

Adaptive LLM Reranking for Multimodal Information Retrieval

Salvi Storladi

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis,
OR, USA.

sstorladi@oregonstate.edu

Cakrawala Prayoga

Department of Computer Science, University of Houston, Houston, TX, USA.

cakrawala.prayoga648@uh.edu

Febe Carro

School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.

febework@uc.edu

Breandan Shortt

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL,
USA.

breandan.work@uab.edu

Abstract

Multimodal information retrieval is increasingly expected to operate across text, images, audio, video, structured records, and sensor streams, often within the same query session. Conventional retrieval cascades rely on modality-specific encoders, approximate nearest-neighbor search, and lightweight rerankers, but these pipelines struggle when intent is ambiguous, modality relevance is uneven, or context shifts rapidly. Large language models offer a promising reranking substrate because they can reason over heterogeneous evidence, follow nuanced instructions, and synthesize relevance judgments across modalities. Yet their integration into retrieval infrastructure raises difficult questions about latency, cost, energy, privacy, fairness, robustness, and governance. This paper examines adaptive LLM reranking as a system-level capability rather than a isolated model upgrade. It develops an architectural account of modality-aware recall, adaptive orchestration, cross-modal evidence fusion, and policy-constrained ranking. It then analyzes data and representation governance, deployment and sustainability trade-offs, robustness and fairness evaluation, and organizational implications for search platforms, enterprises, healthcare, education, and public-sector systems. The central argument is that adaptive reranking must be treated as socio-technical infrastructure. Quality gains are real, but they are contingent on explicit trade-off management, continuous evaluation, transparent reporting, and governance mechanisms that connect model behavior to institutional accountability.

Keywords

multimodal retrieval; large language models; adaptive reranking; information retrieval; governance; sustainability; robustness; fairness; socio-technical systems.

1. Introduction

Multimodal information retrieval has moved from a specialized research problem to a core expectation of modern digital infrastructure. Users now search across documents, images, tables, video, audio, geospatial records, and conversational context within a single interaction. A query may begin as text, become an image upload, and then evolve through follow-up questions that refer to earlier results. Classical retrieval architectures separate these modalities into distinct pipelines, which can be efficient but often weak at preserving cross-modal semantics. Transformer-based models have enabled deeper contextual representation [1], and bidirectional pretraining improved text understanding [2]. Vision-language pretraining further aligned images and text at scale [3], while later systems connected frozen image encoders with large language models for flexible multimodal reasoning [4]. Instruction-tuned vision-language models expanded the range of tasks that can be expressed in natural language [5]. These developments create the technical conditions for adaptive reranking, but they also make system design more complex.

Reranking has traditionally served as a precision-oriented stage after candidate generation. Early neural rerankers demonstrated that contextual language models could substantially improve passage relevance [6], and late-interaction architectures offered a middle ground between efficiency and effectiveness [7]. Dense retrieval methods changed recall by representing queries and documents in shared embedding spaces [8], while retrieval-augmented generation connected retrieval outputs to generative reasoning [9]. The next step is adaptive reranking, in which the system decides when to invoke a large language model, which modalities to compare, how much evidence to expose, and what constraints should govern the final ordering. A multi-role collaborative reranking workflow illustrates how multiple model roles can be coordinated rather than relying on a single monolithic judgment [10]. This paper treats that coordination as an infrastructure problem.

The central claim is that adaptive LLM reranking cannot be evaluated only through ranking accuracy. It must be assessed as a socio-technical system that mediates access to knowledge, allocates computational resources, encodes institutional priorities, and creates new dependencies. The system-level challenge is to balance relevance, latency, cost, privacy, fairness, and resilience. A reranker that performs well in a benchmark but fails under domain shift, excludes minority languages, or consumes unsustainable energy is not a successful deployment. The discussion below therefore moves from architecture to governance, from model behavior to institutional accountability, and from short-term performance to long-term maintainability.

