

Adaptive Query Routing for Domain-Specific Large Language Models with Retrieval-Augmented Generation

Juni Bjorklund

School of Computing, Clemson University, Clemson, SC, USA.
bjorklund958@clemson.edu

Loni Hellwig

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
loni479@uab.edu

Clíodhna J. Cloone

Department of Computer Science, University of Central Florida, Orlando, FL, USA.
cliodhnaclloone881@ucf.edu

Abstract

Large language models are increasingly deployed in domain-specific environments where accuracy, provenance, latency, cost, and regulatory accountability must be balanced simultaneously. Retrieval-augmented generation has become a central architectural pattern for grounding such models in external knowledge, yet most implementations assume a single retrieval pathway or a fixed model configuration. This paper examines adaptive query routing as a system-level capability for domain-specific large language models that use retrieval-augmented generation. Adaptive routing dynamically selects among models, indexes, retrieval strategies, and generation pathways according to query intent, domain constraints, operational conditions, and governance requirements. The analysis treats routing not merely as an optimization technique but as an infrastructural and policy-bearing layer that shapes fairness, robustness, sustainability, and institutional trust. It discusses structural trade-offs among accuracy, interpretability, latency, energy use, and maintainability, and it considers how heterogeneous retrieval sources can be fused without collapsing domain distinctions. The paper also addresses deployment resilience, evaluation under distribution shift, auditability, and cross-domain comparisons in medicine, law, engineering, and public administration. It argues that adaptive query routing should be designed as a governed, observable, and reversible component of socio-technical infrastructure rather than as an opaque accelerator. The conclusion emphasizes that domain-specific retrieval-augmented systems require continuous operational learning, transparent routing rationales, and policy-aware design if they are to remain reliable, fair, and sustainable at scale.

Keywords

adaptive query routing; retrieval-augmented generation; domain-specific language models; system architecture; governance; infrastructure; fairness; robustness; sustainability.

1. Introduction

Domain-specific large language models are moving from experimental prototypes into operational infrastructures where their outputs influence professional judgment, resource allocation, and public communication. In such settings, a general-purpose model augmented with retrieval is rarely sufficient, because domain knowledge is distributed across heterogeneous sources, access controls, formatting conventions, and temporal validity regimes. Retrieval-augmented generation has emerged as a practical response by coupling parametric generation with non-parametric retrieval [1]. Dense passage retrieval and related methods enable semantic matching beyond exact term overlap [2], while fusion-in-decoder and retrieval-enhanced pretraining approaches show that external evidence can be integrated at different stages of the model pipeline [3]. These developments have made retrieval a first-class architectural concern rather than an optional add-on.

However, as retrieval-augmented systems grow more complex, a new problem becomes visible: not every query should be routed through the same retrieval index, the same generator, or the same reasoning pathway. A domain-specific system may have access to clinical guidelines, legal statutes, engineering standards, internal incident reports, and public web sources, each with different reliability, freshness, and authorization requirements. Fixed routing policies therefore risk either over-retrieving from irrelevant sources or under-retrieving from sources that are necessary for high-stakes decisions. Adaptive query routing responds to this challenge by treating routing as a dynamic decision about which knowledge sources, retrieval methods, and generation models are appropriate for a given query and context [6]. The routing layer thus becomes a point where technical design and institutional policy intersect.

This paper argues that adaptive query routing for domain-specific retrieval-augmented generation should be understood as a socio-technical infrastructure problem. The central tensions are not only about retrieval accuracy but also about governance, fairness, sustainability, and robustness. Foundation-model research has shown that scale and generality introduce systemic risks and dependencies [13], while critical scholarship has warned that large models can reproduce and amplify historical inequities [15]. Governance frameworks such as the NIST Artificial Intelligence Risk Management Framework provide a vocabulary for managing these risks through mapping, measurement, and monitoring [20]. The remainder of this paper develops a system-level account of adaptive routing, examining its architecture, structural trade-offs, deployment implications, and policy consequences.

