

Multi-Agent Collaboration for Uncertainty-Aware RAG: Dynamic Evidence Verification and Risk-Controlled Generation

Gianni Siffredi

Department of Computer Science, University of North Texas, Denton, TX, USA.
hellogianni@unt.edu

Spencer Mcpherson

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
spencer993@oregonstate.edu

Abstract

Retrieval-augmented generation has become a central architecture for grounding large language models in external knowledge, yet its operational reliability remains fragile when retrieval quality, source authority, and generation confidence diverge. This paper examines multi-agent collaboration as a system-level strategy for uncertainty-aware retrieval-augmented generation, with emphasis on dynamic evidence verification and risk-controlled generation. The argument is that trustworthy generation cannot be achieved through a single retrieval pass or a monolithic model call. Instead, it requires an orchestrated division of labor among specialized agents responsible for query interpretation, evidence discovery, source validation, contradiction analysis, uncertainty estimation, policy adjudication, and final response synthesis. Such an architecture introduces new structural trade-offs involving latency, cost, auditability, fairness, and deployment sustainability. The paper develops a conceptual framework for multi-agent uncertainty-aware retrieval-augmented generation, analyzes the governance implications of risk-controlled generation, and considers how dynamic evidence verification can be integrated into public-sector, clinical, legal, financial, and scientific workflows. It further assesses infrastructure requirements, failure modes, evaluation metrics, and policy constraints that shape real-world adoption. The central finding is that multi-agent collaboration improves trustworthiness only when uncertainty is treated as a first-class operational signal and when verification is embedded throughout the generation pipeline rather than appended as a post hoc filter. Risk control must therefore be understood not merely as model alignment but as socio-technical infrastructure involving institutional accountability, contestability, and continuous monitoring. The paper concludes that uncertainty-aware retrieval-augmented generation should be designed as a governed evidence system, not simply as a faster or larger language model application.

Keywords

retrieval-augmented generation, multi-agent systems, uncertainty quantification, evidence verification, risk-controlled generation, AI governance, trustworthy AI, socio-technical infrastructure.

1. Introduction

Retrieval-augmented generation emerged as a response to the limitations of parametric language models in knowledge-intensive tasks. By combining a retriever with a generator, the approach allows models to access external documents and condition outputs on evidence that may be more current, domain-specific, or institutionally authoritative than pretrained parameters alone [1]. This design has been widely adopted because it appears to address hallucination, stale knowledge, and lack of source attribution. However, retrieval augmentation does not eliminate uncertainty. It relocates uncertainty across a pipeline of query formulation, document selection, ranking, passage extraction, synthesis, and citation. A retrieved document may be irrelevant, outdated, internally contradictory, adversarially placed, or correct but misinterpreted. A generator may ignore evidence, overgeneralize from a single passage, or present uncertain claims with unwarranted confidence. Therefore, the central engineering question is not whether retrieval improves generation, but how the system can know when its evidence is sufficient, when it is ambiguous, and when it should abstain, escalate, or qualify its output.

The Transformer architecture provided the computational foundation for large-scale language modeling and later became the substrate for retrieval-augmented systems [2]. Yet the same flexibility that enables fluent synthesis also enables persuasive fabrication. Research on stochastic parrots and large language models has shown that scale alone does not guarantee social or epistemic reliability [3]. Surveys of hallucination in natural language generation further demonstrate that factual errors arise from multiple causes, including data sparsity, source conflict, decoding strategies, and evaluation blind spots [4]. Recent work on uncertainty-aware retrieval-augmented generation argues that trustworthy generation requires explicit risk control rather than post hoc confidence claims [5]. This paper builds on that premise but extends it through a multi-agent systems lens. Rather than treating uncertainty as a single score attached to an answer, it treats uncertainty as a distributed property of a collaborative evidence process.

