

Knowledge Graph-Augmented LLM Agents for Explainable Scientific Discovery and Automated Hypothesis Generation

Aleks Banys

Department of Computer Science, George Mason University, Fairfax, VA, USA.

aleksbanys@gmu.edu

Katariina Kaikkonen

School of Computing, Clemson University, Clemson, SC, USA.

katariina.work@clemson.edu

Abstract

Scientific discovery increasingly depends on the ability to integrate vast, heterogeneous knowledge sources with computational agents capable of reasoning, generating hypotheses, and designing validation pathways. Large language models have demonstrated remarkable capacities for synthesis, abstraction, and conversational reasoning, yet their tendency toward hallucination, shallow provenance, and unstable factual grounding limits their direct use in high-stakes scientific workflows. Knowledge graphs offer structured, semantically explicit, and auditable representations of entities, relations, and evidence, making them a natural complement to generative models. This paper examines knowledge graph-augmented large language model agents as a system-level architecture for explainable scientific discovery and automated hypothesis generation. It analyzes the structural trade-offs among retrieval precision, coverage, latency, cost, and interpretability; discusses agent orchestration, provenance tracking, governance, deployment, sustainability, robustness, fairness, and policy implications; and compares applications across biomedicine, chemistry, materials science, and the social sciences. The central argument is that explainable discovery agents should not be treated as monolithic models but as socio-technical infrastructures that combine graph curation, tool use, human oversight, auditability, and reproducible evaluation. Such systems can expand the scale and diversity of hypothesis generation, but only if they are governed by transparent evidence standards, domain-specific validation, and institutional accountability. The paper concludes with forward-looking design principles for building knowledge graph-augmented agents that are scientifically useful, epistemically responsible, and operationally sustainable.

Keywords

knowledge graphs, large language models, scientific discovery, hypothesis generation, explainable artificial intelligence, agentic systems, provenance, research infrastructure, governance.

1. Introduction

The contemporary scientific enterprise faces a paradox of abundance. Digital repositories, instruments, publications, clinical registries, materials databases, and simulation environments produce data at rates that exceed the interpretive capacity of individual research teams. At the same time, many of the most consequential discoveries depend on subtle cross-

domain connections that are difficult to detect through conventional literature review or single-domain analysis. Large language models have emerged as powerful interfaces for navigating this complexity because they can summarize, translate, code, and reason across textual corpora [1], [2]. Their few-shot and chain-of-thought capabilities have made them attractive candidates for assisting scientific ideation, experimental planning, and evidence synthesis [3], [4]. However, their generative fluency can obscure weak evidentiary foundations, and their outputs often lack the provenance required for scientific accountability [6]. This tension motivates the development of knowledge graph-augmented agent architectures that combine the flexibility of language models with the structure, traceability, and domain semantics of curated graphs [7], [8].

Knowledge graphs are not merely databases. They encode entities, relations, ontologies, evidence statements, and sometimes probabilistic or temporal qualifications in forms that support querying, inference, and explanation [8]. They can represent genes, proteins, diseases, compounds, materials, citations, experimental conditions, and causal hypotheses in a shared semantic space. Their construction and maintenance, however, involve difficult problems of entity resolution, schema evolution, incompleteness, bias, and provenance decay [9]. When integrated with large language model agents, knowledge graphs can serve as retrieval substrates, constraint systems, memory structures, and validation references. The resulting systems can potentially generate hypotheses that are both creative and grounded, while also producing explanations that trace claims back to source evidence. Yet this integration is not a simple engineering recipe. It changes the epistemology of discovery, redistributes authority among human researchers, computational agents, and institutional databases, and introduces new forms of infrastructural dependency [5], [20].

This paper develops a system-level analysis of knowledge graph-augmented large language model agents for explainable scientific discovery and automated hypothesis generation. It argues that the relevant unit of analysis is not the language model alone but the entire socio-technical stack: graph curation pipelines, retrieval mechanisms, agent orchestration, tool interfaces, provenance ledgers, human review protocols, deployment environments, and governance arrangements. The paper examines how design choices in this stack affect scientific validity, interpretability, robustness, fairness, sustainability, and policy. It also considers cross-domain differences, because discovery in medicine, chemistry, materials science, and social science imposes distinct evidentiary standards, ethical constraints, and validation cycles [13], [22]. The discussion is organized around architecture, augmentation, agency, explainability, deployment, safety, equity, evaluation, and case-based comparison. The central claim is that explainable discovery agents should be built as auditable infrastructures rather than opaque oracles, with knowledge graphs serving as semantic anchors and human oversight serving as the final epistemic authority.

