

Agentic Retrieval-Augmented Generation for Autonomous Research and Evidence-Grounded Decision Making

Chameli Golla

School of Computing, Clemson University, Clemson, SC, USA.
chameli632@clemson.edu

Fintan Galvin

Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV,
USA.
fintangalvin299@unr.edu

Tai Hoskins

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

Abstract

Agentic retrieval-augmented generation is emerging as a system-level paradigm for combining large language models with dynamic retrieval, tool use, memory, and autonomous control. Unlike conventional retrieval-augmented generation, which primarily conditions generation on retrieved passages, agentic systems plan searches, select sources, critique evidence, revise queries, and coordinate multiple tools across extended tasks. This paper examines the architectural, infrastructural, governance, and policy dimensions of agentic retrieval-augmented generation for autonomous research and evidence-grounded decision making. It argues that the central challenge is not simply improving retrieval accuracy or model capability, but designing socio-technical infrastructures that make autonomous inquiry traceable, contestable, robust, and sustainable. The analysis begins with conceptual foundations, distinguishing between retrieval as a support mechanism and retrieval as an agentic capability. It then examines structural trade-offs involving latency, precision, source heterogeneity, memory, verification, and human oversight. The paper further considers autonomy and evidence in research workflows, emphasizing uncertainty communication, provenance, and the risk of overconfident synthesis. Governance and fairness are analyzed through auditability, accountability, representation, and contestability, while deployment concerns include compute intensity, data freshness, technical debt, and long-term maintainability. Robustness and evaluation are framed as multidimensional requirements spanning groundedness, calibration, resilience, fairness, and policy compliance. The paper concludes that agentic retrieval-augmented generation can support more transparent and evidence-grounded decision making only if autonomy is bounded by institutional controls, provenance infrastructure, and continuous evaluation.

Keywords

agentic retrieval-augmented generation, autonomous research, evidence-grounded decision making, AI governance, socio-technical systems, retrieval infrastructure, accountability, robustness, fairness, sustainability.

1. Introduction

Retrieval-augmented generation has become a foundational approach for connecting large language models to external knowledge sources, reducing reliance on purely parametric memory, and improving factual grounding in knowledge-intensive tasks [1]. Early systems demonstrated that dense retrieval could be integrated with generative models to answer open-domain questions more effectively [4], while pretraining with retrieval showed that language models could learn to attend to external evidence during training rather than only during inference [5]. Subsequent work extended retrieval augmentation to large-scale corpora and open-domain generation, establishing retrieval not merely as a preprocessing step but as a core component of model behavior [2], [3]. These developments created the technical basis for systems that can retrieve, synthesize, and cite evidence in real time.

However, conventional retrieval-augmented generation remains largely reactive. A model receives a query, retrieves a set of passages, and generates an answer. This pattern is insufficient for autonomous research, where goals are often ambiguous, evidence is distributed across heterogeneous sources, and intermediate findings reshape the direction of inquiry. Agentic retrieval-augmented generation addresses this limitation by treating retrieval as one capability within a broader control loop. In such systems, language models may plan, decompose tasks, select tools, issue multiple queries, evaluate retrieved evidence, revise search strategies, and decide when additional verification is required [7], [8], [9]. This shift from retrieval as augmentation to retrieval as agency introduces new opportunities for discovery and decision support, but it also raises difficult questions about reliability, governance, and infrastructure.

The movement toward autonomous agents has been reinforced by work on language models that use tools, reflect on failures, and simulate extended behavior [8], [9], [11]. Surveys of autonomous agents describe architectures that combine planning, memory, action, and observation, often with retrieval as a central mechanism for grounding [10]. Yet the literature remains fragmented across model-centric, agent-centric, and governance-centric perspectives. A system-level account is needed because agentic retrieval-augmented generation is not merely a model architecture. It is an assembly of data pipelines, retrieval indexes, orchestration layers, evaluation regimes, institutional policies, and human oversight practices. The central argument of this paper is that autonomous research and evidence-grounded decision making depend less on isolated model improvements than on the design of trustworthy socio-technical infrastructures.

