

Uncertainty-Aware Retrieval-Augmented Generation for Hallucination Detection and Factuality Calibration

Arif Haering

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

Urszula Bahr

Department of Computer Science, University of New Hampshire, Durham, NH, USA.
ubahr@unh.edu

Sanya Manne

Department of Electrical Engineering and Computer Science, University of Missouri,
Columbia, MO, USA.
sanya.manne@missouri.edu

Tola Meler

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

Abstract

Retrieval-augmented generation has become a central architecture for grounding large language models in external evidence, yet it does not eliminate hallucination. Retrieval can return stale, contradictory, or weakly relevant passages; generation can misread evidence; and fusion mechanisms can privilege fluent synthesis over verifiable attribution. This paper develops a system-level account of uncertainty-aware retrieval-augmented generation for hallucination detection and factuality calibration. It argues that uncertainty should be treated as a distributed property of the retrieval pipeline, the generator, the fusion layer, and the human-machine interface rather than as a single post hoc confidence score. The paper examines architectural trade-offs among sparse and dense retrieval, multi-source fusion, reranking, evidence attribution, calibration, and real-time deployment. It then connects those technical choices to governance, auditability, fairness, sustainability, and policy. The central claim is that reliable factuality requires calibrated uncertainty signals that are actionable at the point of decision, visible to users, and accountable within institutions. By integrating uncertainty estimation with retrieval provenance, attention-based fusion, and verification workflows, systems can detect hallucination more robustly and calibrate factuality in ways that support trust without implying infallibility. The paper concludes with future directions for adaptive retrieval, multimodal grounding, and regulated deployment.

Keywords

retrieval-augmented generation, hallucination detection, uncertainty calibration, factuality, multi-source retrieval, governance, socio-technical systems.

1. Introduction

Large language models have demonstrated remarkable fluency across knowledge-intensive tasks, but their tendency to produce plausible yet unsupported statements remains a foundational obstacle to dependable deployment [1]. Hallucination is not merely an isolated model error; it is a socio-technical failure that emerges when generated language is disconnected from verifiable evidence, institutional accountability, or user expectations [6]. Retrieval-augmented generation has emerged as a practical response by coupling generative models with external memory, document stores, or search infrastructure [1], [2]. The promise is straightforward: if the model can retrieve relevant evidence before generating, then answers can be grounded in sources rather than solely in parametric memory. In practice, however, retrieval introduces new sources of uncertainty, including index staleness, retrieval ambiguity, passage conflict, and fusion errors that can create a false sense of evidential support [7].

The problem is especially acute in high-stakes settings such as medicine, law, public administration, finance, and scientific research, where an unsupported claim can lead to material harm. A system that retrieves a genuine document but misinterprets it may appear factual because it cites a source, even though the generated conclusion is not entailed by that source. Earlier work on factuality and faithfulness in summarization shows that fluency and source overlap are insufficient indicators of truthfulness [8]. Consequently, hallucination detection cannot be reduced to whether a response contains citations. It must evaluate whether the generated content is supported by retrieved evidence and whether the system communicates its own uncertainty in a calibrated manner.

Uncertainty-aware retrieval-augmented generation addresses this gap by treating uncertainty as a first-class signal across the entire pipeline [10], [11]. Instead of generating an answer and then assigning a confidence score, an uncertainty-aware system monitors retrieval relevance, evidence consistency, generation stability, and attribution strength. These signals can be fused into user-facing indicators, abstention policies, or escalation pathways. The design challenge is not simply to estimate uncertainty but to make it operationally meaningful. A confidence score that is invisible, poorly calibrated, or disconnected from downstream decisions does little to reduce hallucination risk [12]. Governance frameworks for artificial intelligence increasingly demand traceability, contestability, and harm mitigation, which means uncertainty must be embedded in architecture, workflow, and policy [15], [18].

This paper advances a system-level perspective on uncertainty-aware retrieval-augmented generation. It first reviews the foundations of retrieval-augmented generation and its evolution toward multi-source and heterogeneous retrieval. It then analyzes uncertainty across retrieval, generation, and fusion. Subsequent sections examine hallucination detection as a socio-technical control problem, factuality calibration at the interface, infrastructure trade-offs, governance implications, evaluation challenges, and cross-domain case comparisons. The paper argues that factuality is not a binary property of text but an institutional and technical achievement produced through calibrated uncertainty, provenance, verification, and accountability [17].