2. Retrieval-Augmented Generation as a Systems Architecture

Retrieval-augmented generation is often described as a model technique, but in operational systems it is better understood as an architecture composed of separable services. A query enters an orchestration layer, is transformed into one or more retrieval representations, is matched against indexes, and then supplies evidence to a generator that produces a response. The foundational RAG design demonstrated that retrieval can improve knowledge-intensive tasks without full retraining [1]. Dense passage retrieval provided a scalable semantic matching mechanism [2], while later work showed that generation can be conditioned on multiple retrieved passages through fusion [3]. Retrieval-enhanced pretraining and retrieval-augmented language model pretraining further blurred the boundary between parametric memory and non-parametric access [4], [5]. In practice, these components are not monolithic; they are distributed across vector databases, search engines, caches, authorization services, and model endpoints.

Within this architecture, adaptive query routing introduces a control plane that decides which retrieval and generation resources should be activated. A simple query about a public regulation may be routed to a sparse keyword index because exact statutory language matters [9]. A complex clinical question may require dense retrieval over guideline summaries, followed by a domain-tuned generator [2]. A multi-source engineering query may need heterogeneous retrieval fusion, where semantic spaces and attention-based fusion mechanisms align evidence from technical manuals, safety reports, and maintenance logs [7]. The routing layer must therefore reason about source type, query intent, risk level, user role, temporal constraints, and service health. This is not merely a classification task; it is a resource-allocation and accountability mechanism.

The architectural consequences are significant. First, routing creates dependencies among components that may evolve at different speeds. A change in embedding model can alter retrieval behavior across many domains, while a change in authorization policy can invalidate previously cached evidence. Second, routing affects observability. If a response is generated from multiple sources, the system must preserve provenance for each retrieved item and for the routing decision itself. Third, routing influences maintainability. As machine learning systems accumulate hidden technical debt, opaque routing rules can become a source of silent failure [16]. A well-designed architecture therefore treats routing as an explicit, versioned, and auditable component rather than as an embedded heuristic.

3. Domain-Specific Constraints and Query Routing

Domain-specific environments impose constraints that general retrieval benchmarks do not fully capture. In medicine, terminology is dense, evidence hierarchies matter, and patient-specific context may be protected by privacy regulation. In law, jurisdiction, precedent, and temporal validity determine relevance, so a semantically similar passage may be legally inapplicable. In engineering, standards and safety cases may require exact revision numbers and traceable approval histories. Heterogeneous benchmarks such as BEIR show that retrieval models often fail to generalize across domains without adaptation [10]. This means that routing cannot rely solely on semantic similarity; it must incorporate domain metadata, source authority, and policy constraints.

Datasheets for datasets and model cards for model reporting offer useful precedents for documenting the assumptions and limitations of components that routing may select [18], [19]. If a routing layer chooses a particular index or generator, users and auditors should be able to understand what data the component was trained on, what evaluation it passed, and what known biases or gaps it carries. Without such documentation, domain-specific routing becomes difficult to justify in regulated settings. The routing decision also interacts with access control. A clinician may be authorized to retrieve patient records, while a billing analyst may not. A legal researcher may access public case law, while an internal compliance officer may access privileged memoranda. Adaptive routing must therefore be policy-aware rather than purely performance-driven.

The required reference on unified semantic space and attention-based fusion for multi-source heterogeneous retrieval-augmented generation illustrates the growing technical sophistication of this problem [7]. When sources differ in modality, vocabulary, and reliability, routing must decide not only which sources to consult but also how to align their representations without erasing meaningful distinctions. A unified semantic space can improve cross-source retrieval, but it can also obscure provenance if fusion is too aggressive. The design challenge is to preserve source identity while enabling comparative reasoning. This is especially important in

domains where contradictory evidence must be surfaced rather than averaged away. Governance frameworks such as the NIST AI Risk Management Framework encourage mapping such dependencies and monitoring them over time [20].

