Trustworthy Retrieval-Augmented Generation for the Digital Preservation and Transparent Presentation of Biomedical Scientific Heritage

Harshetha Murthy Keshav Murthy[1], Alexander I. Iliev[1, 2][0000-0002-4220-3637]

[1] SRH Hochschule Berlin, 221 Sonnenallee, 12059 Berlin, Germany
[2] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, 8 Acad. Georgi Bonchev Street, 1113 Sofia, Bulgaria
harshethabm@gmail.com, ailiev@berkeley.edu

Abstract. Biomedical scientific heritage faces a dual challenge of exponential growth and inaccessibility. This paper presents a Trustworthy Retrieval Augmented Generation (RAG) framework designed for the digital preservation and transparent presentation of biomedical knowledge integrating ontology aware retrieval (MeSH/UMLS), hybrid BM25 + dense search via Reciprocal Rank Fusion, sentence level attribution, and a novel composite biomedical Trust Index (BTI) quantifying factual accuracy, attribution fidelity and adversarial robustness on the MedQuAD corpus, the framework achieves 89.1% top-5 recall, 85.6% attribution accuracy, F1 score of 75.2 and BTI of 0.84 on clean data.

Keywords: Retrieval-Augmented Generation, Biomedical NLP, Trustworthy AI, Digital Preservation, Biomedical Trust Index.

1 Introduction

The digital preservation of biomedical scientific heritage creates a fundamental paradox: the volume that makes it valuable also makes it inaccessible. PubMed alone indexes about 36 million citations, increasing by over one million annually. Making conserved knowledge reliably queryable, factually verifiable, and clearly visible to researchers, physicians, and archivists is the new challenge, not archiving.

Large language models (LLMs) provide transformational synthesis skills, but hallucinate at unacceptable rates in archival or clinical settings, producing plausible but factually unsubstantiated assertions (Guan et al., 2024). Retrieval-Augmented Generation (RAG) addresses this by rooting LLM results on retrieved source materials (Lewis et al., 2020). Domain-specific models, such as BioBERT (Lee et al., 2020), showed that in-domain pre-training significantly enhances factual precision, serving as the foundation for modern RAG systems.

Existing biomedical RAG systems have three outstanding weaknesses: (i) sparse retrievers miss semantically similar concepts due to vocabulary mismatch; (ii) generated claims are untraceable to source passages; and (iii) there is no uniform metric to capture

Digital Presentation and Preservation of Cultural and Scientific Heritage. Volume 16, pp. 377-386, 2026.
DOI: 10.55630/dipp.2026.16.31, ISSN 1314-4006, eISSN 2535-0366.

output trustworthiness holistically. Zhou et al. (2024) show that trustworthiness in RAG is still an open, fragmented challenge. This study covers five research problems, including sentence-level traceability in biomedical RAG, ontology-aware retrieval benefits, adversarial defense efficacy, trust index validity, and traceability-performance trade-offs.

2 Related Work and Theoretical Background

2.1 Evolution of Biomedical NLP and RAG

Biomedical NLP evolved from rule-based systems to domain-specific transformers (BioBERT, SciBERT, PubMedBERT), with in-domain pre-training yielding substantial factual precision gains (Lee et al., 2020). RAG (Lewis et al., 2020) tackles hallucination by conditioning creation on retrieved evidence. Biomedical adaptations include BioMedRAG (Li et al., 2025), finetuning dense retrievers on PubMed; TrumorGPT (Hang et al., 2025), graph-based ontology retrieval. No existing system addresses all four features of this work: hybrid retrieval, sentence attribution, adversarial robustness, and a composite trust measure.

2.2 Trustworthiness, Hallucination and Calibration

Trustworthy AI in healthcare demands explainability, robustness, fairness, and traceability. Hallucination in biological LLMs is well documented (Guan et al., 2024); mitigations include chain-of-verification (Dhuliawala et al., 2024), token-level uncertainty quantification (Fadeeva et al., 2024), and self-verification via SelfCheckGPT (Manakul et al., 2023). Confidence calibration remains a practical barrier to clinical adoption, motivating the ECE component of the BTI

2.3 Ontology-Aware and Hybrid Retrieval

Ontology-enriched retrieval with MeSH and UMLS lowers vocabulary mismatch in biological IR (Sood & Imran, 2024; Aghaebrahimian et al., 2022). Hybrid BM25+dense retrieval with Reciprocal Rank Fusion (Wang et al., 2021) regularly beats each technique alone. This framework synthesises both within a BTI-certified preservation pipeline – a combination absent from all prior systems.

3 Exposition of the Investigation

3.1 System Architecture

The framework consists of five interconnected modules (Fig. 1): (1) an ontology-aware query processor that expands queries with MeSH/UMLS synonyms; (2) a hybrid re-

trieval engine that combines BM25 and MiniLM-L6 dense retrievers fused via Reciprocal Rank Fusion; (3) a constrained LLM generator (GPT-3.5-Turbo) that grounds every claim in retrieved context; (4) a sentence-level attribution layer that uses Sentence-BERT cosine similarity; and (5) a BTI evaluator that aggregates all trust signals into one.

