

Retrieval-Augmented Multimodal Language Models for Reducing Factual Hallucinations

Mieszko Kinal

School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
mieszkok@uc.edu

Antonie Drubin

Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV,
USA.
drubin1972@unr.edu

Abstract

Retrieval-augmented multimodal language models promise to reduce factual hallucinations by coupling generative inference with external evidence drawn from text, images, audio, video, structured databases, and domain-specific repositories. Yet the integration of retrieval into multimodal generation is not a simple corrective layer. It reshapes model architecture, data governance, latency budgets, storage infrastructure, evaluation practices, and organizational accountability. This paper presents a system-level analysis of retrieval-augmented multimodal language models as socio-technical infrastructures for factual grounding. It examines how cross-modal indexing, query formulation, evidence ranking, fusion, attribution, and refusal mechanisms interact with the probabilistic tendencies of large generative models. The discussion emphasizes structural trade-offs among accuracy, coverage, latency, cost, privacy, fairness, and sustainability. It also considers robustness under distribution shift, adversarial retrieval poisoning, stale knowledge, modality conflict, and degraded network conditions. Governance implications are analyzed through documentation regimes, audit trails, content provenance, access control, and policy constraints. Case illustrations from clinical decision support, public-sector information services, industrial maintenance, journalism verification, and educational tutoring demonstrate that retrieval augmentation is most effective when treated as an operational capability rather than a one-time model modification. The paper concludes that reducing factual hallucinations requires an integrated architecture of evidence acquisition, multimodal alignment, uncertainty communication, human oversight, and continuous evaluation. Without such integration, retrieval can create an illusion of grounding while displacing hallucination from generation to retrieval, ranking, or governance failures.

Keywords

retrieval-augmented generation; multimodal language models; factual hallucination; grounding; information retrieval; governance; socio-technical systems; robustness; fairness; deployment.

1. Introduction

Large language models have achieved remarkable fluency across diverse tasks, but their generative confidence often exceeds their factual reliability. Hallucination, understood as fluent output that is unsupported by evidence or contrary to available knowledge, remains a persistent obstacle to high-stakes deployment [3]. Retrieval-augmented generation has emerged as a pragmatic architectural response, combining parametric memory with non-

parametric evidence retrieved at inference time [1]. In text-only settings, this approach can improve factuality, enable attribution, and support domain adaptation without retraining the entire model. In multimodal settings, however, the problem becomes substantially more complex because evidence may be visual, spatial, temporal, or embodied, and because grounding requires alignment across heterogeneous representations [9], [11].

The rise of multimodal language models has expanded the scope of factual error. A system may correctly identify an object in an image but misstate its function, provenance, or safety implications. It may transcribe a chart accurately while misinterpreting its trend. It may describe a medical scan using plausible but unsupported terminology. These failures are not merely linguistic; they arise from the interaction among perception, retrieval, reasoning, and generation. Recent work on attention dynamics and token suppression in large vision-language models illustrates that hallucination can be influenced by internal representational mechanisms as well as external evidence [5]. This finding reinforces the need for system-level interventions that address retrieval quality, multimodal fusion, and governance rather than treating hallucination as a single model defect.

This paper argues that retrieval-augmented multimodal language models should be analyzed as socio-technical infrastructures. Their factual reliability depends not only on model weights or retrieval algorithms but also on the institutions that curate evidence, the policies that determine access, the interfaces that communicate uncertainty, and the operational practices that maintain indexes over time. The discussion is organized around architecture, retrieval quality, trade-offs, robustness, fairness, governance, deployment, evaluation, and cross-domain cases. The central claim is that retrieval can reduce factual hallucinations only when evidence acquisition, multimodal alignment, uncertainty representation, and accountability mechanisms are designed together.

2. Conceptual Foundations and Problem Definition

Factual hallucination in multimodal language models cannot be reduced to a single causal pathway. It may originate in pretraining data gaps, visual encoding errors, cross-modal misalignment, retrieval failure, ranking bias, context-window truncation, or overly permissive decoding. A model may retrieve a relevant document but ignore it, retrieve an irrelevant image but treat it as authoritative, or synthesize evidence from incompatible modalities without acknowledging conflict. Consequently, evaluation must distinguish among unsupported generation, contradicted generation, and underdetermined generation [15], [16]. The first occurs when no evidence supports the claim, the second when evidence contradicts it, and the third when available evidence is insufficient for a definitive answer.