4. Adaptive Routing Mechanisms and Structural Trade-offs

Adaptive routing mechanisms can be organized around several dimensions: query understanding, resource selection, evidence fusion, and response synthesis. Query understanding may involve intent classification, domain detection, risk scoring, and temporal normalization. Resource selection may choose among sparse indexes, dense indexes, graph stores, relational databases, or external application programming interfaces. Evidence fusion may combine passages from multiple sources using rank aggregation, attention, or learned re-ranking [3], [8]. Response synthesis may select a general generator, a domain-tuned generator, or a constrained decoder. Each decision introduces trade-offs among accuracy, latency, cost, interpretability, and energy consumption.

Sparse retrieval remains valuable because it is efficient, interpretable, and strong for exact terminology [9]. Dense retrieval improves semantic recall but can struggle with rare identifiers and out-of-domain queries [2]. Late-interaction models such as ColBERT offer a middle ground by preserving token-level matching while enabling contextual representations [8]. Approximate nearest neighbor indexes such as hierarchical navigable small world graphs and GPU-accelerated similarity search make large-scale dense retrieval feasible, but they introduce recall-latency trade-offs and operational complexity [11], [12]. Adaptive routing must therefore decide when exactness is worth the cost of slower retrieval and when approximate recall is sufficient. These decisions cannot be made once and forgotten; they must adapt to workload, service level objectives, and domain risk.

The structural trade-offs are not only technical. Routing can improve fairness by directing queries to sources that better represent underrepresented populations, but it can also worsen fairness if historical biases are encoded in source authority rankings. Research on the ethical and social risks of language models highlights the danger of opaque systems that allocate benefits and harms unevenly [14]. Routing can improve robustness by avoiding a single failing index, but it can also create new failure modes when fallback paths are poorly tested. Work on generalization challenges the assumption that models will behave predictably on unseen data, reinforcing the need for routing policies that are conservative under uncertainty [17]. Technical debt accumulates when routing logic is spread across configuration files, prompt templates, and ad hoc scripts rather than managed as a coherent subsystem [16]. A mature design therefore treats routing as a governed interface with explicit contracts, monitoring, and rollback procedures.

5. Governance, Fairness, and Policy Implications

Governance of adaptive query routing must address both procedural and substantive questions. Procedurally, who defines routing policies, who can change them, and how are changes recorded? Substantively, what counts as a legitimate source, how are conflicts between sources resolved, and what recourse exists when a user receives an inaccurate or harmful response? These questions are especially urgent in public-sector and regulated industries where decisions may affect rights, safety, or access to services. The NIST AI Risk Management Framework provides a structured approach to mapping, measuring, managing, and monitoring risks, which can be extended to routing decisions [20]. Model cards and

datasheets provide documentation practices that can be adapted to routing components [18], [19].

Fairness implications arise because routing determines which knowledge is visible and which models are used. If a routing policy consistently sends queries from certain communities to lower-quality indexes or older models, it can produce disparate outcomes even if each component is individually unbiased. Ethical risk research emphasizes that harms often emerge from system interactions rather than isolated model errors [14]. Critical analysis of large language models further warns that scale can concentrate power and marginalize underrepresented perspectives [15]. Adaptive routing should therefore include fairness testing across user groups, domains, and query types. It should also provide explanations that are meaningful to non-engineers, such as stating that a response relied on a particular policy document, clinical guideline, or internal record.

Policy implications extend beyond internal governance. Regulators may require audit trails that show which sources informed a decision and which model generated a response. In the European context, risk-based artificial intelligence regulation places obligations on high-risk systems that may apply to domain-specific retrieval-augmented applications. Even where formal regulation is absent, professional standards and liability regimes may demand traceability. Routing policies should therefore be designed for contestability. A user or auditor should be able to challenge a routing decision, request a re-run with different sources, and receive a clear record of what changed. This kind of reversibility is a hallmark of accountable infrastructure. It also aligns with foundation-model research that calls for greater transparency about capabilities, limitations, and downstream dependencies [13].

6. Deployment, Sustainability, and Infrastructure Resilience

Deployment of adaptive routing introduces infrastructure challenges that are often underestimated. Retrieval indexes must be updated as domain knowledge changes, embeddings must be regenerated when models are upgraded, and caches must be invalidated when policies change. GPU-accelerated similarity search and approximate nearest neighbor structures can reduce latency, but they also create specialized operational dependencies [11], [12]. When routing selects among multiple generators, each with different memory and compute profiles, the serving infrastructure must manage heterogeneous workloads without violating service level objectives. This is a systems optimization problem, but it is also a sustainability problem. Energy consumption, carbon intensity, and hardware utilization become relevant because routing can either reduce unnecessary computation or amplify it through redundant fallback paths.

