

Knowledge Graph and Large Language Model Fusion for Personalized Biomedical Literature Discovery

Ole Sanders

Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.

olesanders893@unr.edu

Brent Marshall

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

hellogrant@clemson.edu

Sawyer Barnett

Department of Computer Science, Colorado State University, Fort Collins, CO, USA.

sawyerbarnett25@colostate.edu

Abstract

The biomedical literature is growing at a rate that makes manual curation and comprehensive literature review increasingly difficult. Personalized discovery systems must reconcile the breadth of scientific evidence with the specific informational needs of researchers, clinicians, and policy analysts. This paper presents a system-level examination of knowledge graph and large language model fusion for personalized biomedical literature discovery. It argues that neither symbolic knowledge organization nor parametric language generation alone is sufficient; their integration must be designed around structured biomedical relations, retrieval-augmented inference, user memory, and accountable deployment practices. The discussion begins with architectural foundations, considering graph encoding, language model query interpretation, and evidence grounding. It then addresses personalization mechanisms, including explicit preference models and implicit behavioral memory, and illustrates their use in discovery workflows for clinical and translational contexts. The paper further analyzes governance, safety, fairness, and sustainability implications, including hallucination control, privacy, reproducibility, and environmental cost. A system-level evaluation perspective is developed, emphasizing expert-in-the-loop assessment, continuous monitoring, and the separation of retrieval relevance from evidentiary authority. The conclusion identifies future directions for adaptive knowledge infrastructure that maintains scientific rigor while improving the efficiency of literature navigation. Throughout, the emphasis is on structural trade-offs and the institutional conditions required for trustworthy deployment in real biomedical settings.

Keywords

biomedical literature discovery, knowledge graphs, large language models, retrieval-augmented generation, personalization, governance, scientific information systems.

1. Introduction

The growth of biomedical science increasingly manifests as an acceleration in the number of published articles, clinical guidelines, preprints, and structured database records. Bibliometric analyses have documented a long-term expansion in scientific publication volume, with biomedical fields contributing substantially to this trend [1]. The result is a discovery environment in which even expert readers cannot maintain comprehensive awareness of relevant findings. The problem is not merely the volume of documents but the fragmentation of evidence across heterogeneous sources, the use of inconsistent terminology, and the rapid evolution of disease models, drug mechanisms, epidemiological data, and clinical best practices. Personalized literature discovery must therefore move beyond generic keyword search toward systems that understand both the domain and the individual user.

Large language models have attracted considerable attention for biomedical question answering and clinical knowledge tasks because they can encode fluent conceptual associations and support conversational interactions [2]. These models are built on transformer architectures that allow attention-based representation learning across long sequences, enabling them to capture complex linguistic patterns and generate context-sensitive responses [3]. However, their parametric memory is inherently limited by training data cutoffs, corpus biases, and a tendency to produce plausible but ungrounded statements. Biomedical applications intensify these risks because incorrect assertions can influence clinical reasoning or research priorities. Consequently, large language models require external knowledge structures that can anchor their outputs in curated biomedical entities and relationships.

Knowledge graphs offer a complementary representational framework by encoding entities such as genes, diseases, drugs, pathways, and experimental methods as nodes, with typed relations connecting them [4]. A knowledge graph can represent a biomedical claim as a structured path rather than as an unstructured sentence, which enables precise retrieval, provenance tracking, and logical constraints. When knowledge graphs are combined with language models, the language model can interpret fuzzy user intent while the graph constrains the answer space and supports source attribution. Retrieval-augmented generation has emerged as a practical pattern for this combination, in which retrieved documents or graph subgraphs are injected into the language model context to improve factual grounding [5]. Domain-adapted biomedical encoders such as BioBERT have further shown that pretraining on scientific text improves entity recognition and relation extraction, which are essential for constructing and refreshing biomedical knowledge graphs [6]. Parallel developments in search systems have identified persistent user memory as a mechanism for improving query matching and intent modeling over time, suggesting that personalized literature discovery should maintain dynamic models of user interests rather than treating each query as an isolated event [7].