Retrieval augmentation introduces a division of labor between parametric and non-parametric memory. Parametric memory supports generalization, style, and reasoning patterns, while non-parametric memory supplies updatable, inspectable, and domain-specific evidence [1], [6]. Dense retrieval methods can match semantic intent beyond lexical overlap, but they also inherit embedding biases and may fail on rare entities or specialized notation [7]. Multimodal encoders extend this logic to images, video, and audio, yet they often compress rich perceptual information into embeddings that privilege certain visual cultures or captioning conventions [8], [10]. The resulting system is not a neutral pipeline; it is a set of representational choices that shape what counts as evidence.

A second conceptual issue is grounding. In text-only retrieval, grounding often means that an answer can be traced to a retrieved passage. In multimodal retrieval, grounding must specify

whether the support is textual, visual, spatial, temporal, or relational, and how those forms of support interact. A retrieved image may corroborate the presence of an object but not its ownership. A video may show an event but not its cause. A table may provide a value but not its uncertainty. Therefore, factual grounding requires an explicit evidence model that records provenance, modality, time, jurisdiction, and confidence. Without such a model, retrieval may merely add more plausible context to an already fluent hallucination.

3. Architecture of Retrieval-Augmented Multimodal Systems

A retrieval-augmented multimodal language model typically includes several interacting layers: ingestion and preprocessing, multimodal indexing, query formulation, retrieval and ranking, evidence fusion, generation, attribution, and post-generation verification. Each layer introduces failure modes and governance obligations. Ingestion determines what enters the knowledge base and how it is normalized. Images may require object detection, optical character recognition, captioning, geospatial tagging, or de-identification. Audio may require transcription, speaker separation, and temporal alignment. Video may require keyframe extraction, scene segmentation, and event annotation. These processes are not merely technical; they encode institutional priorities about which evidence is visible and which is excluded [17].

Indexing is equally consequential. A unified vector index can support cross-modal similarity, but it may obscure modality-specific structure. Separate indexes for text, images, tables, and time series can preserve specialized retrieval logic, but they require a fusion layer that reconciles rankings across spaces. Hybrid architectures often combine sparse lexical retrieval, dense semantic retrieval, graph traversal, and rule-based filters. Such designs improve recall and precision in complex domains, yet they increase operational complexity and make end-to-end auditing more difficult [12]. The system must also decide whether to retrieve before generation, during generation, or after generation as a verification step. Pre-generation retrieval provides context, mid-generation retrieval can correct drifting reasoning, and post-generation retrieval can flag unsupported claims [14].

Evidence fusion is the point at which multimodal grounding becomes most fragile. Simply concatenating retrieved passages and images into a context window does not guarantee that the model will privilege reliable evidence. The model may attend to visually salient but irrelevant content, overweight captions, or ignore tabular constraints. Attention-head analyses and token-level interventions suggest that internal routing mechanisms can suppress or amplify hallucinated continuations [5]. System architects therefore need fusion strategies that encode source reliability, recency, jurisdiction, and semantic role. Attribution mechanisms should expose which evidence supports which part of the output, and refusal mechanisms should activate when evidence is conflicting or insufficient. The architecture must treat abstention as a valid outcome rather than a failure of user experience.

4. Retrieval Quality and Multimodal Grounding

Retrieval quality is the foundation of factual reliability. If the retriever returns irrelevant or outdated evidence, generation may become confidently wrong. If it returns only popular or easily indexed sources, the system may reproduce structural inequalities in knowledge production [4]. In multimodal settings, retrieval quality depends on query formulation across modalities. A user may ask a question about an image, but the system must translate visual content into retrievable concepts. This translation can lose spatial relations, subtle context, or cultural meaning. Conversely, a text query may require visual evidence that is difficult to

describe in words. Cross-modal retrieval therefore demands robust alignment between language and perception, not merely a shared embedding space [8], [10].