2. Conceptual Foundations and Reference Architecture

A knowledge graph-augmented language model agent can be understood as a layered system. The lowest layer contains data sources, ontologies, graph stores, vector indexes, and provenance records. Above this lies a retrieval and reasoning layer that maps natural language queries into graph traversals, semantic searches, or hybrid operations. The agent layer includes planning, tool invocation, memory management, critique, and dialogue. The human interface layer provides explanation views, uncertainty indicators, evidence summaries, and intervention controls. Finally, the governance layer defines permissions, audit trails, validation rules, and compliance procedures. This architecture differs from standalone

retrieval-augmented generation because the graph is not only a source of passages but also a constraint space that can reject implausible relations, expose missing evidence, and generate structured explanations [7], [8], [21].

The reference architecture must balance expressiveness against tractability. A richly axiomatized graph can support logical inference and semantic consistency, but it may be expensive to maintain and brittle under domain change. A lightweight property graph can be easier to deploy, but it may lack the formal semantics needed for cross-domain interoperability. Agentic systems add another layer of complexity because they can autonomously decide when to query the graph, when to call external tools, and when to ask for human clarification. Recent work on scaling agentic models for full-scenario applications suggests that such systems are moving toward broader task coverage and more general tool use [5]. However, generality can undermine domain fidelity if the agent is not constrained by curated scientific knowledge and validation protocols. The design challenge is therefore not to maximize autonomy but to calibrate it against the cost of error, the availability of evidence, and the reversibility of decisions.

The architecture also determines what kinds of explanations are possible. Post hoc explanations generated by a language model may be plausible but unfaithful to the actual computation. In contrast, graph-mediated explanations can expose the specific entities, relations, paths, and evidence records that influenced a hypothesis. A graph path might show that a compound is associated with a protein target, that the target is implicated in a disease pathway, and that a clinical trial reported an adverse event under certain conditions. The explanation is not merely a narrative; it is a trace through a structured evidence space. This traceability is essential for scientific review, regulatory scrutiny, and cumulative knowledge production [14], [15]. It also enables failure analysis: when a hypothesis proves wrong, researchers can inspect whether the error arose from missing graph edges, outdated evidence, faulty entity resolution, or an agent’s overgeneralization.

3. Knowledge Graph Augmentation and Retrieval Mechanisms

The first technical challenge is graph construction and curation. Scientific knowledge graphs are assembled from publications, databases, ontologies, experimental records, and expert annotations. Automated extraction can scale coverage, but it introduces noise, duplication, and context loss. Manual curation improves precision but is expensive and difficult to sustain. Knowledge graph refinement methods address entity resolution, relation extraction, link prediction, error detection, and schema alignment [9]. In an agentic discovery system, these refinement processes become part of the operational loop. The agent may propose new relations, identify contradictions, or request expert review. The graph is therefore not a static backdrop but a dynamic memory that evolves with each discovery cycle. This dynamism improves adaptability but creates challenges for versioning, reproducibility, and accountability.

Retrieval mechanisms determine how language models access graph content. Sparse retrieval, dense vector search, graph traversal, and hybrid methods each have trade-offs. Dense retrieval can capture semantic similarity but may miss precise domain terminology. Graph traversal can follow explicit relations but may be limited by missing edges. Hybrid retrieval can combine the strengths of both, but it increases system complexity and latency. Retrieval-augmented generation provides a general framework for grounding language model outputs in external knowledge [7]. When the external knowledge is a graph, retrieval can return subgraphs, paths, or evidence subgraphs rather than isolated passages. This structured context

helps reduce hallucination and supports explanation generation [6]. However, retrieval quality depends on graph completeness, index freshness, and query understanding. An agent that misinterprets a biomedical relation can retrieve a technically correct but scientifically irrelevant subgraph, leading to confident but misleading hypotheses.

Semantic alignment and interoperability are also critical. Scientific domains use overlapping but inconsistent terminologies. A gene may have multiple identifiers, a disease may have evolving diagnostic criteria, and a material may be characterized under different measurement conditions. Ontologies and controlled vocabularies help align these representations, but perfect interoperability is rare. The FAIR principles emphasize findability, accessibility, interoperability, and reusability, providing normative guidance for data stewardship [12]. In knowledge graph-augmented agents, FAIR compliance affects not only data sharing but also the agent's ability to reason across domains. If a graph cannot express uncertainty, temporality, or negative results, the agent may overstate causal relationships. If provenance is incomplete, explanations may point to sources that cannot be verified. Thus, graph augmentation is not only a retrieval problem; it is an epistemic infrastructure problem.