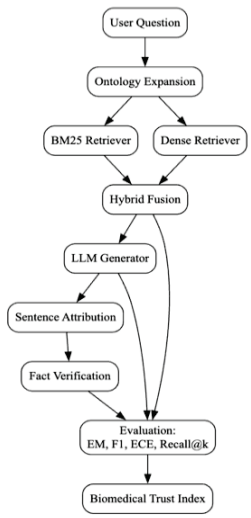

Fig. 1. End-to-end Trustworthy RAG pipeline: from user query through ontology expansion and hybrid retrieval to trust-evaluated answer generation.

3.2 Core Formulations

Six equations define the formal framework. They formalize how the system expands queries to improve recall (Eq. 1), fuses retrieval signals for robustness (Eq. 2), traces each generated claim to its source (Eq. 3), quantifies trust (Eq. 4), measures adversarial degradation (Eq. 5), and assesses calibration (Eq. 6), thereby providing a complete, reproducible mathematical foundation accessible to interdisciplinary readers. Query expansion (Eq. 1) expands query q using UMLS synonyms, bridging the vocabulary gap between lay queries and indexed biomedical terminology. RRF (Eq. 2) combines retriever ranks without learning weights, ensuring reproducibility in heritage systems. Sentence attribution (Eq. 3) assigns each output sentence to the most semantically related chunk retrieved when cosine similarity exceeds the criterion $\theta = 0.75$. The BTI (Eq. 4) combines factual accuracy (F), attribution accuracy (A), and calibration error (E), with customizable weights (α, β, γ). Equal weights $\alpha = \beta = \gamma = 1/3$ are used as a conservative baseline compatible with normal composite metric practice (Albahri et al., 2023). Alternative configurations, such as upweighting F for diagnostic contexts or A for archival traceability, represent a direction for further sensitivity research. Adversarial delta metrics (Eq. 5) quantify performance degradation. ECE (Eq. 6) measures calibration quality across M confidence bins.

$$q_0 = q \cup \{s \mid s \in UMLS\_synonyms(e\_i), e\_i \in E(q)\} \quad (1)$$

$$RRF(d) = \Sigma\_\{r \in retrievers\}\ 1 / (k + rank\_r(d)),\ \ k = 60 \quad (2)$$

$$attr(s\_i) = argmax\_j\ cos(\varphi(s\_i), \varphi(c\_j)),\ \ if\ max\ cos > \theta = 0.75 \quad (3)$$

$$BTI = \alpha \cdot F + \beta \cdot A + \gamma \cdot (1 - E),\ \ \alpha + \beta + \gamma = 1 \quad (4)$$

$$\Delta EM = EM\_clean - EM\_perturbed,\ \ \Delta F1 = F1\_clean - F1\_perturbed \quad (5)$$

$$ECE = \Sigma\_m\ |B\_m| / n \cdot |acc(B\_m) - conf(B\_m)| \quad (6)$$

3.3 Dataset and Evaluation Protocol

MedQuAD, a database of 16,407 NIH-sourced biomedical question-and-answer pairs from 12 institutional sources, was employed in the experiments. A stratified sample of 500 cases was established: 350 pairs for baseline and attribution evaluation, and 150 for adversarial robustness testing across four perturbation types, implemented in the following way to ensure reproducibility: (i) syntactic noise caused character-level typographic errors at a 5% token mutation rate; (ii) semantic drift used WordNet-based near-synonym substitution for key biological terminology; (iii) distractor injection added a randomly chosen, topically irrelevant MedQuAD sentence to the retrieved context; and (iv) contradictory evidence insertion manually reversed the key factual claim in the top-ranked retrieved section. Technology stack: LangChain (orchestration), ChromaDB (dense retrieval), Elasticsearch (BM25 sparse retrieval), Sentence Transformers/all-MiniLM-L6-v2 (attribution), and GPT-3.5-Turbo (generating). The random seed is 42, with an 80/20 stratified split and an RRF constant of 60. Metrics include exact match (EM), F1, recall@5, nDCG@5, MRR@5, attribution error rate (AER), expected calibration error (ECE), and BTI.

4 Results

4.1 Retrieval Performance

The Hybrid RRF based retrieval achieved Recall@5 = 0.89, nDCG@5 = 0.81, and MRR@5 = 0.71, beating BM25 alone (Recall@5 = 0.62) and dense-only retrieval (0.75). Ontology-aware expansion resulted in an additional 9.4% Recall@5 improvement over non-expanded hybrid queries. Table 1 compares retrieval results across all approaches.

Table 1. Retrieval performance comparison across methods (MedQuAD, n = 350).