Multi-agent systems offer a useful conceptual vocabulary because they foreground coordination, role specialization, communication protocols, and conflict resolution [6]. In retrieval-augmented generation, a single model call can be decomposed into multiple agents: a planner that interprets user intent, a retriever that formulates searches, a source critic that evaluates authority, a verifier that checks claims against evidence, a contradiction analyst that identifies conflicting passages, a risk controller that applies policy constraints, and a synthesizer that produces a bounded final response. This division of labor can improve robustness, but it also introduces coordination overhead, error propagation, and accountability ambiguity. The design challenge is to determine which forms of uncertainty should be resolved internally, which should be surfaced to users, and which should trigger refusal or human escalation. The remainder of this paper develops a system-level analysis of these trade-offs, with attention to architecture, governance, infrastructure, fairness, and policy.

2. Conceptual Foundations of Uncertainty-Aware RAG

Uncertainty in retrieval-augmented generation can be understood through several overlapping distinctions. The first is uncertainty about user intent. A query may be underspecified, ambiguous, or based on a false premise. The second is uncertainty about retrieval adequacy. The system may lack documents, retrieve irrelevant passages, or retrieve only a narrow subset of relevant perspectives. The third is uncertainty about evidence quality. Sources differ in authority, recency, methodological rigor, and institutional standing. The fourth is uncertainty about synthesis. Even when evidence is adequate, the generator may compress, overstate, or

misinterpret it. The fifth is uncertainty about consequence. An error in a cooking recommendation has different risk than an error in medical triage, legal advice, or public health guidance. Uncertainty-aware generation therefore requires not only probabilistic confidence but contextual risk assessment.

Existing work on retrieval-augmented generation surveys has mapped many of these challenges, including retrieval quality, augmentation strategies, and evaluation gaps [12]. Self-reflective approaches such as Self-RAG attempt to let models decide when to retrieve, when to critique, and when to generate, thereby introducing internal control signals [13]. Related work on language models knowing what they know suggests that models can sometimes estimate their own knowledge boundaries, but calibration is imperfect and domain-dependent [14]. TruthfulQA shows that models may reproduce common falsehoods even when they have access to broad world knowledge [15]. These findings imply that uncertainty cannot be reduced to a single confidence score. It must be decomposed into evidence coverage, source reliability, contradiction level, claim specificity, and action risk.

Multi-agent collaboration addresses this decomposition by assigning different forms of uncertainty to different agents. A retrieval agent may focus on recall and coverage. A source evaluation agent may focus on authority and provenance. A contradiction agent may compare passages for consistency. A risk agent may translate uncertainty into policy actions, such as answer, qualify, ask a clarifying question, or refuse. A synthesis agent may then produce a response that reflects the system's collective confidence. This architecture is not merely a technical convenience. It is an epistemic design. It determines what counts as evidence, who is authorized to verify it, and how disagreement is represented. In high-stakes settings, the architecture becomes a governance mechanism because it encodes institutional priorities about safety, transparency, and accountability.

However, multi-agent uncertainty-aware retrieval-augmented generation also risks creating an illusion of rigor. Multiple agents can share the same underlying model biases, retrieval index limitations, or source hierarchies. If all agents rely on the same corrupted index, their agreement may be false consensus. If the verifier is trained on the same distribution as the generator, it may fail to detect the generator's characteristic errors. If the risk controller is optimized only for liability reduction, it may over-refuse and undermine access to legitimate information. Therefore, the conceptual foundation must include adversarial self-examination. The system should ask not only whether agents agree, but whether their disagreement is informative, whether their evidence is independent, and whether their roles introduce structural blind spots.

3. Multi-Agent Architectures for Dynamic Evidence Verification

A dynamic evidence verification architecture typically begins with a query understanding layer that decomposes user intent into sub-questions, constraints, and risk categories. This layer may be implemented as a planner agent or a set of specialized intent agents. It then dispatches retrieval tasks to multiple search agents that query different corpora, indexes, or source types. Retrieval may include dense vector search, sparse lexical search, structured databases, and curated knowledge bases. The purpose of multiplicity is not redundancy alone but epistemic diversity. A clinical query may require guidelines, systematic reviews, drug labels, and institutional protocols. A legal query may require statutes, case law, regulatory guidance, and secondary commentary. A public policy query may require demographic data, legislative history, budget documents, and stakeholder submissions.

