

Graph-Augmented Retrieval and Uncertainty Estimation for Explainable Scientific Question Answering

Christel Lochel

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

Tory Goggin

Department of Computer Science, University of Houston, Houston, TX, USA.
torywork@uh.edu

Lucinda McNeill

Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
lucindamcneill781@colostate.edu

Alejandro Sampson

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

Abstract

Scientific question answering increasingly relies on large language models and retrieval-augmented generation, but scientific domains demand more than fluent answers. They require provenance, calibrated uncertainty, and explanations that reflect evolving evidence. This paper examines graph-augmented retrieval and uncertainty estimation as complementary architectural strategies for explainable scientific question answering. It argues that knowledge graphs and structured scientific resources can improve retrieval precision, contextual grounding, and traceability, while uncertainty estimation can signal when an answer should be trusted, revised, or deferred. The paper develops a system-level analysis of trade-offs involving graph construction, hybrid retrieval, latency, storage, governance, and sustainability. It considers how graph augmentation expands explainability by exposing relational evidence paths but also introduces maintenance burdens, coverage gaps, and bias risks. It further examines uncertainty estimation methods such as verbalized confidence, calibration, and conformal prediction, noting that they must be integrated into user interfaces, auditing routines, and risk controls rather than treated as isolated model outputs. Through cross-domain comparisons in biomedicine, climate science, materials science, and policy research, the paper shows that scientific question answering is a socio-technical infrastructure. Successful deployment depends not only on model accuracy but also on institutional accountability, data stewardship, equitable access, and robust evaluation. The conclusion proposes design principles for graph-augmented, uncertainty-aware scientific question answering systems that balance performance with transparency, fairness, and long-term sustainability.

Keywords

scientific question answering; graph-augmented retrieval; uncertainty estimation; retrieval-augmented generation; explainable artificial intelligence; knowledge graphs; trustworthy systems; socio-technical infrastructure; research infrastructure; algorithmic governance.

1. Introduction

Scientific question answering has become a critical interface between computational systems and knowledge production. In domains such as biomedicine, climate science, materials science, and public policy, users do not simply seek a plausible sentence; they seek an answer that can be traced to evidence, evaluated against disciplinary standards, and revised when new findings emerge. Retrieval-augmented generation has emerged as a dominant architecture for grounding language models in external corpora, allowing systems to combine parametric knowledge with retrieved documents and thereby improve factual specificity in knowledge-intensive tasks [1]. Earlier retrieval-augmented pre-training and dense passage retrieval established that learned retrieval can substantially improve open-domain question answering by aligning queries with relevant passages rather than relying on lexical overlap alone [2][3]. Subsequent work on generative readers and large-scale retrieval demonstrated that models can synthesize evidence across multiple passages and very large corpora, but also that retrieval quality, corpus freshness, and source selection remain central determinants of answer reliability [4][5].

Despite these advances, scientific question answering remains vulnerable to hallucination, overconfidence, and opaque reasoning. Surveys of natural language generation and large language models have documented diverse hallucination patterns, including fabricated citations, incorrect numerical claims, conflated entities, and unsupported causal assertions [6][7]. In scientific contexts, such failures are not merely conversational inconveniences. They can mislead researchers, distort policy deliberation, or propagate into downstream publications. Recent work has begun to integrate uncertainty estimates directly into retrieval-augmented generation pipelines, treating risk control as a first-class objective rather than a post hoc evaluation metric [8]. This shift is important because scientific users often need to know not only what a system claims but also how confident it is, which evidence supports the claim, and what alternative interpretations remain plausible.

This paper argues that graph-augmented retrieval and uncertainty estimation should be understood as complementary components of explainable scientific question answering. Graph augmentation introduces structured relations among entities, concepts, measurements, methods, and publications, enabling retrieval to move beyond isolated passages toward evidence paths and contextual neighborhoods. Uncertainty estimation, in turn, provides a language for expressing epistemic limits, detecting distributional shift, and triggering human review or abstention. Together, these strategies can improve transparency, but they also create new system-level trade-offs. Graph construction is expensive, incomplete, and politically inflected; uncertainty signals can be miscalibrated, difficult to communicate, or strategically ignored. The central question is therefore not whether graph augmentation or uncertainty estimation improves benchmark performance in isolation, but how they reshape the architecture, governance, and social legitimacy of scientific knowledge infrastructures.