2. Foundations of Retrieval-Augmented Generation

Retrieval-augmented generation combines a parametric language model with a non-parametric memory that can be queried at inference time [1]. In early formulations, a retriever selected passages from a document collection and a generator conditioned its output on both the query and retrieved evidence. This architecture improved performance on open-domain question answering and knowledge-intensive tasks by allowing models to access information beyond their training parameters [3]. Dense passage retrieval further advanced this line by

learning embeddings that map questions and passages into a shared semantic space, enabling retrieval based on meaning rather than exact lexical overlap [4]. Retrieval-augmented generation therefore became not only a method for improving accuracy but also a mechanism for updating knowledge without retraining the entire model.

Subsequent systems scaled retrieval to trillions of tokens and explored different fusion strategies, including concatenation, cross-attention, and latent integration [2], [5]. These developments revealed an important structural tension. Retrieval improves coverage, but it also increases the complexity of evidence selection. A generator may receive many passages, some relevant, some redundant, and some contradictory. The model must then decide which evidence to prioritize, how to reconcile conflicts, and when to acknowledge that the retrieved context is insufficient. When retrieval is treated as an undifferentiated context window, long-context models may attend unevenly to evidence, sometimes neglecting important material in the middle of the input [22]. Thus, retrieval quality and context organization are not secondary concerns; they directly shape factuality.

Contemporary retrieval-augmented generation has moved toward hybrid and multi-source architectures. Systems may combine sparse retrieval, dense retrieval, structured databases, knowledge graphs, and domain-specific indexes. A unified semantic space and attention-based fusion framework for multi-source heterogeneous retrieval-augmented generation illustrates the growing need to align evidence from different modalities, formats, and provenance structures [7]. This shift complicates uncertainty estimation because different sources carry different reliability profiles. A peer-reviewed article, a government database, a customer support transcript, and a social media post may all be retrievable, but they should not be treated as equally authoritative. A system-level design must therefore represent source trust, recency, jurisdiction, and semantic compatibility alongside relevance scores.

The expansion of retrieval-augmented generation has also produced surveys that map retrieval, generation, and augmentation techniques across tasks [21]. These surveys highlight that retrieval is not a single operation but a pipeline involving query formulation, indexing, retrieval, reranking, evidence selection, and generation. Each stage can introduce uncertainty. Query formulation may omit critical context. Indexing may exclude relevant documents. Retrieval may return superficially similar but evidentially weak passages. Reranking may optimize for relevance while ignoring authority or contradiction. Generation may then overstate the coherence of fragmented evidence. An uncertainty-aware architecture must therefore instrument each stage rather than rely on a single final confidence score. This distributed view is essential for hallucination detection because errors often arise from accumulated small uncertainties rather than from one catastrophic failure.

3. Uncertainty in Retrieval, Generation, and Fusion

Uncertainty in retrieval-augmented generation can be divided conceptually into data uncertainty, model uncertainty, and decision uncertainty. Data uncertainty arises from missing, ambiguous, outdated, or conflicting evidence. Model uncertainty arises from the generator’s limited knowledge, reasoning instability, or sensitivity to prompt structure. Decision uncertainty arises when the system must choose whether to answer, abstain, qualify, or escalate. Classical work on neural network calibration shows that modern models can be overconfident even when incorrect, making raw probabilities unreliable without post hoc calibration [12]. Bayesian approximations and deep ensembles provide complementary ways to estimate uncertainty by capturing variability across model states or training runs [13], [14].

In retrieval-augmented systems, these techniques must be extended to retrieval and fusion components, not only to the generator.

Retrieval uncertainty is particularly consequential because retrieval errors can be difficult for users to detect. A retrieved passage may be topically related but not answer-bearing. Multiple passages may support competing claims. A source may be authoritative in one jurisdiction but irrelevant in another. If the system presents retrieved evidence without indicating heterogeneity or conflict, users may interpret citation presence as proof. Uncertainty-aware retrieval should therefore estimate relevance confidence, source reliability, temporal validity, and contradiction risk. These signals can be used to rerank evidence, trigger additional retrieval, or generate caveats. The multi-source fusion problem becomes especially complex when sources use different vocabularies, schemas, or modalities [7]. A unified semantic space can improve alignment, but it can also obscure provenance unless the architecture preserves source-level metadata throughout the pipeline.