2. Conceptual Foundations of Agentic Retrieval-Augmented Generation

Agentic retrieval-augmented generation can be understood as the integration of retrieval mechanisms into an autonomous or semi-autonomous control loop that pursues a goal over time. In conventional retrieval-augmented generation, the retrieval step is usually triggered by a user query and remains bounded by the generation episode. In agentic systems, retrieval becomes recursive and strategic. The system may generate sub-questions, choose among multiple corpora, evaluate source credibility, compare conflicting claims, and decide whether to retrieve more evidence or to stop. This recursive character is what distinguishes agentic retrieval from simple query expansion. It also makes the system more capable of handling complex research tasks, such as literature synthesis, regulatory analysis, and multidisciplinary problem framing.

A useful conceptual distinction is between retrieval as memory access and retrieval as inquiry. In the first case, retrieval supplements a model with facts that it lacks. In the second, retrieval is part of a deliberative process in which the agent formulates hypotheses, seeks disconfirming evidence, and updates its internal state. Work on reasoning and acting in language models illustrates how interleaving thought and action can improve task performance and interpretability [7]. Tool use further expands the action space by allowing models to query databases, calculators, search engines, or specialized application programming interfaces [8]. Reflection mechanisms add a layer of self-correction by enabling agents to critique their own outputs and revise strategies [9]. Together, these capabilities suggest that agentic retrieval-augmented generation is not a single technique but a family of architectures organized around iterative evidence seeking.

Heterogeneous source integration is a defining challenge. Research and decision-making environments rarely rely on a single corpus. They involve scientific papers, institutional reports, legal texts, structured databases, sensor feeds, and expert commentary. A unified semantic space and attention-based fusion framework for multi-source heterogeneous retrieval-augmented generation demonstrates the need to align representations across source types while preserving source-specific signals during fusion [6]. This is important because naive concatenation of retrieved passages can obscure provenance, dilute relevance, and increase the risk of contradictory synthesis. Agentic systems must therefore manage semantic alignment, source weighting, and conflict resolution as first-class design concerns.

The conceptual foundation also includes memory. Agentic systems often maintain short-term working memory, long-term episodic memory, and retrieval indexes that function as externalized knowledge. Generative agents have shown how memory, reflection, and planning can produce coherent long-horizon behavior in simulated environments [11]. In research contexts, memory can help agents track hypotheses, prior searches, unresolved contradictions, and user preferences. However, memory also introduces risks of stale beliefs, privacy leakage, and path dependency. A system that remembers its earlier assumptions may become less willing to revise them in light of new evidence. Consequently, the design of memory is not merely a technical choice but a normative one, shaping what the system treats as relevant, authoritative, and worth retaining.

3. System Architecture and Structural Trade-Offs

The architecture of an agentic retrieval-augmented generation system typically includes a query planner, a retrieval orchestrator, one or more indexes, a fusion and ranking layer, a generation and synthesis component, a verification module, a memory store, and a governance layer. Each component introduces trade-offs. The query planner may improve coverage by decomposing complex questions, but aggressive decomposition can fragment context and produce redundant retrieval. The retrieval orchestrator may increase recall by querying multiple sources, but broader retrieval can introduce noise, latency, and conflicting evidence. The fusion layer may resolve heterogeneity through semantic alignment, but excessive normalization can erase distinctions that matter for interpretation. These tensions cannot be eliminated entirely; they must be managed through explicit system objectives.

Latency and accuracy form a persistent trade-off. Autonomous research agents may perform dozens of retrieval and reasoning steps before producing an answer. This can improve evidence coverage but may be unacceptable in time-sensitive decision environments such as emergency response or clinical triage. Active retrieval strategies attempt to decide when retrieval is necessary rather than retrieving at every step, thereby reducing unnecessary

computation [14]. Self-reflective systems go further by learning when to retrieve, when to critique, and when to generate without external evidence [13]. These approaches illustrate a broader principle: autonomy should be calibrated to the cost of delay and the cost of error. In high-stakes settings, slower and more verifiable deliberation may be preferable to rapid but weakly grounded output.