Grounding also requires temporal reasoning. A retrieved document from an earlier period may be accurate for that period but misleading for the present. A medical guideline may have been superseded. A map may no longer reflect political boundaries. A product manual may describe an older model. Retrieval-augmented systems must track validity intervals, version histories, and jurisdiction-specific applicability. They must also distinguish primary evidence from secondary commentary, official records from user-generated content, and verified data from speculation. These distinctions are essential in domains such as public health, law, finance, and engineering, where factual errors can cause material harm [19].

Another challenge is evidence conflict. Multimodal systems may retrieve a text passage that contradicts an image, or two authoritative sources may disagree because they use different definitions or measurement periods. A robust system should not average conflicting evidence into a seamless but false narrative. It should identify the conflict, explain the source of disagreement, and either present multiple positions or abstain. This behavior requires an evidence representation that includes provenance, confidence, and scope conditions. Without such representation, retrieval augmentation can create an epistemic illusion: the answer appears grounded because citations are present, even though the citations do not logically support the claim.

5. System-Level Trade-offs and Infrastructure

Retrieval-augmented multimodal systems operate under severe trade-offs. Higher retrieval depth improves recall but increases latency and cost. More sophisticated ranking improves precision but may reduce coverage of minority perspectives. Larger multimodal indexes improve domain specificity but demand substantial storage, energy, and cooling infrastructure. Frequent index updates improve recency but complicate reproducibility and auditability. These trade-offs are not merely engineering details; they determine who can deploy such systems, under what conditions, and with what environmental footprint. A hospital, a newsroom, and a municipal agency may all need factual grounding, but their latency tolerances, privacy constraints, and budget envelopes differ sharply.

Infrastructure choices also shape governance. Cloud-based retrieval offers scalability but may concentrate sensitive data in external jurisdictions. On-premises retrieval supports data sovereignty but may limit access to state-of-the-art models. Federated retrieval can preserve local control while enabling cross-institutional search, yet it raises complex questions about authentication, query privacy, and result aggregation. Edge deployment can reduce latency and bandwidth dependence, but it constrains index size and model capacity. These architectural decisions influence fairness because under-resourced organizations may be unable to maintain high-quality evidence repositories or continuous evaluation pipelines [18], [20].

Sustainability is an often-overlooked dimension. Multimodal retrieval requires encoding images, audio, and video at scale, maintaining duplicate indexes, and serving frequent queries. If every user request triggers broad retrieval across multiple modalities, energy consumption can grow rapidly. More efficient architectures may use cached evidence, tiered indexes, query routing, and modality-specific retrieval only when necessary. However, efficiency gains must not come at the cost of factual reliability. The system should optimize for grounded utility per

unit of computation, not raw retrieval volume. This requires lifecycle assessment that includes data ingestion, model training, index maintenance, inference, and eventual decommissioning.

6. Robustness, Fairness, and Governance

Robustness in retrieval-augmented multimodal systems must address adversarial and non-adversarial threats. Retrieval poisoning occurs when malicious or corrupted content enters the index and is ranked highly. Query manipulation can steer retrieval toward unreliable sources. Modality spoofing can introduce manipulated images, audio, or video that appear authoritative. Distribution shift can degrade retrieval when language, visual conventions, or domain terminology changes. A system that performs well on benchmark data may fail in a new clinical setting, a different legal jurisdiction, or a region with different media ecosystems [20]. Robustness therefore requires continuous monitoring, anomaly detection, source reputation tracking, and rapid index rollback capabilities.

Fairness concerns arise at multiple points. Retrieval may underrepresent certain languages, dialects, regions, disabilities, or cultural practices. Visual encoders may perform unevenly across skin tones, lighting conditions, or clothing styles. Text corpora may privilege dominant institutions and marginalize community knowledge. If these biases enter the evidence base, retrieval augmentation can amplify them by presenting biased claims with citations. Fairness therefore cannot be reduced to model fine-tuning; it requires inclusive data collection, participatory indexing, multilingual retrieval, and disaggregated evaluation [4], [17]. It also requires transparency about what evidence is absent, not only what is retrieved.