The remainder of this paper examines fusion architectures, personalization mechanisms, workflow design, governance, deployment, and evaluation. The central argument is that robust biomedical literature discovery depends on the careful integration of symbolic structure, statistical language understanding, and institutional safeguards. This requires not only technical innovation but also attention to data provenance, model transparency, operational sustainability, and the epistemic norms of biomedical research.

2. Related Work and Systems Context

Biomedical literature discovery has been shaped by several distinct research traditions. Early information retrieval systems relied on controlled vocabularies such as Medical Subject

Headings and on term-weighting functions that estimate relevance from lexical overlap. These systems are interpretable but often fail to capture the conceptual variability of biomedical language. Subsequent vector-based retrieval and neural ranking models improved semantic matching, yet they introduced new challenges related to opacity and domain generalization. Domain-specific language model pretraining has become a central technique for adapting general-purpose language models to biomedical corpora, producing models that better recognize specialized entities and synonyms [8]. Such models support query understanding, entity normalization, and evidence extraction, but they do not by themselves provide a curated relational substrate.

Knowledge graphs have been developed extensively for biomedical knowledge integration, linking genes, proteins, diseases, drugs, and phenotypic traits, with initiatives such as systematic integrative networks enabling drug repurposing analyses [9]. Those integrative efforts reveal a fundamental tension between the desire for broad coverage and the need for editorial control. Automatically extracted relations can rapidly expand a graph but introduce false positives, whereas manually curated relations are more reliable but expensive to maintain. Biomedical retrieval evaluation has similarly evolved. Campaigns such as the TREC clinical trials track have demonstrated that tasks like patient-to-trial matching require hybrid strategies that combine structured eligibility criteria with potentially relevant free-text fields [10]. The evaluation evidence suggests that even strong semantic encoders perform inconsistently when query intent is underspecified or when the target documents are long, heterogeneous, and deeply technical. Personalization has been studied in general web search for decades, with early systems analyzing past interactions to re-rank results by inferred topical interests [11]. The central insight is that a user's history provides a prior over future information needs, reducing ambiguity in short queries. In biomedical settings, the same principle applies but must account for the high cost of false negatives and the frequent need to discover adjacent topics beyond a narrowly defined profile. Recent work on language model-based user memory extends this idea to interactive systems, where the model maintains a latent summary of prior sessions and uses it to refine query matching and intent interpretation [7]. In biomedicine, such memory could encode a researcher's preferred organisms, disease areas, evidence types, or methodological constraints. The challenge is that biomedical relevance is often domain-grounded and cannot be inferred from click behavior alone.

3. Architectural Foundations of Knowledge Graph and Language Model Fusion

At the architectural level, a fusion system typically includes three interacting layers: a representation layer that encodes queries and documents, a retrieval layer that selects candidate evidence, and a generation or ranking layer that synthesizes a response. Transformer-based bidirectional encoders such as BERT have become standard building blocks because they capture contextual meaning through attention over entire sequences rather than relying on local word order [12]. In biomedical literature, these encoders can map ambiguous phrases like diabetes complications or inflammatory response into dense vectors that align with semantically related passages even when lexical overlap is low. Knowledge graph involvement can occur at multiple points. One strategy is to embed graph entities and relations into vector space so that language representations are informed by known biomedical relationships. Unified models such as KEPLER integrate text and knowledge graph objectives during pretraining, thereby encouraging the representation of a disease entity to reflect both its textual contexts and its relations to symptoms, drugs, and genes [13]. More advanced language-knowledge graph pretraining systems jointly optimize masked language

modeling and link prediction, allowing relational constraints to shape semantic encoding while preserving linguistic flexibility [14]. These joint representations support tasks such as entity disambiguation and relation-aware retrieval.