Source breadth and evidentiary precision also stand in tension. Retrieving from many sources can reduce blind spots, but it can also amplify low-quality or adversarial content. A system that prioritizes diversity may surface minority perspectives, while a system that prioritizes authority may reproduce existing power hierarchies. The challenge is not to choose one objective but to make the weighting explicit and auditable. Survey work on retrieval-augmented generation emphasizes that retrieval quality, ranking, and integration strategies strongly influence downstream factuality and coherence [12]. In agentic settings, these choices become dynamic: the system may need to shift from broad exploratory retrieval to narrow confirmatory retrieval as a decision approaches. Such shifts should be governed by transparent criteria rather than hidden heuristics.

Long-context behavior introduces another architectural concern. Even when relevant evidence is retrieved, language models may attend unevenly across long inputs, sometimes performing better on information at the beginning or end of a context window than in the middle [15]. This phenomenon complicates the assumption that more retrieved text necessarily improves grounding. Agentic systems may need to structure evidence into ranked sections, use intermediate summaries, or employ multiple focused generation passes. Research on when to trust parametric versus non-parametric memory further suggests that models do not always use retrieved evidence reliably, especially when parametric knowledge conflicts with external sources [16]. Architecture must therefore include mechanisms for conflict detection, source comparison, and explicit uncertainty marking, rather than assuming that retrieval automatically yields truth.

4. Autonomy, Evidence, and Decision Making

Evidence-grounded decision making requires more than retrieving relevant documents. It requires assessing the quality, reliability, and applicability of evidence to a specific decision context. Agentic systems can support this process by triangulating sources, identifying contradictions, and generating structured summaries. However, they can also produce fluent syntheses that appear more authoritative than the underlying evidence warrants. Research on model self-knowledge indicates that language models can sometimes recognize what they do not know, but this capacity is imperfect and context-dependent [17], [18]. In autonomous research, calibration is therefore a core safety property. An agent that can say “the evidence is insufficient” or “sources conflict” is more useful than one that always produces a confident answer.

Autonomous research workflows illustrate both the promise and the difficulty. An agentic system might conduct a literature review by searching multiple databases, extracting claims, mapping citation networks, and identifying gaps. It might generate hypotheses, propose experimental designs, and simulate peer review. Yet each of these activities depends on domain conventions that are not fully captured in retrieved text. Scientific validity involves methodological rigor, statistical reasoning, replication, and ethical approval, all of which may be absent from the retrieved passages. Foundation model research has highlighted broad opportunities and risks across domains, including the danger that general-purpose systems may be deployed without sufficient domain-specific validation [19]. For agentic retrieval-

augmented generation, this implies that autonomy should be bounded by domain expertise and institutional review.

Decision making in public policy, medicine, law, and engineering adds further complexity. In these settings, evidence is contested, values are plural, and decisions have distributional consequences. An agentic system may retrieve relevant statutes, clinical trials, or engineering standards, but it cannot resolve normative disagreements by retrieval alone. Ethical and social risk analyses of language models show that harms can arise from representational bias, misinformation, privacy leakage, and unequal access [20]. Agentic systems can exacerbate these harms by scaling synthesis across many decisions. Therefore, evidence-grounded decision making should be understood as a human-machine process in which agents provide structured evidence and reasoning while humans retain responsibility for judgment, especially when rights, safety, or public resources are at stake.