2. Foundations of Scientific Question Answering, Uncertainty, and Structured Knowledge

Scientific question answering differs from general open-domain question answering because its evidence base is heterogeneous, hierarchical, and constantly contested. A single query may

require interpretation of experimental protocols, statistical results, clinical guidelines, material properties, regulatory documents, and expert disagreement. Systems that retrieve only flat text passages may miss relational context, such as whether a drug interacts with a protein pathway, whether a climate model depends on a particular emissions scenario, or whether a materials finding has been replicated under different conditions. Uncertainty estimation is therefore foundational rather than peripheral. Research on language model self-knowledge has shown that models can sometimes recognize the limits of their knowledge, although such signals are uneven and sensitive to prompting and domain [9]. Verbalized uncertainty offers an intuitive interface for users, but it can be vague, inconsistent, or rhetorically overconfident unless constrained by training and evaluation [10].

Calibration research provides a more formal basis for understanding whether confidence scores correspond to empirical correctness. Modern neural networks are often poorly calibrated, especially under distribution shift or when trained on imbalanced scientific data [11]. Conformal prediction offers distribution-free uncertainty sets under exchangeability assumptions, but scientific evidence often violates those assumptions because papers are temporally ordered, studies are not independent, and publication bias shapes the available record [12]. These limitations do not eliminate the value of uncertainty estimation; rather, they imply that confidence must be contextualized by domain, provenance, and institutional review. A scientific question answering system should not present a single confidence number as an objective truth. It should expose the conditions under which the estimate was produced, the evidence that supports it, and the known failure modes that could undermine it.

Structured knowledge resources provide one route to richer grounding. Knowledge graphs represent entities and relations in a form that supports traversal, aggregation, and provenance tracking [13]. Surveys of knowledge graph representation and acquisition describe methods for extracting entities, relations, and ontologies from text, databases, and expert curation, while also emphasizing incompleteness and schema heterogeneity as persistent challenges [14]. Knowledge graph embeddings and related representation learning techniques can support link prediction and semantic similarity, but they often compress nuanced scientific relations into vector spaces that are difficult to interpret [15]. For scientific question answering, the most promising use of graphs may not be end-to-end embedding similarity alone. Instead, graphs can serve as scaffolding for retrieval, evidence composition, and explanation, allowing systems to connect documents, datasets, methods, and claims through explicit paths that users can inspect.

Evaluation foundations also matter. Benchmarks for knowledge-intensive language tasks and fact extraction and verification have advanced the measurement of retrieval, attribution, and claim support [16][17]. Yet many benchmarks simplify scientific reasoning by using static corpora, clean entity boundaries, and short evidence passages. They may not capture the temporal dynamics of scientific consensus, the prevalence of negative results, or the interpretive disagreements that characterize frontier research. A robust scientific question answering system must therefore be evaluated not only on answer accuracy but also on provenance quality, uncertainty calibration, abstention behavior, fairness across disciplines, and resilience to adversarial or outdated evidence.

3. Graph-Augmented Retrieval Architecture

A graph-augmented retrieval architecture typically combines conventional document retrieval with structured representations of entities, relations, concepts, and evidence provenance. In a scientific system, the graph may include citations, author networks, experimental methods,

chemical compounds, genes, proteins, climate variables, geographic regions, policy instruments, and measurement units. Retrieval can begin with lexical or dense matching over documents, then expand through graph neighbors to identify related studies, contradictory findings, or methodological caveats. This expansion can improve recall for long-tail queries and multi-hop questions, but it also increases the risk of topic drift, where graph traversal pulls in semantically related but evidentially irrelevant material. The design challenge is to balance exploratory reach with precision and to ensure that every expanded node contributes to an auditable explanation.

The architecture must also decide where graph structure is created and maintained. Automated extraction from scientific text can scale across large corpora, but it produces noisy relations, duplicate entities, and uncertain schema alignments. Expert-curated graphs offer higher precision but are expensive, unevenly distributed across fields, and slow to update. A hybrid approach may combine automated extraction, community curation, and versioned provenance, but this introduces governance questions about who controls the schema, who resolves conflicts, and how updates are audited. In biomedical domains, for example, curated pathways and ontologies coexist with rapidly growing literature, requiring continuous reconciliation. In climate and policy domains, graph relations may encode contested assumptions, such as causal links between emissions pathways and regional impacts. Graph augmentation therefore does not merely improve retrieval; it makes epistemic and institutional choices visible in system architecture.