Governance mechanisms should operationalize accountability across the retrieval lifecycle. Datasheets for datasets and model cards provide documentation templates that can be extended to retrieval indexes, ranking policies, and multimodal evidence sources [17], [18]. Audit trails should record queries, retrieved items, ranking scores, fusion decisions, generated claims, and human overrides. Content provenance standards can help verify the origin and integrity of media. Access control policies should specify who may retrieve sensitive records and under what legal basis. Finally, redress mechanisms should allow affected individuals and communities to contest erroneous outputs and request corrections [19]. Governance is not a bureaucratic add-on; it is a condition for trustworthy factual grounding.

7. Deployment, Sustainability, and Policy Implications

Deployment introduces pressures that can erode factual safeguards. Product teams may prioritize response speed, engagement, or cost reduction over abstention and verification. Users may treat fluent answers as authoritative even when citations are present. Organizations may lack the expertise to maintain domain-specific indexes. Policy responses therefore need to address both transparency and capability. Regulations that require disclosure of automated decision-making, provenance of synthetic media, and human oversight can create incentives for better retrieval governance. However, overly rigid rules may freeze architectures before best practices mature. A more effective approach combines risk-based obligations, sector-specific standards, and safe harbors for documented auditing [20].

Public-sector deployment deserves particular attention. Governments may use multimodal language models to answer questions about benefits, permits, public health, or emergency response. In such settings, factual hallucination can deny services, misallocate resources, or undermine civic trust. Retrieval from official records can improve accuracy, but it also requires clear chains of authority. If a model retrieves a draft policy rather than a final regulation, or an outdated webpage rather than a current statute, the result may be legally

misleading. Policy frameworks should therefore require source hierarchies, version control, and explicit uncertainty statements for public-facing systems. They should also protect against surveillance risks when retrieval systems aggregate personal data across modalities.

Sustainability policy is intertwined with deployment scale. If retrieval augmentation becomes the default for every multimodal query, energy and bandwidth demands will rise. Policymakers and system designers should encourage tiered architectures that retrieve only when necessary, cache high-value evidence, and use smaller models for routine verification. They should also require reporting on computational intensity and data retention. Sustainability is not opposed to factual reliability; inefficient retrieval can be both environmentally costly and epistemically noisy. The goal is disciplined retrieval that improves grounding without indiscriminate expansion.

8. Evaluation, Auditing, and Operational Metrics

Evaluating retrieval-augmented multimodal models requires metrics beyond answer accuracy. System operators need to measure retrieval recall, ranking precision, evidence sufficiency, attribution correctness, conflict detection, abstention calibration, and robustness under distribution shift [15], [16]. Multimodal evaluation should test whether the model correctly integrates text, image, audio, and tabular evidence, and whether it appropriately distinguishes correlation from causation. It should also assess whether citations actually support the generated claims rather than merely accompanying them. Human evaluation remains essential because automated metrics may reward fluent but unfaithful outputs.

Auditing should be continuous rather than episodic. Retrieval indexes drift as documents are added, deleted, or modified. Models are updated, user populations change, and adversarial actors adapt. A static audit at deployment cannot capture these dynamics. Continuous auditing can use canary queries, shadow retrieval, counterfactual testing, and community feedback. It should also examine disaggregated performance across languages, regions, disabilities, and demographic groups. When failures are detected, the system should support rollback, re-ranking, index correction, and user notification [18]. Operational metrics should be tied to governance thresholds, so that degraded factual reliability triggers review rather than silent continuation.

9. Case Illustrations and Cross-Domain Comparisons

In clinical decision support, retrieval-augmented multimodal systems might combine patient records, imaging, laboratory results, and medical literature. The stakes are high, and hallucination can cause direct harm. Effective systems require strict access control, de-identification, evidence hierarchies, and clinician oversight. Retrieval can reduce errors by surfacing guidelines and contraindications, but it must not override clinical judgment or obscure uncertainty. In industrial maintenance, multimodal retrieval might combine equipment manuals, sensor data, images of corrosion, and repair histories. Here, factual errors can cause safety incidents or downtime. Grounding requires version-specific documentation and spatial alignment between diagrams and physical components.

In journalism and public verification, retrieval augmentation can help identify manipulated media, trace claims to sources, and assemble context. However, it may also amplify rumor if retrieval prioritizes engagement metrics or unverified social media. Newsroom systems need provenance verification, editorial oversight, and transparent correction workflows. In education, multimodal tutors may retrieve historical maps, scientific diagrams, and primary sources. Factual hallucination can entrench misconceptions, especially for young learners.