A second strategy relies on retrieval-augmented generation, in which the language model does not permanently store all factual knowledge but instead receives retrieved graph subgraphs and document passages at inference time [5]. This separation reduces the burden on parametric memory and permits evidence provenance to be attached to generated statements. For personalized biomedical discovery, the retrieval layer can be conditioned on user profiles, selecting graph neighborhoods that reflect individual research interests. However, retrieval-augmented architectures introduce latency and increase prompt length, which must be balanced against the benefit of grounding. The composition of the context window therefore becomes a system design problem involving trade-offs among recall, precision, computational cost, and the model's ability to distinguish relevant from distracting material. The reliability of such systems is shaped by the fact that large language models can produce fluent biomedical text without an underlying evidentiary commitment [15]. This characteristic makes it necessary to design architectural boundaries that prevent the model from overriding retrieved evidence. A common approach is to treat the language model as an interface rather than as the source of biomedical truth. The graph retrieval subsystem returns typed paths with provenance, and the language model may translate those paths into natural language but should not invent relations or disregard contradictions. Such boundaries are not simply technical; they reflect an epistemic stance about what counts as evidence in biomedical inquiry.

4. Personalization, User Memory, and Intent Modeling

Personalization in biomedical literature discovery requires more than a static interest profile. Researchers often move among topics, shift between method-focused and disease-focused questions, and maintain long-term projects alongside exploratory side interests. A personalized system must model both persistent preferences and session-specific intent. Early personalized search research established that historical interaction data can be used to infer topical preferences and improve ranking without requiring users to articulate every constraint [11]. Recent work has extended this idea through language model-based user memory systems that capture query patterns and intent shifts across sessions, enabling more precise matching between current information needs and available literature [7]. In biomedicine, such memory could represent a researcher's typical evidence hierarchy, preferred species or model systems, tolerance for preprint material, and interest in observational versus interventional study designs.

User memory can be implemented as explicit preference schemas, implicit behavioral summaries, or a combination of both. Explicit preference settings give users control and make the system's assumptions inspectable, but they impose interaction costs and may become stale. Implicit methods infer preferences from reading time, downloads, citation behavior, and query reformulation, yet they risk producing profiles that overfit recent activity or reinforce unproductive biases. The technical challenge is to align the memory representation with graph query formulation. For example, a user who repeatedly opens papers about kinase inhibitors and renal cell carcinoma should cause the system to expand graph paths around those entities, but not so aggressively that it excludes evidence from other disease contexts where similar molecular mechanisms appear. Privacy is a central constraint in this layer. User memory necessarily stores traces of individual behavior, which could reveal unpublished research

directions, clinical interests, or institutional priorities [16]. Biomedical institutions often operate under strict data protection regimes, and personalization systems must support forgetting, disclosure controls, and user-visible memory management. Decentralized or local memory storage may reduce the concentration of sensitive behavioral data, but it complicates model improvement and cross-user learning. System designers must therefore choose between centralized adaptive models and privacy-preserving alternatives such as federated user modeling or locally maintained profiles. These choices are not purely operational because they affect the system's ability to provide recommendations across user groups and to detect data drift.

A further concern is the potential for personalization to narrow discovery. Although precision is valuable, biomedical innovation often arises from noticing connections outside an established specialization. Purely similarity-based recommendation may suppress serendipitous discoveries. Therefore memory systems should include exploration mechanisms that intentionally introduce controlled diversity, graph-based analogical paths, and counterfactual associations. Balancing exploitation of known interests against exploration of adjacent knowledge is an architectural and governance question, not merely a parameter tuning problem.

5. Biomedical Discovery Workflows and Case Illustrations

Personalized fusion systems can support several distinct biomedical workflows. In literature surveillance, a researcher may wish to receive a weekly digest of new findings relevant to a long-running project. The system must cluster newly ingested abstracts, map them to graph entities, and determine whether they fall within a personalized interest envelope without generating excessive noise. In this workflow, precision and timeliness are paramount, and the language model may serve as a summarizer that highlights how a new finding connects to the user's saved graph subnetwork. In clinical evidence synthesis, an analyst may need to compile all trial results, systematic reviews, and guidelines relevant to a comparative effectiveness question. The fusion system must distinguish between study designs, outcome measures, and confidence levels. A knowledge graph that models interventions, comparators, outcomes, and populations can support structured filtering. The language model can translate natural language questions into graph patterns and can synthesize retrieved evidence, but it must preserve the distinction between evidence summaries and recommendations. LitCovid and similar curated literature resources illustrate how rapid indexing and structured metadata can support specialized review when combined with query understanding systems [19].