From a systems perspective, graph-augmented retrieval changes latency, storage, and failure profiles. Graph traversal, entity linking, and relation ranking add computation and may create bottlenecks under high query loads. Caching can help for common queries, but scientific questions are often long-tailed and context-dependent. Graph storage may require specialized databases, version control, and access management, especially when proprietary or sensitive data are involved. Failure can occur at multiple layers: entity resolution may link the wrong gene, relation extraction may invert a causal claim, retrieval may miss a crucial negative result, or the generator may ignore graph evidence. Explainability depends on exposing these layers rather than presenting a single final answer. A graph-augmented system should therefore log retrieval paths, confidence at each stage, and the provenance of graph edges, enabling auditors and users to reconstruct how an answer was assembled.

4. Uncertainty Estimation and Risk Control

Uncertainty estimation in scientific question answering must address both aleatoric and epistemic uncertainty, even if the system does not use those terms explicitly. Aleatoric uncertainty arises from noise, missing measurements, or inherent variability in scientific phenomena. Epistemic uncertainty arises from incomplete knowledge, model limitations, or disagreement among sources. Retrieval-augmented systems add another layer: uncertainty about whether the retrieved evidence is relevant, current, and representative. Recent frameworks for trustworthy generation and risk control emphasize that uncertainty should be integrated into the generation pipeline, not appended as a superficial confidence score [8]. This integration can take several forms, including calibrated confidence, verbalized hedging, conformal prediction sets, and abstention thresholds. Each has trade-offs. Calibrated probabilities can support quantitative risk management but may be poorly calibrated under domain shift [11]. Verbalized uncertainty is user-friendly but can be manipulated or inconsistently expressed [10]. Conformal prediction can provide coverage guarantees under assumptions, but scientific evidence often violates exchangeability [12].

A key architectural question is where uncertainty is estimated and how it flows through the system. Retrieval uncertainty may reflect ambiguity in the query, sparse coverage in the corpus, or disagreement among retrieved passages. Graph uncertainty may reflect missing edges, low-confidence relation extraction, or conflicting provenance. Generation uncertainty may reflect model confidence, evidence sufficiency, or internal inconsistency. If these signals are aggregated into a single number, important distinctions are lost. A better approach is to expose a structured uncertainty profile: what is known, what is inferred, what is contested, and what requires expert review. This profile can drive user interface design, such as highlighting unsupported claims, showing alternative evidence paths, or recommending abstention when confidence falls below a domain-specific threshold.

Risk control also requires operational policies. In clinical or regulatory settings, an uncertain answer may need to be routed to a human expert, while in exploratory research it may be presented as a hypothesis with clear caveats. The same confidence level can imply different actions depending on context, consequence, and institutional tolerance for error. Uncertainty estimation therefore cannot be separated from governance. It must be connected to audit logs, escalation procedures, and accountability structures. If a system is deployed in a high-stakes environment, it should document how uncertainty thresholds were chosen, how often abstention occurs, and whether error rates differ across disciplines, languages, or user groups. Without such documentation, uncertainty estimation risks becoming a technical ritual that legitimizes opaque decisions rather than improving them.

5. Explainability, Provenance, and Scientific Trust

Explainability in scientific question answering is not equivalent to generic model interpretability. Scientists need explanations that connect claims to evidence, methods, and uncertainty. Graph-augmented retrieval can support this by exposing relational paths, such as a chain from a clinical trial to a drug, a dosage, a patient population, and an outcome. These paths can be rendered as evidence maps, citation trails, or structured summaries. Knowledge graph research emphasizes that graph structure can support provenance and provenance-aware reasoning, but also that graphs are social artifacts shaped by inclusion and exclusion decisions [13][14]. An explanation that appears objective because it is graph-based may hide gaps, biases, or contested relations. Explainability must therefore include not only the path that was followed but also the paths that were pruned, the sources that were unavailable, and the assumptions embedded in entity resolution.