4. Agentic Hypothesis Generation and Experimental Design

Agentic hypothesis generation involves more than prompting a language model to produce candidate statements. It requires problem framing, literature synthesis, gap detection, analogy formation, constraint checking, feasibility assessment, and experimental design. A useful agent architecture may include a planner that decomposes a research question, a retriever that gathers graph evidence, a generator that proposes hypotheses, a critic that tests logical and empirical consistency, and an executor that interfaces with simulation or laboratory tools. Chain-of-thought reasoning can help structure these steps [3], but it must be validated against graph evidence and domain constraints. The agent's memory can store prior hypotheses, rejected paths, experimental outcomes, and expert feedback, allowing the system to improve over time [21]. This cumulative memory is valuable, but it also raises risks of lock-in, bias amplification, and overfitting to past successes.

Automated hypothesis generation has historical precedents in computational scientific discovery. Early systems demonstrated that machines could formulate laws from data, design experiments, and refine theories [10], [11]. Contemporary language model agents extend this tradition by operating over natural language and heterogeneous knowledge sources. In chemistry, autonomous agents have been used to plan and execute experiments, analyze results, and propose next steps [22]. In biomedicine, the complexity of disease mechanisms, patient heterogeneity, and regulatory requirements makes full autonomy inappropriate, but agents can still assist with target identification, drug repurposing, and clinical trial design [13]. The key is to match the level of autonomy to the validation environment. In silico exploration may tolerate more speculative hypotheses, while wet-lab or clinical translation requires stricter evidence thresholds and human approval.

The agentic loop also changes the economics of hypothesis generation. If a system can produce hundreds of candidate hypotheses in a day, the bottleneck shifts from generation to evaluation. This can be beneficial if evaluation is also augmented by tools, simulations, and graph-based reasoning. But it can be harmful if the system floods researchers with low-quality suggestions that consume attention without improving decision-making. Explainability becomes a filtering mechanism: hypotheses accompanied by clear evidence paths, uncertainty estimates, and feasibility rationales are easier to prioritize. Conversely, hypotheses generated without provenance or constraint checking may be scientifically interesting but operationally

useless. A well-designed system should therefore optimize for evaluability, not just novelty. It should generate hypotheses that are specific enough to test, connected enough to existing knowledge to be plausible, and novel enough to justify investigation.

5. Explainability, Provenance, and Epistemic Governance

Explainability in scientific discovery is not merely a user interface concern. It is a condition for trust, reproducibility, and cumulative knowledge. A knowledge graph-augmented agent can provide several forms of explanation: evidence paths that connect a hypothesis to source data, counterfactual explanations that show what evidence would weaken the claim, uncertainty explanations that identify missing or conflicting graph regions, and process explanations that describe how the agent selected tools and queries. These explanations are more defensible than free-form rationales because they are anchored in structured records. However, they can still be incomplete or misleading if the graph itself is biased, outdated, or curated with hidden assumptions. Explainability therefore depends on graph transparency as much as model transparency.

Provenance and auditability are central governance mechanisms. Model cards and datasheets for datasets provide structured documentation of intended use, limitations, and ethical considerations [14], [15]. Internal algorithmic auditing frameworks extend these ideas by defining accountability processes across the lifecycle of an artificial intelligence system [16]. In a discovery agent, provenance should record the graph version, source documents, extraction methods, confidence scores, retrieval queries, agent decisions, human interventions, and validation outcomes. This record supports reproducibility, regulatory review, and dispute resolution. It also enables meta-science: researchers can study which types of graph structures, prompts, or agent strategies lead to successful discoveries. Without such records, the system may produce impressive demonstrations that cannot be independently verified or improved.

Epistemic governance addresses who controls the knowledge base, who validates hypotheses, and who bears responsibility for errors. Human-centered artificial intelligence emphasizes meaningful human control, transparency, and user empowerment [17]. In scientific contexts, this means that researchers should be able to inspect, challenge, and override agent recommendations. It also means that institutions must define review boards, data stewardship policies, and liability arrangements. The governance layer should distinguish between exploratory hypotheses, which may be shared with appropriate caveats, and actionable claims, which require stronger evidence. This distinction is especially important in medicine and public health, where a misleading hypothesis can influence clinical practice or policy [13]. Governance is not a bureaucratic add-on; it is part of the system architecture that determines how knowledge is created, validated, and used.