Generation uncertainty is equally multifaceted. A model may be uncertain because the evidence is incomplete, because the question is ambiguous, or because the answer requires reasoning that exceeds the retrieved context. Research on whether language models know what they know suggests that models can sometimes recognize their own ignorance, but this capacity is imperfect and context-dependent [10]. Verbalized uncertainty offers a more user-facing form of calibration, where models express confidence in natural language [11]. However, verbalized confidence can be inconsistent, culturally variable, and vulnerable to prompt framing. It should not replace internal uncertainty estimates but can complement them by making uncertainty more interpretable to non-expert users. The key system requirement is consistency between internal confidence, external expression, and downstream action.

Fusion uncertainty arises when the system combines evidence, retrieval scores, and generated claims. Attention-based fusion can weigh sources dynamically, but it can also create opacity. If the fusion layer prioritizes fluency or semantic similarity over source authority, the resulting answer may be well-written but weakly grounded. An uncertainty-aware fusion mechanism should produce not only an answer but also an attribution map indicating which claims depend on which sources and how strongly. This map becomes the basis for hallucination detection. If a generated sentence cannot be traced to any retrieved passage, or if it depends on a passage with low relevance and low authority, the system should mark it as high risk. In this sense, uncertainty is not merely a numerical confidence score; it is a structured representation of evidentiary support.

4. Hallucination Detection as a Socio-Technical Control Problem

Hallucination detection has often been framed as a model evaluation problem, but in deployed systems it is better understood as a control problem. The goal is not only to classify an output as factual or non-factual after generation. The goal is to prevent, detect, and mitigate unsupported claims within an operational workflow. Surveys of natural language generation hallucination distinguish between intrinsic and extrinsic errors, between faithfulness and factuality, and between source-based and world-knowledge-based evaluation [6]. These distinctions matter because retrieval-augmented systems may hallucinate in different ways. They may ignore retrieved evidence, misattribute a claim to the wrong source, or blend retrieved facts with unsupported parametric knowledge. A detection layer must therefore examine both the generated text and its relationship to the evidence set.

Attribution measurement provides a useful foundation for detection. Methods for measuring attribution in natural language generation models assess whether generated statements are supported by cited or retrieved sources [9]. Fact verification datasets such as FEVER have similarly advanced the automated evaluation of claims against evidence [19]. In a retrieval-augmented system, these techniques can be adapted to produce claim-level support labels. Each sentence or proposition can be classified as supported, partially supported, contradicted, or unsupported. This granularity is important because a single answer may contain both accurate and inaccurate statements. A binary document-level label hides the locations and causes of hallucination. Claim-level attribution allows the system to highlight risky segments, request additional evidence, or route the response to human review.

Yet automated detection remains limited by the quality of retrieval and the ambiguity of natural language. A claim may be entailed by a source under one interpretation but not another. Numerical values, temporal references, and causal assertions are particularly difficult to verify. A system may retrieve a passage that contains the correct entities but the wrong relation, leading to a plausible but false synthesis. For this reason, hallucination detection should combine automated verification with human oversight in high-stakes contexts. The socio-technical control problem involves designing escalation thresholds, review queues, and audit trails. If uncertainty exceeds a threshold, the system might abstain, ask a clarifying question, retrieve additional sources, or present competing interpretations. The appropriate response depends on domain, user role, and consequence of error.

This control perspective also changes how success is measured. A system that answers fewer questions but abstains when evidence is weak may be more reliable than a system that maximizes coverage while hiding uncertainty. In customer service, a qualified answer may be sufficient. In clinical decision support, no answer may be safer than a confident but unsupported recommendation. In legal research, competing precedents may need to be displayed rather than resolved. Hallucination detection therefore cannot be evaluated solely by detection accuracy. It must be evaluated by decision quality, including the costs of false reassurance, unnecessary abstention, and delayed human intervention. The design of uncertainty thresholds is thus an ethical and institutional choice, not merely an engineering parameter.