Sustainability is not merely an environmental concern; it is also an economic and institutional concern. A routing policy that always selects the largest model and the most comprehensive retrieval set may achieve high accuracy on benchmarks but prove unsustainable in daily operations. Conversely, an overly aggressive cost-saving policy may route high-stakes queries to weak resources and create downstream costs in rework, liability, or loss of trust. The foundational RAG architecture showed that retrieval can reduce the need for purely parametric memorization [1], and retrieval-enhanced pretraining suggests that external memory can complement model scale [4]. Adaptive routing can extend this logic by matching query difficulty to resource intensity. However, such matching requires careful measurement of both quality and cost, including human oversight costs.

Resilience is another deployment concern. Domain-specific systems may face outages in external databases, degraded network conditions, or adversarial queries designed to exhaust retrieval resources. Routing can improve resilience by providing fallback pathways, but fallbacks must be tested and governed. If a fallback uses a less authoritative source, the system should disclose that shift. If a fallback uses a smaller model, it should adjust confidence estimates or recommend human review. Technical debt in machine learning systems often arises from hidden dependencies and undeclared feedback loops [16]. Adaptive routing should therefore be instrumented with health checks, circuit breakers, and provenance-preserving logs. It should also be evaluated under stress conditions, not only on static benchmarks.

7. Evaluation, Robustness, and Operational Learning

Evaluation of adaptive routing cannot rely solely on end-to-end answer accuracy. It must assess routing precision, source selection quality, latency distribution, cost per query, fairness across groups, and robustness to distribution shift. Heterogeneous retrieval benchmarks provide a starting point for measuring cross-domain retrieval performance [10], while late-interaction and dense retrieval studies offer methods for comparing semantic matching strategies [2], [8]. Zero-shot dense retrieval approaches show that models can generalize without relevance labels, but their behavior in specialized domains may still require validation [6]. In operational settings, evaluation should include counterfactual analysis: what would have happened if a different route had been chosen? This requires logging routing decisions and outcomes in a way that supports retrospective audit.

Robustness is closely tied to uncertainty. When a query is ambiguous, adversarial, or outside the domain, the routing layer should be able to detect this and respond conservatively. It may retrieve from multiple sources, lower confidence, or escalate to a human. Research on generalization suggests that high performance on training distributions does not guarantee reliable behavior on rare or shifted inputs [17]. For domain-specific systems, rare cases are often the most important: an unusual drug interaction, a novel legal question, or an unprecedented engineering failure. Routing policies should therefore include out-of-distribution detection and explicit escalation pathways. They should also be tested against scenarios where sources conflict, where metadata is missing, and where authorization boundaries are ambiguous.

Operational learning should be continuous but controlled. User feedback, incident reports, and audit findings can inform routing updates, but uncontrolled updates can introduce instability. A governed routing lifecycle should include versioning, canary deployment, rollback, and documentation. Model cards and datasheets provide templates for documenting changes and limitations [18], [19]. The required heterogeneous retrieval fusion framework illustrates that multi-source integration is technically possible, but it also implies that evaluation must account for fusion effects [7]. If fusion improves average accuracy while obscuring source-level failures, the system may appear robust while hiding local weaknesses. Operational learning should therefore preserve granular metrics and allow domain experts to inspect routing paths.

8. Cross-Domain Comparisons and Forward-Looking Perspectives

Cross-domain comparison reveals both commonalities and differences in adaptive routing. In medicine, routing must prioritize evidence quality, privacy, and temporal validity. In law, routing must respect jurisdiction, precedent, and privilege. In engineering, routing must track

standards, revisions, and safety cases. In public administration, routing must balance transparency, accessibility, and due process. Despite these differences, all domains benefit from explicit source authority models, audit trails, and fallback policies. Retrieval-augmented generation provides a general pattern [1], but domain-specific routing is what makes the pattern operational. Fusion-in-decoder and retrieval-enhanced pretraining show that integration can occur at multiple levels [3], [4], while REALM demonstrated the value of retrieval during pretraining [5]. The next generation of systems will likely combine these insights with policy-aware orchestration.