Retrieval can support inquiry-based learning, but the system must distinguish established knowledge from contested interpretation. Across these domains, the common lesson is that retrieval quality, governance, and user interface design determine whether augmentation reduces or merely relocates hallucination.

10. Future Directions

Future research should develop evidence-aware architectures that represent modality, provenance, time, jurisdiction, and confidence as first-class variables. This may require new retrieval models that can reason over conflicting sources and new fusion mechanisms that allow language models to cite specific visual regions or temporal segments. Another direction is uncertainty communication. Instead of returning a single confident answer, systems might present evidence maps, confidence intervals, or alternative interpretations. Such interfaces could help users calibrate trust and detect unsupported claims. Research should also examine federated and privacy-preserving retrieval, especially in health, finance, and public administration.

Finally, the field needs stronger socio-technical evaluation. Benchmarks should include realistic infrastructure constraints, adversarial conditions, multilingual and multicultural evidence, and long-term maintenance scenarios. Studies should examine how organizations actually use retrieval-augmented systems, how they interpret citations, and how they handle corrections. Interdisciplinary collaboration among computer scientists, information scientists, legal scholars, ethicists, and domain experts is essential. The reduction of factual hallucinations is not a purely technical objective; it is an institutional commitment to evidence, transparency, and accountability.

11. Conclusion

Retrieval-augmented multimodal language models offer a credible path toward reducing factual hallucinations, but only when retrieval is treated as part of a broader system of evidence governance. The architecture must align multimodal indexing, query formulation, ranking, fusion, attribution, and abstention. The infrastructure must balance accuracy, latency, cost, privacy, fairness, and sustainability. The governance must provide documentation, audit trails, provenance, access control, and redress. Evaluation must measure not only whether answers are fluent or accurate but whether they are grounded, attributable, and appropriately uncertain. Case illustrations from healthcare, industry, journalism, and education show that the same technical capability can support or undermine trust depending on institutional design.

The central conclusion is that retrieval does not automatically eliminate hallucination. It can reduce unsupported generation when evidence is relevant, reliable, and correctly integrated. It can also create a new form of epistemic risk when irrelevant evidence is cited, conflicts are hidden, or indexes are biased. Therefore, the responsible development of retrieval-augmented multimodal language models requires continuous attention to the entire socio-technical pipeline. Factual reliability is not a feature that can be added at the end; it is an ongoing property of systems that connect models, data, institutions, and users in accountable ways.

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. In *Advances in Neural Information Processing Systems* 33 (pp. 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. In *Advances in Neural Information Processing Systems* 30 (pp. 5998-6008).
3. 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. <https://doi.org/10.1145/3571730>
4. 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). <https://doi.org/10.1145/3442188.3445922>
5. Chen, Y., Li, J., Yao, W., Ma, X., Du, G., Wang, W., & Li, J. (2026, July). Vocabulary hijacking in lnlms: Unveiling critical attention heads by excluding inert tokens to mitigate hallucination. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 38458-38485).
6. 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* (pp. 3929-3938).
7. 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). <https://doi.org/10.18653/v1/2020.emnlp-main.550>
8. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8748-8763).
9. Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., ... Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems* 35 (pp. 23716-23736).
10. Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 19730-19742).
11. Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y., & Scialom, T. (2023). Augmented language models: A survey. *Transactions on Machine Learning Research*. arXiv:2302.07842
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:2312.10997
13. Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., Schulman, J., & Kingdom, H. (2021). WebGPT: Browser-assisted question-answering with human feedback. arXiv:2112.09332

14. Menick, J., Trebacz, M., Mikulik, V., Aslanides, J., Song, F., Chadwick, M., Glaese, M., Young, S., Campbell-Gillingham, L., Irving, G., & McAleese, N. (2022). Teaching language models to support answers with verified quotes. arXiv:2203.11147
15. 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). <https://doi.org/10.18653/v1/N18-1074>
16. 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). <https://doi.org/10.18653/v1/2022.acl-long.229>
17. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92. <https://doi.org/10.1145/3458723>
18. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220-229). <https://doi.org/10.1145/3287560.3287596>
19. 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:2108.07258
20. 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). <https://doi.org/10.1145/3531146.3533088>