5. Factuality Calibration and User-Facing Reliability

Factuality calibration concerns the correspondence between a system’s expressed confidence and its actual correctness. A model that is 80 percent confident should be correct about 80 percent of the time on comparable cases. Modern neural networks often violate this expectation, producing overconfident predictions that require temperature scaling or other calibration methods [12]. In retrieval-augmented generation, calibration must account for multiple uncertainty sources. A system may be confident in its language generation but uncertain about retrieval relevance. It may be confident in a retrieved source but uncertain about whether the source applies to the user’s jurisdiction. A single calibrated probability for the entire answer may therefore be misleading. More useful is a calibrated profile that separates answer correctness, evidence sufficiency, source authority, and temporal validity.

Verbalized uncertainty provides one interface for communicating these distinctions. When models express uncertainty in words, users may better understand when an answer is tentative or incomplete [11]. However, verbal expressions must be grounded in internal estimates and validated against outcomes. Otherwise, they become stylistic hedges rather than reliable signals. Research suggesting that language models can sometimes know what they know

provides a basis for self-assessment, but it also shows that self-knowledge is uneven [10]. Calibration should therefore be evaluated across domains and user groups. A phrase such as “likely” may be interpreted differently by a physician, a lawyer, and a consumer. Interface design must align linguistic expressions with numerical or visual indicators where appropriate, while avoiding false precision.

For factuality calibration to improve reliability, uncertainty must be actionable. A calibrated score is useless if the system cannot change its behavior in response. An uncertainty-aware system might lower confidence when retrieved sources conflict, trigger a second retrieval pass when support is weak, or present an answer with explicit caveats when evidence is partial. Dense retrieval can be used to fetch additional passages that either confirm or contradict a claim [4]. Ethical and social risk frameworks emphasize that language models can cause harm through misinformation, bias, and over-reliance, which makes transparent uncertainty communication a governance requirement as much as a user-experience feature [16]. Users should not be led to believe that a citation guarantees truth. They should be able to see what is known, what is uncertain, and what would be required to resolve the uncertainty.

Calibration also has temporal and institutional dimensions. Factual status can change as laws, scientific consensus, public health guidance, and organizational policies evolve. A statement that was correct last year may be outdated today. A system calibrated only on static benchmarks may therefore become miscalibrated in deployment. Uncertainty-aware retrieval can incorporate recency, versioning, and source validity periods. It can also record when an answer was generated and which evidence version was used. This provenance supports auditability and enables later correction. Factuality calibration is thus not a one-time model adjustment but a continuous operational process involving monitoring, feedback, and governance.

6. Architecture and Infrastructure Trade-Offs

Uncertainty-aware retrieval-augmented generation imposes architectural demands. The system must support retrieval, reranking, evidence fusion, generation, attribution, calibration, and logging without introducing unacceptable latency. Large-scale retrieval from trillions of tokens demonstrates the potential of massive external memory, but it also raises infrastructure costs and complexity [2]. Hybrid retrieval can improve robustness by combining lexical matching with dense semantic search, but it requires maintaining multiple indexes and fusion logic. Multi-source heterogeneous retrieval adds further complexity because sources may differ in schema, language, modality, and update frequency [7]. A unified semantic space can reduce friction, but it must preserve provenance and source-specific reliability metadata.

Deployment trade-offs are central. Real-time applications such as customer support or search assistance require low latency, which limits the number of retrieval passes and verification steps. High-stakes applications such as clinical decision support or legal analysis may tolerate higher latency in exchange for stronger evidence checking. Caching can improve responsiveness, but cached answers may become stale or contextually inappropriate. Streaming retrieval can reduce initial delay, but it complicates attribution because evidence arrives while generation is underway. The architecture must therefore expose configurable uncertainty thresholds that align with domain requirements. A single universal configuration is unlikely to satisfy both a low-risk informational query and a high-risk professional recommendation.