Multi-agent frameworks such as generative agents, AutoGen, and MetaGPT provide early models for role-based collaboration among language agents [7][8][9]. Reflexion and ReAct show how agents can use verbal feedback and iterative reasoning-acting cycles to improve task performance [10][11]. In an uncertainty-aware retrieval-augmented generation system, these ideas can be adapted into verification loops. A verifier agent may extract claims from a draft answer and check each claim against retrieved evidence. A contradiction agent may identify passages that support, undermine, or complicate the claim. A provenance agent may assess whether sources are primary, secondary, peer-reviewed, governmental, commercial, or user-generated. A calibration agent may estimate whether the system's confidence is justified by evidence coverage and source agreement. The final synthesizer then produces an answer that includes qualifications, citations, and residual uncertainty.

The key structural trade-off is between depth and latency. Each additional verification loop improves reliability in principle, but it increases computational cost, response time, and coordination complexity. In interactive applications, users may tolerate only a few seconds of delay. In clinical decision support or legal research, longer deliberation may be acceptable if it reduces harm. Therefore, architectures should be risk-adaptive. Low-risk queries may use lightweight verification, while high-risk queries invoke additional agents, slower models, human review, or formal escalation. This requires a policy layer that maps query categories to verification intensity. The policy layer must be transparent enough to audit but flexible enough to accommodate domain-specific standards.

A second trade-off concerns centralization versus federation. A centralized multi-agent orchestrator can enforce consistent verification protocols, but it may become a bottleneck and a single point of failure. A federated architecture allows domain-specific agents to operate closer to local data and governance regimes, but it complicates coordination and evidence reconciliation. For example, a hospital network may require patient data to remain within a secure enclave, while a legal team may require privileged documents to remain within a controlled repository. Multi-agent collaboration can be designed with federated verification, where agents exchange uncertainty summaries and provenance attestations rather than raw sensitive data. This approach aligns with privacy-preserving infrastructure but requires standardized protocols for trust, identity, and audit.

A third trade-off concerns human roles. Human-in-the-loop verification can improve accountability, but it can also become a form of automation bias if reviewers defer to system outputs. Human-on-the-loop monitoring may be more scalable but less protective in high-consequence cases. The architecture should therefore specify escalation triggers based on uncertainty, risk category, and contestability. If evidence is contradictory or source authority is unclear, the system should not simply force an answer. It should present the conflict, request clarification, or route the query to a qualified human. This design treats uncertainty as actionable information rather than as a defect to be hidden.

4. Risk-Controlled Generation and Governance

Risk-controlled generation begins with the recognition that not all errors are equivalent. A factual error in a benign context may be tolerable if corrected quickly. A factual error in a medical, legal, financial, or security context can cause direct harm. Risk control therefore requires a taxonomy of consequence, reversibility, affected populations, and institutional liability. This taxonomy should be encoded in policy rather than left entirely to model intuition. Governance frameworks such as the NIST Artificial Intelligence Risk Management Framework provide a structured approach to mapping, measuring, managing, and monitoring

risks across the AI lifecycle [17]. In retrieval-augmented generation, this maps onto evidence governance, model governance, and operational governance.

Evidence governance defines which sources are admissible for which claims. It includes source whitelists, authority hierarchies, recency requirements, conflict-of-interest disclosures, and provenance metadata. Model governance defines how agents are trained, evaluated, updated, and constrained. It includes red-teaming, calibration testing, bias evaluation, and drift monitoring. Operational governance defines how the system is deployed, who is accountable, how incidents are reported, and how users can contest outputs. These layers are interdependent. A perfectly calibrated model can still produce harmful outputs if its retrieval corpus is biased. A well-curated corpus can still be misused if the synthesis layer overstates certainty. A transparent interface can still mislead if users lack the expertise to interpret uncertainty.

Ethical and policy scholarship on algorithms and AI principles emphasizes transparency, fairness, accountability, and human oversight [18][19]. These principles are often stated at a high level, but retrieval-augmented generation requires operational translation. Transparency may mean not only disclosing that an answer was generated by AI but also showing which sources were used, which were rejected, and where disagreement remains. Fairness may mean ensuring that retrieval does not systematically underrepresent marginalized communities or non-English sources. Accountability may mean assigning responsibility to a named human or institution for high-risk deployments. Human oversight may mean designing interfaces that encourage critical engagement rather than uncritical acceptance.