Another case involves drug repurposing exploration, where a researcher looks for indirect paths between an approved drug and a disease mechanism. Systematic integrative networks have demonstrated that graph paths connecting drug targets, disease-associated genes, and intermediate phenotypes can generate repurposing hypotheses [9]. A fusion system can use a language model to interpret a loosely formulated research question, retrieve ranked graph paths, and explain the biological plausibility of each path in readable form. The risk is that the language model may overstate the causal strength of a path that merely reflects data curation artifacts. Therefore the graph layer must track edge sources, evidence codes, and confidence levels, and the language interface must be constrained not to elide those qualifications. For translational researchers, the system can support biomarker discovery by linking literature-derived findings to gene expression datasets and pathway annotations. This requires integrating literature graphs with experimental data graphs, a task complicated by identifier mismatch, batch effects, and dynamic curation. The fusion architecture should treat

experimental data and literature claims as separate evidentiary layers with different reliability models. A system-level perspective recognizes that the value of personalization emerges not from any single retrieval result but from repeated interaction across a project lifecycle, during which the system learns which graph edges and evidence types the user trusts.

6. Governance, Fairness, and Epistemic Integrity

Biomedical literature discovery systems occupy a sensitive position in the production of scientific knowledge. Their outputs may influence hypothesis generation, grant proposal priorities, clinical decisions, and policy briefs. Governance therefore begins with transparency about model limitations and the provenance of retrieved claims. Model cards and system documentation practices offer a mechanism for communicating intended use, training data, and evaluation results [17]. A fusion system should document not only the language model but also the knowledge graph sources, extraction pipelines, update frequency, and personalization logic. Without such documentation, users cannot appropriately interpret a summary that blends evidence from different confidence levels.

Fairness concerns arise at multiple levels. The underlying biomedical literature is itself biased by publication practices, funding priorities, and historical underrepresentation of certain populations. A knowledge graph built from this literature may propagate those biases into entity coverage and relation density. A language model pretrained on broad text may further amplify stereotypes or produce default assumptions that are inappropriate for individual patients or communities. Personalization can worsen or mitigate these effects depending on whether the user model is transparent and whether the system exposes diversity across disciplines, geographies, and study populations. Governance mechanisms should include bias audits of graph coverage, evaluation across underrepresented disease areas, and interventions to surface evidence from neglected subfields. Epistemic integrity requires that the system distinguish between citing a paper, summarizing a claim, and endorsing that claim. Automated models can mistake citation frequency for scientific consensus or confuse correlation with causation. A fusion system should preserve metadata indicating study type, sample size, journal, retraction status, and replication findings. When generating a synthesis, the model must avoid flattening contested evidence into a single confident narrative. Human experts must retain the ability to inspect the retrieved evidence, expand the graph context, and challenge the system's selection. The goal is not to automate biomedical judgment but to support accountable human judgment with better organized evidence.

Institutional policy implications follow from these concerns. Universities, publishers, and public health agencies may need shared standards for provenance logging, retraction propagation, and user data protection. A system that aggregates literature and personal behavior creates a data asset that could be repurposed for surveillance or commercial profiling. Governance should specify data minimization, auditability, and constraints on secondary use. These requirements are not external constraints on an otherwise technical design; they shape the architecture itself by determining where computation occurs, what data may be stored, and how system behavior is explained.

7. Deployment, Sustainability, and Infrastructure

Deploying a biomedically grounded fusion system in production involves substantial infrastructure decisions. Large language models are computationally intensive, and their environmental and financial costs have been documented in the natural language processing literature [18]. Biomedical deployments may require frequent updates to the literature index,

graph refresh cycles, and retraining of user memory models. A sustainable system design will therefore consider model distillation, caching of frequent query patterns, and selective use of large models only when retrieval evidence is ambiguous or when synthesis demands high fluency. Latency and availability requirements differ across use cases. A clinician seeking evidence at the point of care needs rapid, highly filtered responses; a systematic reviewer may tolerate longer inference if the result includes more comprehensive coverage. These requirements affect whether the language model runs locally or via a remote service, how much context can be retrieved, and whether graph traversal is executed synchronously. Federated deployment options, in which user memory stays within institutional boundaries, can reduce privacy risk but complicate global model updates and performance monitoring.