Infrastructure sustainability is another structural concern. Large-scale retrieval and repeated generation consume energy, storage, and network bandwidth. Maintaining fresh indexes, embeddings, and audit logs adds ongoing operational cost [18]. If uncertainty-aware systems simply retrieve more passages and run more verification models for every query, they may improve factuality at the expense of environmental and economic efficiency. A more sustainable design uses adaptive retrieval: it begins with lightweight retrieval and escalates only when uncertainty is high [21]. Reranking and verification can be selectively applied. Caching can be combined with version checks to avoid stale reuse. These choices are not merely optimizations; they determine whether factuality calibration can be sustained at scale.

Long-context models introduce a related trade-off. Providing more retrieved text can improve coverage, but models may attend unevenly across long inputs, with important evidence sometimes lost in the middle [22]. Simply expanding the context window does not guarantee better grounding. Instead, systems should structure evidence, place key passages strategically, and use attention-based fusion to highlight source relationships. The fusion framework for multi-source heterogeneous retrieval illustrates how attention can integrate diverse evidence, but it also underscores the need for transparency in how sources are weighted [7]. If the system cannot explain why one source dominated another, users and auditors cannot evaluate the basis of the answer. Architecture must therefore balance performance, cost, provenance, and explainability.

7. Governance, Policy, and Sustainability Implications

Governance for uncertainty-aware retrieval-augmented generation must address accountability across the data lifecycle. Who is responsible when a retrieved source is outdated? Who is accountable when a fusion model weights a low-authority source too heavily? Who audits the calibration of confidence expressions? These questions cannot be answered by model documentation alone. They require institutional roles, logging standards, and escalation procedures. AI safety research has long emphasized that specification, robustness, and oversight are interdependent [15]. In retrieval-augmented systems, oversight depends on evidence provenance. If the system cannot reconstruct which sources supported which claims, meaningful review is impossible.

Policy frameworks increasingly demand transparency, contestability, and harm prevention for automated decision systems. Ethical and social risk analyses of language models identify misinformation, bias, privacy leakage, and over-reliance as key concerns [16]. Uncertainty-aware retrieval can mitigate some of these risks by making evidentiary limits visible, but it can also create new governance challenges. A confidence score may be treated as objective even when it reflects incomplete calibration. A citation may be treated as endorsement even when the source is contested. Governance must therefore specify how uncertainty is interpreted, who can override it, and how users can contest automated outputs. This is especially important in public-sector deployment, where legitimacy depends on procedural fairness and clear redress.

Sustainability policy intersects with governance through procurement and infrastructure planning. Organizations may adopt retrieval-augmented generation to reduce retraining costs, since external indexes can be updated without retraining the full model. However, maintaining retrieval infrastructure, embedding pipelines, and verification services creates continuous energy and labor demands [17]. A governance framework should consider total cost of ownership, not just inference cost. It should also address labor: human reviewers, data curators, and domain experts are essential for calibrating uncertainty and correcting

hallucination, yet their work is often invisible in system evaluations. A sustainable socio-technical design recognizes that factuality is maintained by people and institutions as well as by models.

International and cross-jurisdictional deployment adds further complexity. A source that is authoritative in one country may be inadmissible in another. Privacy rules may restrict the retrieval of personal data. Copyright and licensing may limit which documents can be indexed and surfaced. Foundation model research has highlighted the broad societal risks and opportunities of large-scale systems, including concentration of power and dependence on opaque infrastructure [18]. Uncertainty-aware retrieval should therefore include policy metadata such as jurisdiction, license, retention period, and permitted use. Without such metadata, calibration may be technically accurate but legally or ethically invalid. Governance must treat provenance as a policy variable, not only a technical field.

8. Evaluation, Robustness, and Fairness

Evaluating uncertainty-aware retrieval-augmented generation requires benchmarks that go beyond answer accuracy. A system may produce correct answers while failing to detect hallucination, or it may detect uncertainty while failing to communicate it usefully. Evaluation should therefore measure claim-level attribution, calibration error, abstention quality, and user decision outcomes. Hallucination surveys provide taxonomies that can guide metric design, distinguishing between faithfulness, factuality, and source attribution [6]. Faithfulness and factuality in summarization research further shows that different tasks require different notions of support [8]. Attribution measurement methods offer ways to test whether generated statements are actually grounded in cited sources [9]. Fact verification datasets can be adapted to evaluate retrieval-augmented claims against evidence [19].

