, sustainability, fairness, and policy. Deployment should be tiered according to risk, with lightweight paths for routine queries and full multi-role verification for consequential or contested cases. Evaluation should extend beyond accuracy to citation precision, evidence sufficiency, robustness, fairness, and user comprehension. Accountability should be assigned to identifiable institutional roles, and audit trails should enable contestation. If these conditions are met, multi-role retrieval-augmented generation can support more trustworthy search in complex domains. If they are neglected, it may simply produce more elaborate justifications for the same underlying epistemic vulnerabilities.

References

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
2. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 6769–6781). Association for Computational Linguistics.
3. Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 874–880). Association for Computational Linguistics.
4. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., ... Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 2206–2240). PMLR.
5. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 3929–3938). PMLR.

6. Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 39–48). Association for Computing Machinery.
7. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1.
8. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
9. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.
10. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the 11th International Conference on Learning Representations*.
11. Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.
12. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
13. Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K.-W., & Lim, E.-P. (2023). Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (pp. 2609–2634). Association for Computational Linguistics.
14. Jin, C., Peng, H., Zhang, A., Chen, N., Zhao, J., Xie, X., ... & Metaxas, D. N. (2025, May). Rankflow: A multi-role collaborative reranking workflow utilizing large language models. In *Companion Proceedings of the ACM on Web Conference 2025* (pp. 2484–2493).
15. Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157–173.
16. Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E., Schärli, N., & Zhou, D. (2023). Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 31210–31227). PMLR.
17. Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2023). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (pp. 9802–9822). Association for Computational Linguistics.
18. Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Dodds, Z. H., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A.,

- Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., ... Kaplan, J. (2022). Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221.
19. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38.
20. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.