Provenance also matters for temporal scientific trust. Scientific knowledge changes as new studies replicate, revise, or refute earlier findings. A static graph or corpus can become outdated, and a system that confidently retrieves an old result without temporal context may mislead users. Graph-augmented systems can represent temporal validity, versioning, and retraction status, but doing so requires metadata discipline and coordination across publishers, databases, and institutions. Benchmarks for fact verification and knowledge-intensive tasks provide useful starting points for measuring attribution and support [16][17], but they often treat evidence as stable. Scientific explainability should instead communicate the evidential status of a claim: established, preliminary, contested, retracted, or hypothetical. This is a design and governance problem as much as a modeling problem.

Trust is also relational. Users may trust a system because it cites sources, because it is affiliated with a reputable institution, or because it aligns with prior beliefs. Explainability can support warranted trust, but it can also create an illusion of rigor if explanations are verbose, technical, or selectively presented. A graph-augmented interface that shows complex evidence

paths may overwhelm non-specialists while appearing authoritative to specialists. Effective explainability should be layered: a concise answer with uncertainty, a middle layer of key evidence and caveats, and a deeper layer of graph paths, source metadata, and audit logs. This layered approach allows different users to engage at the level of detail appropriate to their expertise and decision context.

6. Governance, Fairness, and Policy Implications

Graph-augmented, uncertainty-aware scientific question answering is a socio-technical infrastructure, not merely a model. It depends on data access policies, copyright regimes, institutional repositories, cloud infrastructure, and expert labor. Governance must address who can contribute to the graph, who can query it, who can audit it, and who bears responsibility when an answer causes harm. Foundation models already raise broad questions about concentration of power, environmental cost, and downstream risks [18]. Taxonomies of language model risks identify harms such as misinformation, bias, privacy violation, and unequal access [19]. In scientific question answering, these risks intersect with existing inequalities in research funding, publication access, and disciplinary representation. A graph built primarily from English-language, high-impact journals may systematically underrepresent low-resource languages, local knowledge, and negative results.

Fairness considerations are therefore both technical and structural. Technical fairness may involve evaluating retrieval and uncertainty calibration across subfields, institutions, and user populations. Structural fairness requires attention to whose data are included, whose questions are prioritized, and whose labor maintains the knowledge base. Algorithmic auditing frameworks emphasize end-to-end accountability, including documentation, testing, and redress [20]. For scientific question answering, auditing should examine not only answer accuracy but also whether uncertainty estimates are equally reliable across domains, whether graph expansion amplifies dominant paradigms, and whether abstention disproportionately affects marginalized research communities. Policy implications extend to research integrity, peer review, and public communication. If AI systems become routine intermediaries for scientific evidence, their design choices will shape what counts as a valid question and a credible answer.

Governance also needs to address transparency without creating new vulnerabilities. Full graph and model transparency may expose proprietary data, personal information, or dual-use knowledge. Yet excessive secrecy undermines accountability. A workable approach may involve tiered access, independent audit trails, and standardized disclosure of evaluation results, uncertainty thresholds, and known limitations. Regulatory frameworks for AI in high-stakes domains often emphasize risk management, human oversight, and documentation. Scientific question answering could benefit from similar discipline, but it also requires domain-specific governance. Biomedical, climate, and materials communities have different norms, data-sharing practices, and risk profiles. A single global policy is unlikely to fit all. Instead, systems should support configurable governance layers that allow institutions to set retrieval boundaries, uncertainty thresholds, and audit requirements while preserving interoperability.

7. Deployment, Sustainability, and Operational Trade-Offs

Deployment of graph-augmented retrieval and uncertainty estimation introduces significant operational costs. Graph construction, entity linking, relation extraction, and continuous updates require computational resources and human expertise. Dense retrieval and large

language model inference add further energy and latency demands, especially when systems perform multi-hop reasoning or generate long explanations. These costs must be weighed against benefits such as improved accuracy, reduced hallucination, and better user decision-making. In many scientific settings, a slower, more transparent system may be preferable to a fast but opaque one. However, latency can also affect equity: users with slower internet connections or older devices may be unable to access resource-intensive interfaces. Sustainability therefore involves not only energy efficiency but also inclusive access and long-term maintainability.