5. Governance, Fairness, and Accountability

Governance of agentic retrieval-augmented generation must address the full lifecycle of data, models, and decisions. At the data layer, retrieval corpora determine what counts as evidence. If those corpora underrepresent certain languages, regions, or marginalized perspectives, the agent’s outputs will inherit those gaps. Critiques of large language models have emphasized that scale does not automatically produce diversity or fairness; large datasets can encode historical inequalities and amplify dominant voices [21]. In agentic retrieval, this problem is compounded because the system may autonomously choose which sources to consult. Without governance, it may repeatedly privilege sources that are easier to retrieve, more machine-readable, or institutionally powerful.

Accountability requires traceability. A decision supported by an agentic system should be reconstructable in terms of queries issued, sources retrieved, passages used, transformations applied, and uncertainties identified. This is not merely a logging problem. It is a question of institutional design. Work on algorithmic ethics has mapped concerns about opacity, bias, due process, and responsibility [23]. Internal auditing frameworks propose end-to-end processes for documenting, testing, and monitoring algorithmic systems [24]. For agentic retrieval-augmented generation, such frameworks must extend to retrieval provenance, source versioning, and the rationale for source selection. Without these controls, it becomes difficult to contest a decision or to assign responsibility when harm occurs.

Fairness also involves access and participation. If agentic research systems are concentrated in well-resourced institutions, they may widen existing epistemic inequalities. Researchers with limited compute, data access, or technical expertise may be unable to audit or challenge the systems that shape funding, publication, or policy. Floridi and Chiriatti argue that powerful language models raise questions about agency, creativity, and responsibility that cannot be reduced to technical performance [22]. In the context of autonomous research, fairness requires attention to who can use the system, whose questions are prioritized, and whose evidence is treated as authoritative. Governance should therefore include participatory mechanisms, not only technical audits.

6. Infrastructure, Deployment, and Sustainability

Agentic retrieval-augmented generation depends on substantial infrastructure. It requires vector indexes, document stores, embedding models, orchestration services, monitoring systems, and secure computation environments. These components must be maintained over time, and their interactions can create hidden technical debt [25]. Technical debt in machine

learning systems arises from entanglements between data dependencies, configuration, model updates, and feedback loops. In agentic retrieval, debt can accumulate when retrieval indexes become stale, when source schemas change, when embedding models are updated without reindexing, or when evaluation pipelines fail to capture newly relevant evidence. Sustainable deployment therefore requires lifecycle management, not just initial model training.

Sustainability also has environmental and economic dimensions. Large-scale retrieval and iterative reasoning consume compute and energy. While retrieval can reduce the need for parameter scaling in some tasks, agentic systems may compensate by performing many retrieval and reasoning steps. The environmental cost must be weighed against the value of improved evidence grounding. In some applications, caching, hierarchical retrieval, and selective autonomy can reduce unnecessary computation. In others, the cost may be justified by safety or accuracy requirements. The key is to make these trade-offs explicit in deployment planning. A system that is environmentally unsustainable or economically inaccessible will not remain trustworthy over the long term.

Deployment in organizations introduces additional concerns. Agentic systems must integrate with existing workflows, data governance policies, and legal requirements. They may need to operate across cloud and edge environments, respect data residency rules, and provide audit trails for regulators. Infrastructure design should therefore include identity management, access control, encryption, and provenance tracking. It should also support human oversight through dashboards, review queues, and escalation paths. The goal is not to remove humans from the loop but to design interfaces and controls that make human judgment more effective where it matters most.

7. Robustness, Evaluation, and Policy Implications

Robustness in agentic retrieval-augmented generation is multidimensional. A robust system maintains groundedness when sources are noisy, contradictory, or adversarial. It maintains calibration when evidence is sparse. It maintains fairness when retrieval corpora are biased. It maintains security when external tools or documents contain prompt injection attempts. Evaluation must therefore go beyond answer accuracy. It should include retrieval precision, source diversity, citation correctness, contradiction detection, uncertainty calibration, latency, cost, and user trust. Studies of long-context use, self-reflection, and active retrieval provide methodological foundations for such evaluation [13], [14], [15]. They also show that more retrieval is not always better; the quality and organization of evidence matter significantly.