Sustainability also applies to the knowledge graph itself. Biomedical ontologies evolve as new genes are characterized, drugs are renamed, and disease subtypes are refined. Graph maintenance is a socio-technical process requiring expert curation, automated relation extraction, and periodic manual review. Automation can accelerate coverage but may introduce errors that compound over time. The system should therefore log extraction decisions, preserve versioned graph snapshots, and support rollback when a source is corrected. The intellectual labor of curators must be recognized as part of the infrastructure rather than treated as a one-time annotation cost. Interoperability with existing biomedical platforms depends on adherence to open identifiers and data standards. Entity resolution across PubMed, clinical trial registries, electronic health records, and pathway databases remains a persistent challenge. Fusion systems should adopt community identifiers where possible and maintain mapping tables where standards diverge. This is a structural trade-off because strict identifier discipline reduces noise but may exclude novel preprint content or emerging terminology. A well-governed deployment treats mapping as a first-class component, with regular updates and quality metrics.

8. Evaluation and Continuous Monitoring

Evaluation of personalized literature discovery cannot be reduced to a single ranking metric. Offline measures such as recall and precision on benchmark collections provide useful signals but may not capture the quality of a synthesized answer, the user's trust in the system, or the long-term effect on research directions. Biomedical retrieval campaigns have shown that carefully constructed test collections enable reproducible comparisons, yet their relevance judgments often reflect only a narrow slice of user intent [10]. A mature evaluation program therefore combines offline benchmarks, expert panel reviews, and longitudinal user studies. Human-AI interaction research has long emphasized that interface design and system behavior should be evaluated together, not as separate components [21]. In a biomedical discovery context, an ideal interaction might allow the user to inspect graph paths, expand evidence nodes, and correct incorrect entity links. The system should learn from these corrections without overgeneralizing from a single event. Evaluation should track whether users can detect when the language model has omitted an important trial or overstated a mechanistic claim. This shifts the focus from raw retrieval effectiveness to the system's ability to support expert oversight.

Continuous monitoring is necessary because biomedical knowledge and user needs drift over time. A model that performs well at initial deployment may degrade as terminology changes or as the literature becomes dominated by a new topic. Monitoring indicators include entity coverage, retrieval latency, rate of retracted papers appearing in recommendations, and the frequency with which users override system suggestions. Monitoring should also detect

feedback loops in which personalization narrows retrieval so much that the user never encounters contrary evidence. Evaluation and governance are thus intertwined, since the decision to prioritize precision over coverage is a value-laden system choice. Finally, evaluating fairness requires stratified analysis across disease areas, languages, and user populations. A biomedical discovery system that works well for well-resourced cancer genomics may fail for neglected tropical diseases with sparse graph coverage and limited scholarly text. Foundation models for generalist medical tasks have shown promise across diverse domains, but their performance varies and must be assessed in context [20]. Longitudinal evaluation should include not only aggregate accuracy but also the distribution of failures, ensuring that underrepresented areas do not fall further behind.

9. Conclusion

Personalized biomedical literature discovery sits at the intersection of knowledge organization, generative language understanding, and institutional data governance. This paper has argued that fusing knowledge graphs with large language models can improve grounding and interpretability, but only if the architecture is designed to maintain clear evidentiary boundaries, support user memory with privacy safeguards, and treat evaluation as an ongoing epistemic practice. The relational structure of a knowledge graph provides a substrate for accountable retrieval, while the fluent language model offers a flexible interface to evidence. Neither component can guarantee reliability on its own.