Forward-looking perspectives should consider interoperability and federated retrieval. Domain-specific systems may need to query external knowledge bases without centralizing sensitive data. Routing could support federated search across institutions, with local authorization and provenance preservation. This would require shared semantics for source metadata, risk levels, and audit events. Foundation-model research highlights the importance of understanding systemic dependencies [13], and ethical risk research emphasizes the need for inclusive design [14]. Governance frameworks such as the NIST AI Risk Management Framework can provide a common vocabulary for risk mapping and monitoring [20]. Datasheets and model cards can document components, while routing logs can document decisions [18], [19]. Together, these practices can make adaptive routing more accountable.

A final forward-looking concern is the relationship between routing and human expertise. In high-stakes domains, routing should not replace professional judgment; it should structure access to evidence and expose uncertainty. The system should help experts see which sources were used, which were excluded, and why. It should also allow experts to override routing decisions and record the rationale. This human-in-the-loop perspective aligns with the broader insight that socio-technical systems succeed when technical optimization is embedded in institutional practices. Adaptive routing for domain-specific retrieval-augmented generation is therefore not simply a matter of improving search. It is a matter of designing infrastructure that supports reliable, fair, and contestable knowledge work.

9. Conclusion

Adaptive query routing is an emerging systems capability for domain-specific large language models that use retrieval-augmented generation. It decides which queries go to which models, indexes, and evidence sources, and it therefore shapes accuracy, latency, cost, fairness, and accountability. This paper has argued that routing should be treated as a governed architectural layer rather than an opaque optimization. Retrieval-augmented generation provides the foundation [1], and heterogeneous fusion frameworks show how multiple sources can be aligned [7]. However, technical integration alone is insufficient. Governance frameworks, documentation practices, fairness testing, and resilience engineering are necessary to ensure that routing supports institutional trust [20].

The main trade-offs are structural. Sparse retrieval offers exactness and interpretability, while dense retrieval offers semantic flexibility [2], [9]. Large models may improve quality but increase cost and energy use, while smaller models may be more sustainable but less capable. Routing policies must navigate these tensions with explicit criteria and continuous evaluation. They must also remain robust under distribution shift and adversarial conditions [17]. As domain-specific systems scale, adaptive routing will become a point of policy contestation as well as engineering design. The most credible path forward is to build routing layers that are observable, reversible, documented, and open to expert override. Such layers can help

retrieval-augmented generation serve domain-specific needs without sacrificing fairness, sustainability, or accountability.

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. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 6769–6781.
3. Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 874–880.
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. *Proceedings of the 39th International Conference on Machine Learning*, 162, 2206–2240.
5. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). REALM: Retrieval-augmented language model pre-training. *Proceedings of the 37th International Conference on Machine Learning*, 119, 3929–3938.
6. Gao, L., Ma, X., Lin, J., & Callan, J. (2023). Precise zero-shot dense retrieval without relevance labels. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 1762–1777.
7. Deng, Q., Liu, S., Qiu, H., Long, N., Lin, W., & Chen, S. (2026, April). Unified Semantic Space and Attention-Based Fusion Framework for Multi-Source Heterogeneous Retrieval-Augmented Generation. In *2026 6th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)* (pp. 591-597). IEEE.
8. Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 39–48.
9. Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389.
10. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1.
11. Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4), 824–836.
12. Johnson, J., Douze, M., & Jégou, H. (2021). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535–547.

13. 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. <https://arxiv.org/abs/2108.07258>
14. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., ... Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv. <https://arxiv.org/abs/2112.04359>
15. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.
16. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems, 28, 2503–2511.
17. Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning requires rethinking generalization. Communications of the ACM, 64(3), 107–115.
18. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, 220–229.
19. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92.
20. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0). NIST. <https://doi.org/10.6028/NIST.AI.100-1>