Policy implications follow from these evaluation requirements. Regulators may need to require provenance metadata, audit logs, and human oversight for high-stakes applications. Professional bodies may develop standards for evidence grading and disclosure when agentic systems contribute to research or decisions. Organizations may need to define liability when an autonomous agent retrieves outdated information or fails to surface conflicting evidence. Ethical analyses of algorithms emphasize that responsibility cannot be delegated entirely to technical systems [22], [23]. Internal auditing frameworks offer a model for accountability, but they must be adapted to the dynamic and generative nature of agentic retrieval [24].

Cross-domain comparisons reveal different tolerances for error. In exploratory scientific research, false leads may be acceptable if they generate hypotheses and are later corrected. In clinical decision support, false reassurance can cause direct harm. In legal research, missing a precedent can change the outcome of a case. In public policy, biased retrieval can entrench inequality. These differences suggest that agentic retrieval-augmented generation should not

be deployed with a one-size-fits-all autonomy setting. Instead, systems should be designed with domain-specific constraints, escalation rules, and verification thresholds. Autonomy should expand only as evidence, infrastructure, and governance mature.

8. Conclusion

Agentic retrieval-augmented generation represents a significant evolution from static retrieval pipelines toward systems that can plan, search, critique, and synthesize evidence across heterogeneous sources. This evolution offers new possibilities for autonomous research and evidence-grounded decision making, but it also intensifies existing challenges related to grounding, fairness, accountability, and sustainability. The analysis in this paper suggests that progress will depend less on isolated model improvements than on the co-design of architecture, infrastructure, governance, and policy. Retrieval must be treated as an inquiry process, not merely a memory lookup. Autonomy must be bounded by verification, provenance, and human oversight. Infrastructure must be maintained as a long-term socio-technical commitment.

A central conclusion is that evidence-grounded decision making is not achieved by retrieval alone. It requires institutional mechanisms that make evidence traceable, contestable, and accountable. Agentic systems can support this goal by surfacing sources, documenting reasoning, and identifying uncertainty, but they can also obscure it by producing fluent syntheses that mask weak evidence. The design choice is therefore normative as well as technical. Systems should be built to expose their assumptions, respect domain-specific standards, and support human judgment rather than replace it. If developed with these principles, agentic retrieval-augmented generation can contribute to more transparent, rigorous, and inclusive research and decision environments. If developed without them, it risks scaling opacity and error. The path forward requires sustained interdisciplinary work across computer science, information science, law, ethics, and domain practice.

References

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
2. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., ... Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. *Proceedings of the 39th International Conference on Machine Learning*, 162, 2206–2240.
3. Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 874–880.
4. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781.

5. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). REALM: Retrieval-augmented language model pre-training. *Proceedings of the 37th International Conference on Machine Learning*, 119, 3929–3938.
6. Deng, Q., Liu, S., Qiu, H., Long, N., Lin, W., & Chen, S. (2026, April). Unified Semantic Space and Attention-Based Fusion Framework for Multi-Source Heterogeneous Retrieval-Augmented Generation. In *2026 6th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)* (pp. 591-597). IEEE.
7. 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*.
8. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 68539–68551.
9. 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.
10. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), 186345.
11. 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. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22.
12. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
13. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *International Conference on Learning Representations*.
14. Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., & Neubig, G. (2023). Active retrieval augmented generation. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7969–7992.
15. Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157–173.
16. Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2023). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10993–11009.
17. Pan, L., Cremer, M., Heinisch, B., & Beyer, A. (2023). Do large language models know what they don't know? Findings of the Association for Computational Linguistics: *ACL 2023*, 11048–11064.

18. Kadavath, S., Conerly, T., Askill, 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 preprint arXiv:2207.05221.
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 preprint 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). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.
21. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.
22. Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30, 681–694.
23. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1–21.
24. 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. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33–44.
25. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28, 2503–2511.