Several structural trade-offs are likely to shape future systems. The first concerns the balance between parametric knowledge and external retrieval. Larger models may reduce the frequency of retrieval errors but increase latency, cost, and the risk of generating confident statements that disregard retrieved evidence. The second trade-off involves user memory and privacy. More detailed user profiles can improve personalization but create sensitive behavioral records that must be protected, audited, and deletable. The third trade-off concerns the granularity of knowledge graph representation. Fine-grained relations support precise querying but demand more expensive curation, whereas coarse relations scale more easily but provide weaker constraints for evidence synthesis. A fourth trade-off arises between exploration and exploitation. Systems that strictly optimize for past interests may suppress cross-domain discovery, while systems that overemphasize novelty may bury users in irrelevant material. Sustainable deployments will likely require explicit controls over these balances and transparent feedback loops.

Future research should develop hybrid architectures in which graph-based reasoners handle relational constraints, vector retrieval handles surface similarity, and language models manage interpretive flexibility under constrained generation policies. Biomedical institutions should also invest in shared curation pipelines that produce versioned graphs with provenance tied to source documents. The maintenance of such infrastructure is not a peripheral technical task but a core condition for trustworthy discovery. As foundation models become more capable, there will be pressure to replace structured knowledge graphs with parametric memorization. This would be a mistake in biomedical contexts. Scientific literature is not a static training corpus but a living record of contested claims, retractions, and evolving evidence. A graph that persists independently of any language model can represent those changes more faithfully than a set of static weights. The most promising path forward is therefore a governance-aware fusion in which symbolic structure, statistical language understanding, and user memory coexist under accountable institutional oversight.

References

1. Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11), 2215–2222. <https://doi.org/10.1002/asi.23329>
2. Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review. *Healthcare*, 11(6), 887. <https://doi.org/10.3390/healthcare11060887>
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
4. Nicholson, D. N., & Greene, C. S. (2020). Constructing knowledge graphs and their biomedical applications. *Computational and Structural Biotechnology Journal*, 18, 1414–1428. <https://doi.org/10.1016/j.csbj.2020.05.017>
5. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
6. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240. <https://doi.org/10.1093/bioinformatics/btz682>
7. Yu, X. (2026). Enhancing Search Efficiency through LLM-Based User Memory Systems for Query Matching and Intent Modeling.
8. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1), Article 2. <https://doi.org/10.1145/3458754>
9. Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Green, A., Khankhanian, P., Baranzini, S. E., & Greene, C. S. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6, e26726. <https://doi.org/10.7554/eLife.26726>
10. Roberts, K., Demner-Fushman, D., Voorhees, E. M., & Hersh, W. R. (2017). Overview of the TREC 2017 clinical trials track. In *Proceedings of the Twenty-Sixth Text REtrieval Conference*. National Institute of Standards and Technology.
11. Speretta, M., & Gauch, S. (2005). Personalized search based on user search histories. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 622–628). IEEE. <https://doi.org/10.1109/WI.2005.114>
12. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics.
13. Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., & Tang, J. (2021). KEPLER: A unified model for knowledge embedding and pre-trained language representation.

Transactions of the Association for Computational Linguistics, 9, 176–194.
https://doi.org/10.1162/tacl_a_00360

14. Yasunaga, M., Bosselut, A., Ren, H., Zhang, X., Manning, C. D., Liang, P., & Leskovec, J. (2022). Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems*, 35, 37309–37323.
15. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., 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>
16. Kobsa, A. (2007). Privacy-enhanced personalization. *Communications of the ACM*, 50(8), 24–33. <https://doi.org/10.1145/1278201.1278202>
17. 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). ACM. <https://doi.org/10.1145/3287560.3287596>
18. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645–3650). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1355>
19. Chen, Q., Allot, A., & Lu, Z. (2021). LitCovid: An open database of COVID-19 literature. *Nucleic Acids Research*, 49(D1), D1534–D1540. <https://doi.org/10.1093/nar/gkaa952>
20. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956), 259–265. <https://doi.org/10.1038/s41586-023-05881-4>
21. Amershi, S., Weld, D., Vorvoreanu, M., Fournery, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). ACM. <https://doi.org/10.1145/3290605.3300233>