Maintenance is another critical trade-off. Scientific knowledge evolves, journals retract papers, ontologies change, and new measurement standards emerge. A graph-augmented system that is not continuously maintained will degrade. Maintenance requires institutional commitment, clear ownership, and sustainable funding. Open-source communities can contribute, but volunteer labor is uneven and may not cover specialized domains. Public-private partnerships may provide resources but can introduce conflicts of interest. A sustainable architecture should separate core retrieval and uncertainty components from domain-specific graph modules, allowing updates without rebuilding the entire system. It should also support graceful degradation, where the system can still provide useful answers with lower confidence when graph services are unavailable. Operational resilience is thus closely tied to explainability: users should know when the system is operating with incomplete or stale knowledge.

8. Evaluation and Robustness

Evaluation of graph-augmented, uncertainty-aware scientific question answering must extend beyond standard accuracy metrics. Benchmarks for knowledge-intensive tasks and fact verification provide valuable foundations for measuring retrieval recall, answer correctness, and evidence support [16][17]. However, they often assume stable evidence and well-formed queries. Scientific evaluation should include temporal robustness, adversarial queries, distribution shift across subfields, and calibration under uncertainty. It should also measure explanation quality: whether users can identify the evidence behind a claim, understand its limitations, and recognize when the system is uncertain. Human evaluation remains essential because scientific reasoning involves domain expertise, interpretive judgment, and awareness of disciplinary norms.

Robustness testing should probe failure modes at each layer. Retrieval tests can examine whether graph expansion introduces irrelevant evidence or misses contradictory findings. Uncertainty tests can assess whether confidence scores remain calibrated when the corpus is updated, when queries use new terminology, or when evidence is intentionally misleading. Governance tests can review whether audit logs are complete, whether abstention policies are followed, and whether error rates differ across user groups. Because scientific question answering is often deployed in high-stakes contexts, robustness cannot be treated as a one-time certification. It requires continuous monitoring, incident reporting, and iterative improvement. A system that performs well on a static benchmark but fails silently in practice is not trustworthy, even if its average accuracy is high.

9. Conclusion

Graph-augmented retrieval and uncertainty estimation offer a promising path toward explainable scientific question answering, but they are not plug-in enhancements. They reshape the architecture, economics, and governance of knowledge systems. Graph

augmentation can improve grounding and provenance by exposing relational evidence, but it introduces maintenance burdens, coverage gaps, and epistemic choices that must be governed. Uncertainty estimation can support risk control and user trust, but it requires calibration, domain-sensitive thresholds, and integration with human review. Together, these strategies can help scientific question answering move beyond fluent but opaque answers toward systems that communicate evidence, limitations, and confidence in ways that support responsible decision-making.

The broader lesson is that scientific question answering is a socio-technical infrastructure. Its legitimacy depends on fairness, accountability, sustainability, and inclusion as much as on model performance. Future work should develop interoperable graph schemas, uncertainty standards, and audit protocols that can travel across disciplines without erasing local norms. It should also examine how users interpret uncertainty and graph-based explanations in practice, especially when evidence is contested or incomplete. If designed with care, graph-augmented and uncertainty-aware systems can strengthen scientific communication rather than merely automating it. If designed without governance, they risk amplifying existing inequalities and creating new forms of unaccountable authority over what counts as scientific knowledge.

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. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning* (Vol. 119, pp. 3929–3938). PMLR.
3. 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.
4. 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.
5. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J. B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., ... Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning* (Vol. 162, pp. 2206–2240). PMLR.
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.
7. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv*. <https://arxiv.org/abs/2311.05232>

8. 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.
9. 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>
10. Lin, S., Hilton, J., & Evans, O. (2022). Teaching models to express their uncertainty in words. arXiv. <https://arxiv.org/abs/2205.14334>
11. 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 (Vol. 70, pp. 1321–1330). PMLR.
12. Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv. <https://arxiv.org/abs/2107.07511>
13. Hogan, A., Blomqvist, E., Cochez, M., d’Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Labra Gayo, J. E., Navigli, R., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys*, 54(4), 1–37.
14. Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2021). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2), 494–514.
15. Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12), 2724–2743.
16. Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rocktäschel, T., & Riedel, S. (2021). KILT: A benchmark for knowledge intensive language tasks. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 2523–2544). Association for Computational Linguistics.
17. 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.
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. <https://arxiv.org/abs/2108.07258>
19. 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). Association for Computing Machinery.

20. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33-44). Association for Computing Machinery.