Robustness testing must include adversarial retrieval, conflicting sources, temporal drift, and distribution shift. A system may perform well on clean Wikipedia-style evidence but fail when sources contradict each other or when important evidence is missing. It may be vulnerable to prompt injection through retrieved documents, where malicious text instructs the generator to ignore safety rules or produce false claims. Uncertainty estimation can help by flagging unusual evidence patterns, but it is not a complete defense. Language models can also produce confident errors when parametric knowledge conflicts with retrieved evidence, even though they may sometimes recognize their own ignorance [10]. Robustness requires layered defenses including retrieval filtering, source allowlists, contradiction detection, and human escalation.

Fairness is intertwined with factuality. Retrieval corpora may underrepresent certain languages, regions, demographic groups, or perspectives. A system may be more uncertain about questions involving marginalized communities because evidence is scarce, and it may then default to majority perspectives that are not locally valid. Bender and colleagues warn that large language models can encode and amplify biases through training data and design choices [17]. Ethical risk analyses similarly identify discrimination and exclusion as key harms [16]. Uncertainty-aware systems should not simply hide these disparities behind confidence scores. They should monitor calibration across groups and domains. If the system is systematically underconfident for some users and overconfident for others, that is a fairness problem as well as a calibration problem.

Long-context evaluation also matters for fairness and robustness. If important evidence is buried in the middle of a long retrieved context, users who ask complex questions may

receive less reliable answers [22]. Multi-source retrieval can improve coverage, but it can also amplify dominant sources if fusion weights are not audited [7]. Evaluation should therefore include provenance audits: which sources are retrieved, which are cited, and which are ignored. It should examine whether source diversity correlates with answer quality and whether uncertainty is communicated consistently across languages and formats. Factuality evaluation must be continuous, because the evidentiary landscape changes. Static benchmarks alone cannot capture the evolving risks of deployed retrieval-augmented systems.

9. Case Illustrations and Cross-Domain Comparisons

In healthcare, uncertainty-aware retrieval-augmented generation could support clinicians by retrieving clinical guidelines, drug labels, and peer-reviewed literature. However, the cost of hallucination is high. A system that retrieves an outdated guideline or misstates a contraindication could contribute to harm. Dense retrieval can help locate semantically relevant passages [4], but it must be paired with source authority metadata and temporal validity checks. The system should distinguish between general medical knowledge and patient-specific recommendations. Uncertainty signals should trigger clinician review when evidence is conflicting or when the patient context is incomplete. In this domain, abstention and escalation are not failures; they are safety features. Weidinger and colleagues emphasize that language model harms are context-dependent, which means uncertainty thresholds must be calibrated to clinical workflows [16].

In legal research, retrieval-augmented generation can help find statutes, regulations, and case law [1]. Yet legal factuality depends on jurisdiction, hierarchy of authority, and temporal validity. A retrieved case may be overruled or distinguishable. A statute may have amendments not captured by the index. Uncertainty-aware systems should display competing interpretations and indicate when a source is binding, persuasive, or outdated. Attribution measurement can help identify which legal propositions depend on which sources [9]. Fact verification methods can be adapted to legal claims, but they must handle defeasible reasoning and interpretive ambiguity [19]. The system should not present a single confident answer when the legal question is genuinely contested.

In scientific literature review, retrieval-augmented generation can synthesize findings across papers, but scientific consensus is often provisional. A system may retrieve studies with different methodologies, sample sizes, and effect directions. Multi-source heterogeneous fusion can integrate these sources, but it must preserve methodological metadata [7]. Uncertainty should reflect disagreement among studies, not merely retrieval relevance. In customer service and enterprise knowledge management, the stakes may be lower, but the volume is higher. A system can use adaptive retrieval to answer routine questions quickly and escalate uncertain cases [21]. Calibration should be tuned to organizational tolerance for error. In all domains, the common lesson is that factuality is not a universal property. It is defined by the evidentiary standards, legal rules, and professional norms of the setting.