6. Deployment, Infrastructure, and Sustainability

Deploying knowledge graph-augmented agents requires significant computational and organizational infrastructure. Graph stores, vector databases, model serving clusters, workflow orchestrators, provenance ledgers, and monitoring systems must be integrated into a reliable pipeline. Large language models may be hosted locally, in the cloud, or through hybrid arrangements, each with different implications for cost, latency, privacy, and sovereignty. Scientific data often carries legal and ethical restrictions, so deployment must support access control, encryption, and audit logging. Federated or decentralized architectures can allow institutions to share knowledge without moving sensitive data, but they complicate

graph alignment and query planning. The infrastructure must also support versioning, because a hypothesis generated with one graph snapshot may not be reproducible with another.

Sustainability is both environmental and institutional. Training and serving large models consume energy and water, and repeated graph queries can add substantial overhead [20]. A system that generates thousands of hypotheses per day may be scientifically exciting but economically and environmentally costly. Designers should consider caching, model distillation, selective retrieval, and tiered autonomy. Not every hypothesis requires the largest model or the full graph. A lightweight classifier may filter candidates before expensive reasoning is invoked. Institutional sustainability is equally important. Knowledge graphs require ongoing curation, ontology maintenance, and expert feedback. If funding ends, the graph may decay, making the agent’s explanations unreliable. Sustainable systems therefore need governance models that support long-term stewardship, not just initial deployment.

Robustness and safety are closely tied to infrastructure. AI safety research has identified problems such as reward hacking, distributional shift, and unintended side effects [18]. In scientific agents, these risks manifest as overoptimization for novelty, exploitation of graph errors, and premature convergence on popular hypotheses. Deep learning systems can also fail in ways that are difficult to anticipate, especially when they are applied outside their training distribution [19]. A robust discovery agent should include adversarial testing, uncertainty quantification, contradiction detection, and human escalation pathways. It should be able to say that the evidence is insufficient rather than generate a plausible but unsupported claim. Robustness is not only about preventing catastrophic failures; it is about maintaining scientific integrity under uncertainty.

7. Fairness, Equity, and Policy Implications

Knowledge graph-augmented agents can amplify existing inequities in science. Research funding, publication, and data collection are unevenly distributed across regions, languages, diseases, and populations. A graph built primarily from English-language literature may underrepresent research from other linguistic communities. A biomedical graph may overrepresent well-studied genes or diseases while neglecting rare conditions. An agent trained on such a graph may generate hypotheses that reinforce these gaps. Fairness therefore requires attention to representation, annotation, and evaluation. Datasheets and model cards can document known gaps [14], [15], but documentation alone does not correct structural bias. Institutions need policies that incentivize inclusive data collection, multilingual graph construction, and community participation.

Policy implications extend to intellectual property, open science, dual use, and accountability. Automated hypothesis generation may produce outputs that are difficult to attribute to a single human inventor, raising questions about patentability and credit. Open science policies may require sharing graph schemas, model configurations, and evaluation protocols, but commercial interests may resist full transparency. Dual-use concerns arise when discovery agents are applied to dangerous pathogens, chemical weapons, or surveillance technologies. Governance frameworks must therefore include risk assessment, export controls, and ethical review. The FAIR principles provide a foundation for data sharing [12], but they must be complemented by rules for sensitive data, indigenous knowledge, and community consent. Policy should also address the concentration of computational power, because only well-resourced institutions can deploy and maintain advanced discovery agents. Without deliberate intervention, these systems may widen existing inequalities in scientific capacity.

8. Evaluation and Validation Frameworks

Evaluating discovery agents is inherently difficult because scientific success is delayed, contested, and context-dependent. Standard machine learning metrics such as accuracy or perplexity are insufficient. Researchers need metrics for novelty, plausibility, testability, reproducibility, explanatory quality, and downstream impact. A hypothesis may be novel but false, plausible but trivial, or true but already known. Evaluation should therefore combine automated checks, expert review, and prospective validation. Knowledge graphs can support some automated checks by identifying whether a proposed relation already exists, whether it contradicts known constraints, or whether it connects distant literatures. However, graph-based novelty is not the same as scientific significance. Expert judgment remains essential.

Validation frameworks should also assess the agent’s process, not only its outputs. Can reviewers trace the evidence? Are uncertainty estimates calibrated? Does the system acknowledge missing data? Are human interventions recorded and respected? Model cards, datasheets, and audit frameworks provide templates for this kind of documentation [14], [15], [16]. Benchmarks should include adversarial cases, out-of-distribution domains, and low-resource settings. They should also test for robustness to graph poisoning, prompt injection, and contradictory evidence. In chemistry and materials science, automated laboratories can provide faster validation loops [22]. In medicine, validation may require preclinical studies, clinical trials, and regulatory review [13]. The evaluation architecture must therefore be domain-specific while retaining shared standards for provenance and reproducibility.