Risk-controlled generation also has implications for refusal. A system that refuses too often may be safe but unhelpful, while a system that answers too readily may be helpful but dangerous. The appropriate balance depends on context, user expertise, and institutional mandate. In a public health setting, refusal may be appropriate for individualized medical advice but inappropriate for general informational queries. In a legal aid setting, the system may provide procedural information while clearly declining to provide case-specific legal advice. In a financial setting, it may summarize regulations while refusing to recommend specific investments. Multi-agent architectures can support this nuance by separating the generation of candidate answers from the adjudication of release. The risk controller can decide whether an answer should be released, qualified, or withheld based on policy rules and uncertainty thresholds.

Governance must also address the temporal dimension. Evidence changes. Guidelines are updated, laws are amended, scientific consensus shifts, and new data emerge. A retrieval-augmented system that performs well at deployment may degrade as its corpus ages. Dynamic evidence verification must therefore include freshness checks, version tracking, and change detection. When a source is updated or retracted, the system should be able to identify affected claims and trigger re-verification. This is especially important in domains where retractions and corrections are common. The governance challenge is not only to prevent initial errors but to maintain trustworthiness over time.

5. Infrastructure, Deployment, and Sustainability

The infrastructure for multi-agent uncertainty-aware retrieval-augmented generation is substantially more complex than a single model endpoint. It includes document ingestion pipelines, optical character recognition, metadata extraction, chunking strategies, embedding generation, vector databases, lexical indexes, graph stores, orchestration services, policy

engines, monitoring systems, and audit logs. Each component introduces dependencies and failure modes. Poor chunking can separate a claim from its qualification. Poor metadata can obscure source authority. A stale vector index can retrieve obsolete evidence. An orchestration service can become a bottleneck. A policy engine can be misconfigured. These are not merely engineering details; they determine the epistemic quality of the system.

Deployment choices also shape sustainability. Running multiple agents, each potentially based on a large language model, multiplies computational cost and energy consumption. If every query invokes several retrieval and verification passes, the environmental footprint may become significant. Sustainability therefore requires optimization strategies such as caching verified evidence, reusing agent outputs when queries are similar, routing simple queries to smaller models, and batching verification tasks. It also requires architectural restraint. Not every query needs maximal verification. Risk-adaptive orchestration can reduce unnecessary computation while preserving safety in high-risk cases. The goal is not to minimize agents but to match verification intensity to consequence.

Cloud, edge, and hybrid deployments present different trade-offs. Cloud deployment offers scalability, centralized monitoring, and easier updates, but it raises data sovereignty and latency concerns. Edge deployment can reduce latency and keep sensitive data local, but it limits model size and coordination capability. Hybrid deployment may place retrieval and initial verification at the edge while routing high-risk cases to cloud-based or human review. This architecture aligns with institutional boundaries. A hospital may keep patient data local while using a cloud service for de-identified literature search. A government agency may keep classified documents on-premises while using external sources for public information. The design challenge is to maintain coherent uncertainty management across heterogeneous infrastructure.

Observability is another critical infrastructure requirement. Operators need to know not only whether the system answered a query but how it reached an answer. This requires logging retrieval queries, selected documents, verifier decisions, contradiction checks, risk classifications, and final outputs. These logs must be designed for auditability without creating new privacy risks. They must be retained according to policy and protected against tampering. In some domains, logs may be subject to legal discovery or regulatory inspection. Therefore, infrastructure design must integrate privacy, security, and records management from the outset. Multi-agent systems can improve auditability by making roles explicit, but they can also generate enormous volumes of intermediate data. The governance challenge is to determine what must be logged, what may be summarized, and what must be deleted.

6. Robustness, Fairness, and Policy Implications

Robustness in multi-agent retrieval-augmented generation must address adversarial and non-adversarial failures. Adversarial failures include poisoned documents, prompt injection, retrieval manipulation, and source impersonation. If an attacker can insert documents into the retrieval corpus, they may influence the verifier and synthesizer. Multi-agent verification can help if agents use independent sources and provenance checks, but it can also be circumvented if all agents trust the same compromised index. Robustness therefore requires diversity of evidence, integrity controls, and anomaly detection. It also requires red-teaming not only of the generator but of the entire retrieval and verification pipeline.