2. Foundations and Evolution of Multimodal Retrieval and Reranking

The evolution of retrieval systems can be understood as a sequence of recall and precision trade-offs. Sparse lexical retrieval remains valuable for exact matching and interpretability, while dense retrieval captures semantic similarity that lexical methods miss [8]. Hybrid systems combine both signals, often using learned fusion or reciprocal ranking strategies. Multimodal retrieval adds another layer because text, images, audio, and video do not share a natural common representation. Vision-language models such as those trained on image-text pairs create shared embedding spaces that enable cross-modal search [3]. More recent multimodal models use language models as reasoning engines over visual tokens, enabling instruction-following and conversational search [4][5]. These models can answer questions about images, but retrieval systems must still decide which candidates deserve deep inspection.

Reranking is the stage where expensive reasoning is most defensible. A first-stage retriever may return hundreds or thousands of candidates; a reranker can focus on the top subset and apply richer context. Contextual rerankers improved passage ranking by jointly encoding query and document [6], and late-interaction methods reduced the cost of storing and comparing token-level representations [7]. Large language models extend this logic by generating relevance judgments, explaining decisions, or producing listwise orderings. Listwise reranking with open-source models has shown that instruction-tuned systems can approximate proprietary behavior under controlled conditions [12]. However, listwise approaches are sensitive to prompt design, candidate ordering, and context length. They also introduce non-determinism, which complicates reproducibility and service-level guarantees.

Multimodal retrieval-augmented generation connects reranking to downstream synthesis. Systems such as MuRAG retrieve images and text to answer open questions [13], while retrieval-augmented multimodal language modeling integrates external evidence into generative pretraining or inference [14]. Vision-language assistants such as MiniGPT-4 demonstrate the promise of combining visual encoders with large language models for richer interaction [15]. In these systems, reranking is not merely a filter; it determines the evidence that will condition generation. Poor reranking can therefore propagate errors into fluent but unsupported answers. The retrieval infrastructure must preserve provenance, uncertainty, and modality balance so that downstream generation remains grounded.

3. Adaptive LLM Reranking Architecture

An adaptive LLM reranking architecture begins with query understanding. The system must infer intent, modality needs, domain constraints, and risk level. A medical query about a rash may require image retrieval, dermatology references, and privacy controls, while a news query may prioritize recency and source diversity. The architecture then routes the request through modality-specific encoders, sparse and dense retrievers, and metadata filters. Candidate generation should not be a single monolithic index; it is better understood as a federation of indexes with different update rates, access controls, and quality profiles. The reranker receives a candidate set and must decide whether to invoke a large model, a smaller distilled model, or a rule-based fallback.

The adaptive component concerns resource allocation. Not every query needs a large language model. Simple navigational queries may be satisfied by lexical matching, while complex comparative queries benefit from multi-step reasoning. A multi-role collaborative workflow can separate query analysis, evidence selection, relevance judgment, and final ordering [10]. Other work has examined large language models as reranking agents, showing both potential and instability when prompts or candidate orders change [11]. Open-source listwise rerankers provide more control but may lag proprietary systems in generalization [12]. The architectural goal is not to maximize model size but to match reasoning depth to query difficulty, risk, and service constraints.

Cross-modal evidence fusion is another core design problem. Text, images, tables, and audio have different reliability, granularity, and cost profiles. A product search may rely on image similarity, textual descriptions, and structured attributes. A video search may need transcript matching, visual object recognition, and temporal alignment. The reranker must weigh these signals without collapsing them into a single opaque score. It should expose enough structure for auditing and debugging. The system should also support graceful degradation: if an image encoder is unavailable, text retrieval should continue; if a large model times out, a smaller

reranker should preserve basic service. These requirements make adaptive reranking a reliability engineering problem as much as a machine learning problem.

4. Data, Representation, and Semantic Governance

Multimodal reranking depends on data governance because the meaning of a result is shaped by collection, annotation, and representation choices. Image-text pairs may reflect cultural biases, geographic imbalances, or historical inequities. Speech and video data introduce consent, privacy, and surveillance risks. Enterprise documents may contain proprietary information that cannot be sent to external model providers. Healthcare images and records are subject to strict legal and ethical controls. Data governance must therefore address provenance, consent, retention, access, and permitted use across every modality. Datasheets for datasets and model cards provide structured reporting practices that can make these choices visible [20]. Without such documentation, adaptive reranking becomes difficult to audit and easy to misuse.