10. Future Directions

Future research should develop uncertainty representations that are compositional and auditable. Instead of a single confidence score, systems could expose an evidence graph showing which claims depend on which sources, how sources relate, and where contradictions remain. Verbalized uncertainty can make these signals accessible to users [11], while ensemble and Bayesian methods can provide internal stability estimates [14]. The challenge is to align these layers without overwhelming users or auditors. Interfaces should present

uncertainty in ways that support appropriate reliance: enough detail to inform judgment, but not so much that users ignore it. Adaptive retrieval should be governed by explicit policies that determine when additional evidence is required [21].

Another direction is multimodal and cross-lingual grounding. Retrieval-augmented generation will increasingly draw on images, tables, audio, and video. Unified semantic spaces must represent modality-specific uncertainty and provenance. Cross-lingual retrieval raises calibration challenges because evidence quality and source authority vary across languages. Systems must avoid treating high-resource-language sources as universally authoritative. They must also handle translation uncertainty, which can introduce subtle factual errors. Long-context models will continue to expand, but research shows that simply providing more context does not guarantee effective use [22]. Future architectures should combine structured evidence selection with attention-based fusion and explicit provenance.

Finally, governance research should move from principles to operational controls. This includes standard audit logs, calibration reporting, source licensing registries, incident response procedures, and user contestability mechanisms. Policy should require that high-stakes deployments document their uncertainty thresholds and the rationale for abstention. Sustainability should be considered in procurement, with attention to the energy and labor costs of continuous retrieval and verification. The most promising path is not to pursue perfectly factual models, which may be unattainable, but to build systems that recognize, communicate, and manage uncertainty in ways that support human judgment and institutional accountability.

11. Conclusion

Uncertainty-aware retrieval-augmented generation offers a practical framework for hallucination detection and factuality calibration, but only if uncertainty is treated as a distributed and operational property of the entire system. Retrieval, fusion, generation, attribution, interface design, and governance all contribute to whether an answer is trustworthy. A system that retrieves evidence but cannot assess its relevance, authority, or consistency may create false confidence. A system that estimates uncertainty but cannot act on it may fail to prevent harm. The architecture must therefore connect uncertainty signals to concrete behaviors such as reranking, additional retrieval, abstention, caveating, or human escalation. It must also preserve provenance so that claims can be traced, audited, and contested.

The paper has argued that factuality is not simply a model output but a socio-technical achievement. It depends on calibrated confidence, transparent evidence, domain-specific thresholds, and accountable institutions. Multi-source and heterogeneous retrieval increases coverage but also introduces new forms of uncertainty that require unified semantic representation and careful fusion [7]. Evaluation must move beyond accuracy to measure calibration, robustness, fairness, and decision quality. Governance must treat uncertainty as a policy concern, not only a technical metric. By integrating these perspectives, uncertainty-aware retrieval-augmented generation can support more reliable, contestable, and sustainable knowledge systems. The goal is not to eliminate uncertainty, but to make it visible, actionable, and aligned with the consequences of being wrong.

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. 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., Driessche, G., ... & 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.
3. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. W. (2020). Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 3929–3938). PMLR.
4. 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.
5. 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* (pp. 874–880). Association for Computational Linguistics.
6. 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. <https://doi.org/10.1145/3571730>
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. Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919). Association for Computational Linguistics.
9. Rashkin, H., Nikolaev, V., Lamm, M., Aroyo, L., Collins, M., Das, D., Petrov, S., Tomar, G. S., Turc, I., & Reitter, D. (2021). Measuring attribution in natural language generation models. *arXiv preprint arXiv:2112.12870*.
10. Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., ... & Kaplan, J. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
11. Lin, S., Hilton, J., & Evans, O. (2022). Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*. <https://arxiv.org/abs/2205.14334>
12. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1321–1330). PMLR.
13. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning* (pp. 1050–1059). PMLR.

14. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems* 30 (pp. 6402–6413). Curran Associates.
15. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
16. 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. (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214–229). ACM. <https://doi.org/10.1145/3531146.3533088>
17. 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). ACM. <https://doi.org/10.1145/3442188.3445922>
18. 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*.
19. Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 809–819). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N18-1074>
20. Asai, A., Min, S., Zhong, Z., & Chen, D. (2023). Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts* (pp. 41–46). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-tutorials.6>
21. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
22. 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. *arXiv preprint arXiv:2307.03172*.