Non-adversarial failures include distribution shift, source drift, and coordination breakdown. A system trained on one domain may fail when deployed in another. A source that was

authoritative may become outdated. Agents may enter repetitive loops, disagree indefinitely, or converge prematurely. These failures are often harder to detect because they do not involve malicious intent. They require continuous evaluation and human oversight. The system should be able to detect when its confidence is poorly calibrated, when its evidence base is shrinking, or when its agents are producing inconsistent outputs. Such signals should trigger investigation rather than automatic deployment.

Fairness is deeply implicated in retrieval-augmented generation because retrieval determines whose knowledge is visible. If the corpus overrepresents dominant languages, institutions, or perspectives, the system may reproduce epistemic inequality [3]. If the source hierarchy privileges official documents over community knowledge, it may exclude marginalized voices. If the risk controller is trained on historical enforcement data, it may over-police certain populations. Fairness therefore cannot be reduced to model debiasing. It requires examining corpus composition, source selection, ranking algorithms, and policy rules. It also requires participation from affected communities in defining what counts as trustworthy evidence.

Policy implications extend beyond individual systems. Governments, regulators, and professional bodies are increasingly interested in how AI systems make claims and how those claims are justified. Retrieval-augmented generation offers a partial transparency mechanism because it can cite sources. However, citations can be performative if they do not actually support the generated claim. Multi-agent verification can improve citation faithfulness by checking each claim against evidence. This has implications for consumer protection, professional liability, and public communication. In journalism, for example, a multi-agent verification pipeline could help fact-check claims before publication. In education, it could help students distinguish between well-supported and weakly supported statements. In public administration, it could improve the reliability of citizen-facing information services.

Cross-domain comparisons reveal that the same architecture must be tuned differently across sectors. In medicine, uncertainty must be communicated in ways that support clinical judgment without replacing it. In law, evidence must be traceable to authoritative sources and privilege must be protected. In finance, risk disclosures and regulatory constraints may require conservative generation. In scientific research, uncertainty is inherent and should be represented rather than suppressed. In public health, misinformation resistance and accessibility are paramount. A single universal policy is unlikely to work. Instead, multi-agent systems should expose policy modules that can be configured by domain experts and audited by regulators.

7. Evaluation, Operational Metrics, and Case Illustrations

Evaluating multi-agent uncertainty-aware retrieval-augmented generation requires metrics that go beyond fluency and factual accuracy. Useful dimensions include evidence coverage, citation precision, citation recall, contradiction detection, calibration error, abstention appropriateness, risk-weighted harm, latency, cost, and user trust. Evidence coverage measures whether the system retrieved sufficient information to address the query. Citation precision measures whether cited sources actually support the associated claims. Contradiction detection measures whether the system identified conflicting evidence. Calibration error measures whether confidence estimates match observed correctness. Abstention appropriateness measures whether the system refused or escalated when it should have. Risk-weighted harm assigns greater weight to errors with severe consequences.

Case illustrations clarify these metrics. In a clinical decision support scenario, a multi-agent system might retrieve guidelines, drug interaction databases, and recent trials. A verifier agent checks whether the recommended information aligns with current guidelines. A contradiction agent notes that a recent study challenges part of the guideline. The risk controller then decides to present the guideline recommendation while noting the emerging evidence and recommending clinician review. The final output is not a single answer but a bounded evidence summary. In a legal research scenario, the system might retrieve statutes and case law. A provenance agent verifies jurisdiction and date. A contradiction agent identifies conflicting interpretations. The risk controller limits the output to legal information rather than case-specific advice. In a public health scenario, the system might retrieve vaccination guidance, epidemiological data, and community FAQs. A fairness agent checks whether the sources represent affected populations. The final output includes uncertainty about local conditions and directs users to local authorities.