Representation governance is equally important. Embedding spaces are not neutral maps of meaning; they encode the distributional patterns of their training data. A model trained predominantly on English text and Western imagery may perform poorly on other languages, dialects, or visual cultures. Multimodal encoders can also inherit biases from captioning data, which often overrepresents certain objects, activities, and social groups. Foundation model research has highlighted the broad risks of scale, opacity, and downstream reuse [17]. Critiques of large language models have warned that scale alone does not resolve social bias, environmental cost, or concentration of power [16]. Adaptive reranking should therefore include representation audits, language coverage tests, and modality-specific fairness checks.

Semantic governance also concerns the relationship between retrieval and truth. A reranker may prioritize documents that are fluent, popular, or well-optimized rather than accurate. In multimodal settings, misleading images, synthetic media, and out-of-context clips can be ranked highly if engagement signals dominate. Retrieval-augmented generation can amplify these problems because generated answers may cite ranked evidence as if it were verified [9]. Governance mechanisms should distinguish between relevance, credibility, and safety. They should allow policy rules to override learned rankings when legal, ethical, or public-interest constraints apply. This requires an architecture in which ranking is not a final black box but an accountable stage within a broader information quality process.

5. Infrastructure, Deployment, and Sustainability

Adaptive LLM reranking is computationally expensive. Large models require accelerators, high-bandwidth memory, and significant energy for training and inference. Even if reranking is applied to a small candidate set, the aggregate cost across millions of queries can be substantial. Sustainability research in natural language processing has drawn attention to energy consumption and carbon emissions [18], while Green AI advocates for efficiency as a first-class evaluation criterion [19]. Deployment strategies must therefore consider model size, quantization, distillation, caching, batching, and geographic load balancing. A smaller reranker that is slightly less accurate may be preferable if it reduces latency and energy while maintaining acceptable quality.

Infrastructure choices also affect latency and user experience. Cloud-based large models offer elasticity and access to advanced capabilities, but they introduce network delay, vendor dependency, and data transfer risks. Edge deployment can reduce latency and improve privacy, but edge devices have limited memory and power. Hybrid architectures can route

sensitive or simple queries locally and complex or low-risk queries to cloud services. Caching is valuable for repeated queries and popular content, but it can reinforce popularity bias and stale results. The system must define cache invalidation policies that respect freshness, personalization, and deletion rights. These are not merely engineering details; they determine who can use the system, under what conditions, and with what environmental footprint.

Maintainability is another sustainability dimension. Machine learning systems accumulate hidden technical debt through data dependencies, configuration drift, and feedback loops [24]. An adaptive reranker that depends on external models, changing APIs, and evolving prompts is especially vulnerable. Versioning must cover encoders, indexes, prompts, policy rules, and evaluation datasets. Monitoring should detect performance degradation, bias drift, and unexpected cost spikes. Organizations should be able to roll back changes and reproduce past rankings for audits. Sustainability therefore includes operational resilience and institutional capacity, not only energy efficiency. A system that cannot be maintained will eventually fail its users regardless of benchmark performance.

6. Robustness, Fairness, and Evaluation

Robustness in multimodal reranking requires testing under distribution shift, modality failure, adversarial manipulation, and noisy queries. A model may perform well on clean image-text pairs but fail when images are low-resolution, text is colloquial, or metadata is missing. Adversarial actors may inject optimized content to manipulate rankings, especially in advertising, political, or reputation-sensitive contexts. Retrieval systems should include anomaly detection, source diversity constraints, and uncertainty estimation. They should also degrade gracefully when a modality is unavailable. Evaluation must go beyond average ranking metrics and examine worst-case behavior, tail queries, and minority-language performance.