9. Case Illustrations and Cross-Domain Comparisons

In biomedicine, knowledge graph-augmented agents can integrate genomic, proteomic, clinical, and literature data to propose drug repurposing hypotheses or identify biomarkers. The graph can encode pathways, phenotypes, adverse events, and trial outcomes. The agent can retrieve evidence paths and generate explanations for why a compound might affect a disease. However, biomedical validation is slow, expensive, and ethically constrained. A false positive can waste resources, while a false negative can delay treatment. Explainability and human oversight are therefore non-negotiable. The system should be designed to support expert reasoning rather than replace it [13].

In chemistry and materials science, the loop between hypothesis generation and experimentation can be tighter. Autonomous platforms can synthesize compounds, run assays, and feed results back into the graph [22]. The agent can propose reaction conditions, predict material properties, and refine hypotheses based on experimental outcomes. This domain benefits from structured representations of molecules, reactions, and crystal structures. However, safety risks are significant, especially when automated systems can access hazardous materials or execute unsupervised experiments. Robust governance, physical safeguards, and human approval are essential. The knowledge graph must also represent negative results, because failed experiments are often as informative as successful ones.

In the social sciences, discovery agents face different challenges. Data may be qualitative, contested, and context-dependent. Knowledge graphs can represent concepts, theories, methods, and findings, but causal inference is difficult and ethical concerns are pervasive. An agent might synthesize literature or generate hypotheses about social dynamics, but it must avoid reifying biased categories or ignoring power relations. Cross-domain comparison shows that there is no universal autonomy setting. The appropriate level of agent independence depends on the cost of error, the speed of validation, the availability of structured data, and

the social consequences of failure. A mature research infrastructure would allow different domains to configure the same core architecture with different governance and validation parameters.

10. Conclusion

Knowledge graph-augmented large language model agents offer a promising pathway toward explainable scientific discovery and automated hypothesis generation. By combining the generative flexibility of language models with the semantic structure and provenance of knowledge graphs, these systems can generate hypotheses that are grounded in evidence, traceable through structured paths, and open to human scrutiny. However, the promise of these systems depends on architecture, governance, and infrastructure choices that are often treated as secondary. Retrieval precision, graph curation, agent autonomy, provenance, deployment cost, sustainability, robustness, fairness, and policy all shape whether an agent produces useful science or merely plausible text. The required reference on scaling agentic models for full-scenario applications [5] illustrates the broader trend toward general-purpose agents, but scientific discovery demands domain-specific constraints and accountability mechanisms that general benchmarks may not capture.

The paper has argued that knowledge graph-augmented discovery agents should be understood as socio-technical infrastructures. They require continuous graph refinement, transparent provenance, auditable agent behavior, human-centered oversight, and domain-specific validation. They must be evaluated not only by novelty or accuracy but by reproducibility, explanatory quality, and downstream scientific value. They must be governed with attention to equity, dual use, intellectual property, and the concentration of computational power. They must be deployed with sustainable funding and maintenance models, because a decaying knowledge graph can silently erode the reliability of explanations. If these conditions are met, the systems can expand the scale and diversity of scientific hypothesis generation while preserving the epistemic standards that make science cumulative and trustworthy. If they are ignored, the result may be an acceleration of confident but unverifiable claims. The path forward is therefore not to maximize autonomy but to design infrastructures that make discovery more explainable, more inclusive, and more accountable.

References

1. 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* (Vol. 30). Curran Associates.
2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877-1901).
3. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 24824-24837).
4. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.
5. Jiang, Hao, et al. "AgentOmnia: Scaling Agentic Models for Full-Scenario Applications." *arXiv preprint arXiv:2607.23124* (2026).

6. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1-38.
7. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459-9474).
8. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., ... & Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys*, 54(4), 1-37.
9. Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8(3), 489-508.
10. King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., ... & Whelan, K. E. (2009). The automation of science. *Science*, 324(5923), 85-89.
11. Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science*, 324(5923), 81-85.
12. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018.
13. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.
14. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220-229). ACM.
15. 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.
16. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 33-44). ACM.
17. Shneiderman, B. (2020). *Human-centered AI*. Oxford University Press.
18. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
19. Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.
20. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
21. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., ... & Wen, J. R. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), 186345.
22. Boiko, D. A., MacKnight, R., Kline, B., & Gomes, G. (2023). Autonomous chemical research with large language models. *Nature*, 624(7992), 570-578.