These cases show that multi-agent collaboration is most valuable when it makes uncertainty visible rather than merely reducing it. The goal is not to produce an answer that appears certain. The goal is to produce an answer whose confidence is proportionate to evidence and consequence. This requires interfaces that communicate uncertainty without overwhelming users. Visual provenance maps, confidence bands, conflicting source summaries, and escalation recommendations can help. However, interface design must avoid creating false precision. Saying that a system is eighty percent confident may be misleading if confidence is not well calibrated. It may be more useful to say that multiple authoritative sources agree, that one major source disagrees, and that the evidence is current as of a specific date.

Evaluation must also include longitudinal monitoring. A system that performs well on a static benchmark may fail after deployment when sources change, user behavior shifts, or adversaries adapt. Continuous evaluation should track retrieval drift, source retraction rates, user correction reports, and incident logs. It should also include periodic red-teaming and external audit. In high-risk domains, evaluation should involve domain experts who can assess whether the system's uncertainty representations are meaningful. A clinician may care about whether a contradiction is clinically significant. A lawyer may care about whether a source is binding or persuasive. A policymaker may care about whether a risk classification aligns with statutory requirements. Multi-agent systems can support these distinctions by allowing domain-specific agents to participate in verification.

8. Conclusion

Multi-agent collaboration for uncertainty-aware retrieval-augmented generation represents a shift from model-centric AI to evidence-centric AI. The central claim of this paper is that trustworthy generation requires dynamic verification, distributed uncertainty estimation, and risk-controlled release. A single retrieval pass and a single generation call are insufficient for high-consequence domains because they cannot reliably distinguish between supported claims, uncertain claims, and unsupported claims. Multi-agent architectures can improve this distinction by separating retrieval, source evaluation, contradiction analysis, calibration, policy adjudication, and synthesis. However, this separation introduces new trade-offs involving latency, cost, coordination, accountability, and sustainability. These trade-offs cannot be resolved by technical optimization alone. They require governance frameworks that specify admissible evidence, escalation triggers, audit requirements, and human oversight.

The paper has argued that uncertainty should be treated as a first-class operational signal rather than a post hoc confidence score. Dynamic evidence verification should be embedded

throughout the pipeline, not appended as a final filter. Risk control should be adaptive to context and consequence, allowing low-risk queries to proceed efficiently while high-risk queries invoke deeper verification or human review. Fairness should be addressed through corpus composition, source selection, and policy design, not only through model debiasing. Sustainability should be considered through risk-adaptive computation, caching, and deployment choices that avoid unnecessary energy use. Policy implications extend to professional liability, consumer protection, public communication, and democratic governance. The future of retrieval-augmented generation will depend less on the size of the generator and more on the quality of the evidence system that surrounds it. Multi-agent collaboration offers a promising architecture for that system, but only if it is designed with humility, transparency, and institutional 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. 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.
3. 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.
4. 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), Article 248.
5. Zhang, R. (2025, December). Uncertainty-Aware RAG: Trustworthy Generation and Risk Control. In *2025 5th International Symposium on Artificial Intelligence and Big Data (AIBDF)* (pp. 526–531). IEEE.
6. Wooldridge, M. (2009). *An introduction to multiagent systems* (2nd ed.). Wiley.
7. Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (Article 2, pp. 1–22). Association for Computing Machinery.
8. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Awadallah, A., White, R. W., Burger, D., & Wang, C. (2023). AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv*. <https://arxiv.org/abs/2308.08155>
9. Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., Wang, J., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., & Schmidhuber, J. (2023). MetaGPT: Meta programming for a multi-agent collaborative framework. *arXiv*. <https://arxiv.org/abs/2308.00352>

10. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 8634–8652.
11. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. *International Conference on Learning Representations*.
12. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv*. <https://arxiv.org/abs/2312.10997>
13. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv*. <https://arxiv.org/abs/2310.11511>
14. Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., ... Kaplan, J. (2022). Language models (mostly) know what they know. *arXiv*. <https://arxiv.org/abs/2207.05221>
15. Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3214–3252). Association for Computational Linguistics.
16. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv*. <https://arxiv.org/abs/1606.06565>
17. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0). U.S. Department of Commerce.
18. Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1–21.
19. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1).
20. O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.