Fairness is both a ranking outcome and a process concern. Outcome fairness asks whether different groups receive comparable access to relevant information and opportunities. Process fairness asks whether data collection, model training, and deployment decisions include affected communities. In multimodal retrieval, fairness issues can arise from imbalanced training data, inaccessible interfaces, captioning bias, or ranking policies that disadvantage minority viewpoints. Foundation model research emphasizes that downstream harms are often context-dependent and difficult to anticipate [17]. Ethical risk analyses of language models identify hazards such as misinformation, discrimination, and privacy violation [21]. Adaptive reranking should therefore include participatory design, impact assessments, and transparent appeal mechanisms.

Evaluation should combine automated metrics with human judgment and policy review. Offline benchmarks are useful but often fail to capture real user intent, multimodal ambiguity, or long-term societal effects. Online experiments can reveal engagement and satisfaction, but they may optimize for short-term clicks rather than user welfare. A robust evaluation program should include adversarial testing, counterfactual analysis, and longitudinal monitoring. It should also report system limitations through model cards and datasheets [20]. Because adaptive reranking influences what evidence downstream models see, evaluation must trace errors from retrieval through reranking to generation. This end-to-end perspective is necessary for accountability.

7. Governance, Policy, and Organizational Implications

Governance of adaptive LLM reranking must connect technical design to institutional authority. Organizations need clear ownership of ranking policies, escalation paths for harmful results, and review processes for model updates. Policies should specify what data can be used for personalization, how long queries and results are retained, and when human review is required. In regulated domains such as healthcare, finance, and education, reranking may be subject to explainability, non-discrimination, and due process requirements. Public-sector search systems must balance efficiency with transparency, accessibility, and democratic accountability. Policy frameworks should not be an afterthought; they should shape architecture from the beginning.

Procurement is a key governance lever. When organizations rely on external model providers, contracts should address data use, model updates, audit rights, service levels, and environmental reporting. Vendor lock-in can reduce flexibility and obscure accountability. Open-source models offer more inspectability but require internal expertise and infrastructure. A multi-role reranking workflow may distribute responsibility across several models, making it harder to assign accountability for errors [10]. Governance should therefore require provenance tracking across model calls, prompt versions, and policy decisions. It should also define how users can contest rankings that affect their rights or opportunities.

Policy also matters at the societal level. Reranking systems shape access to news, employment, housing, health information, and educational resources. If adaptive models prioritize engagement or commercial objectives, they may amplify polarization, misinformation, or inequality. If they prioritize safety too aggressively, they may suppress legitimate speech or minority perspectives. The appropriate balance is a matter of public deliberation, not only engineering optimization. Regulatory approaches such as transparency obligations, risk assessments, and data protection rules can provide baseline safeguards. However, organizations must go further by building internal review boards, external audits, and stakeholder feedback channels. Governance is ultimately about maintaining legitimacy in systems that mediate knowledge.

8. Case Illustrations and Cross-Domain Comparisons

Web search illustrates the scale and diversity of adaptive reranking. A search engine must handle navigational, transactional, informational, and multimodal queries within milliseconds. It may use large models for complex question answering but smaller models for common queries. Ranking policies must address spam, misinformation, and monetization. Enterprise search presents a different profile. The corpus is smaller but more sensitive, with access controls, terminology, and compliance requirements. Adaptive reranking can improve document discovery, but it must respect permissions and avoid leaking confidential information across departments. The same architecture may require different governance depending on whether the enterprise is a hospital, law firm, or manufacturing company.

Healthcare multimodal retrieval combines medical images, clinical notes, genomic data, and literature. Reranking must prioritize accuracy, safety, and privacy while avoiding bias across patient populations. A fluent but unsupported generated answer can be dangerous. Education retrieval may focus on learning materials, videos, and interactive content. Fairness concerns include accessibility, language support, and equitable representation. Public-sector systems may involve benefits, legal rights, and civic information, where accountability and appeal rights are essential. In each domain, the technical reranking problem is similar, but the acceptable trade-offs differ. Cross-domain comparisons show that adaptive reranking cannot be governed by a single universal policy.

Multimodal retrieval-augmented generation provides a connecting thread across these cases. MuRAG demonstrated how images and text can be retrieved for open question answering [13], and retrieval-augmented multimodal language modeling showed the value of external evidence during generation [14]. MiniGPT-4 illustrated the flexibility of vision-language assistants [15]. Yet these capabilities depend on reliable retrieval and ranking. If the reranker selects biased, outdated, or irrelevant evidence, the generator may produce confident errors. Case studies should therefore report not only answer quality but also retrieval provenance, modality balance, and failure modes. This is especially important when systems are deployed in high-stakes settings.

9. Future Directions

Future research should develop adaptive rerankers that are efficient, auditable, and context-aware. One direction is query-adaptive model selection, where the system predicts whether a large language model is necessary based on query complexity, risk, and available resources. Another is continual learning under policy constraints, so that rerankers can adapt to new domains without erasing fairness safeguards. Federated and privacy-preserving retrieval could allow sensitive data to remain local while still benefiting from shared models. Energy-aware scheduling could shift non-urgent workloads to lower-carbon periods. These approaches require interdisciplinary collaboration among information retrieval, systems engineering, law, ethics, and human-computer interaction.

Evaluation infrastructure is another priority. Benchmarks should include multimodal ambiguity, multilingual queries, accessibility scenarios, and adversarial content. They should measure latency, cost, energy, and fairness alongside relevance. Reproducibility requires open prompts, versioned indexes, and documented data provenance. Human evaluation should include affected communities, not only expert annotators. Model reporting practices such as model cards and datasheets can support transparency [20], but they must be updated continuously. A useful evaluation program will treat deployment as an ongoing experiment with feedback loops and safeguards.

Finally, governance research should examine how organizations actually make trade-offs. It is not enough to list principles; we need empirical studies of procurement, audit, incident response, and user appeal. Adaptive reranking is a site where technical choices become policy choices. The decision to use a large model, retain a query, personalize results, or prioritize source diversity has distributive consequences. Future systems should make these choices explicit, contestable, and reversible. That is the basis for trustworthy multimodal retrieval at scale.

10. Conclusion

Adaptive LLM reranking offers a powerful way to improve multimodal information retrieval, but it also concentrates technical, economic, and political power in ranking infrastructures. This paper has argued that the relevant unit of analysis is not the reranker alone but the socio-technical system that includes data pipelines, encoders, indexes, model providers, policies, and users. A multi-role collaborative reranking workflow can improve coordination among reasoning stages [10], yet its benefits depend on governance, sustainability, robustness, and fairness. Without these supports, adaptive reranking may increase relevance while deepening opacity, bias, and environmental cost. The path forward requires architectures that degrade gracefully, evaluations that measure what matters, and governance that connects model behavior to institutional accountability.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics.
3. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 8748–8763). PMLR.
4. Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 19730–19742). PMLR.
5. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 34892–34916.
6. Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.
7. 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.
8. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (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.
9. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
10. 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).
11. Sun, W., Yan, L., Ma, X., Wang, S., Ren, P., Chen, Z., ... & Ren, Z. (2023). Is ChatGPT good at search? Investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 14918–14937). Association for Computational Linguistics.

12. Pradeep, R., Sharifymoghaddam, S., & Lin, J. (2023). RankVicuna: Zero-shot listwise document reranking with open-source large language models. arXiv preprint arXiv:2309.15088.
13. Chen, W., Hu, H., Chen, X., Verga, P., Cohen, W. W., & Wang, W. (2022). MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 5558–5570). Association for Computational Linguistics.
14. Yasunaga, M., Aghajanyan, A., Shi, H., James, R., Leskovec, J., Liang, P., ... & Fried, D. (2023). Retrieval-augmented multimodal language modeling. In Proceedings of the 40th International Conference on Machine Learning (Vol. 202, pp. 39755–39769). PMLR.
15. Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). MiniGPT-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592.
16. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). Association for Computing Machinery.
17. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
18. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650). Association for Computational Linguistics.
19. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63.
20. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229). Association for Computing Machinery.
21. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.
22. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems, 28, 2503–2511.