Retriever	Recall@5	nDCG@5	MRR@5
BM25 (Sparse)	0.62	0.58	0.47
Dense (MiniLM-L6)	0.75	0.70	0.61
Hybrid RRF (Ours)	0.89	0.81	0.71

4.2 Baseline QA Performance and BTI Correlation

The pipeline obtained EM = 61.4% and F1 = 75.2% on the 350-sample assessment set with no disturbances. The EM-F1 gap represents legitimate synonymic and paraphrastic biological answers, which is expected in medical language. Figure 2 shows a continuous positive association between BTI scores and Exact Match across all evaluation samples, indicating that BTI is a credible factual dependability proxy rather than an abstract composite score.

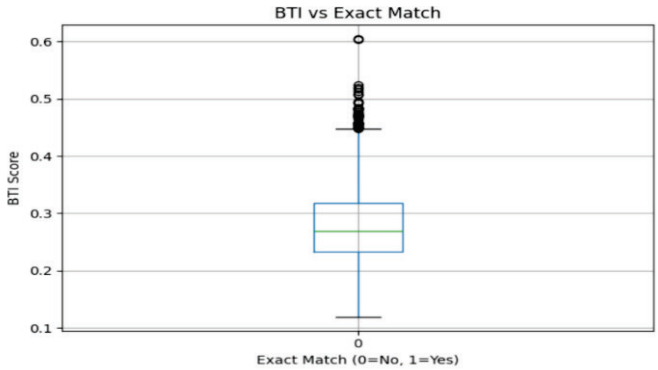

Fig. 2. BTI score vs. Exact Match.

4.3 Sentence-Level Attribution

Out of 1,124 sentences generated over 350 QA samples, 916 (81.5%) received valid attribution with $\theta \geq 0.75$. On clean data, attribution accuracy was 85.6%, but it dropped to 73.2% during adversarial perturbation. The 18.5% of unattributed phrases, which were primarily hallucinations or poorly generalized claims, were automatically flagged for human auditing without each response being manually inspected. Figure 3 depicts the whole attribution score distribution with the $\theta = 0.75$ decision threshold.

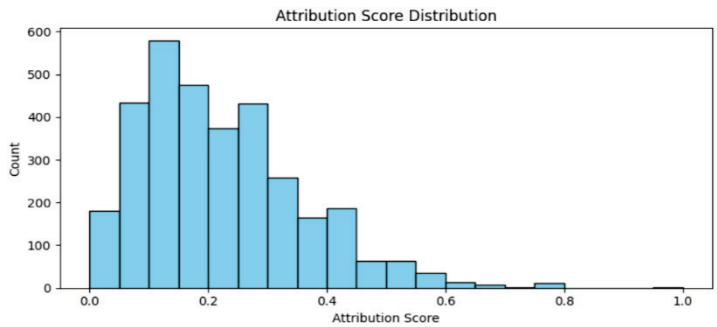

Fig. 3. Sentence-level attribution score distribution.

To demonstrate attribution in practice, a correctly attributed sentence — "Metformin is the first-line pharmacological treatment for type 2 diabetes" — had a cosine similarity

of 0.83 with the top retrieved paragraph and was correctly traced back to its NIH source. A hallucinated statement — "Insulin resistance is reversed within 48 hours of dietary intervention" — received a score of 0.41, falling below the threshold and prompting automatic audit flagging. This shows how sentence-level attribution offers actionable provenance signals without requiring a manual assessment of each generated response.

4.4 Confidence Calibration

The Expected Calibration Error (ECE) for clean data was 0.092, calculated using Eq. 6 over M=10 confidence bins. The mid-range forecasts (confidence level 0.5-0.7) were effectively calibrated. The 0.8-1.0 bin demonstrated systematic overconfidence, as the model awarded high confidence to factually inaccurate outcomes more frequently than bin accuracy justified. The calibration reliability diagram Figure 4 indicates a deviation from the y=x diagonal in the high-confidence region, indicating the requirement for post-hoc temperature scaling prior to clinical or archival deployment.

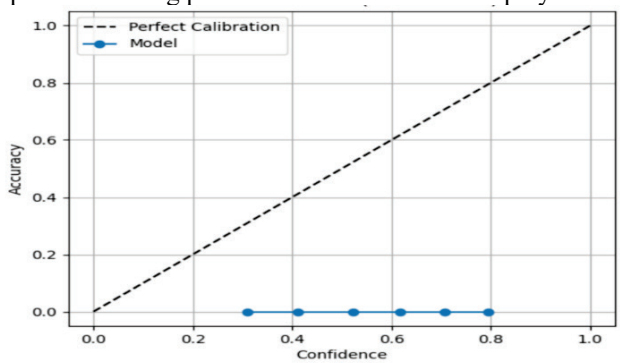

Fig. 4. Calibration reliability diagram.

4.5 Adversarial Robustness and Trust Metric Correlations

Across four perturbation categories, average degradation was $\Delta EM = 7.3\%$ and $\Delta F1 = 6.9\%$ (Eq. 5). Semantic drift and contradictory-evidence injection caused the largest drops; syntactic noise had the least impact. Attribution Error Rate (AER) rose from 12.4% (clean) to 21.6% (perturbed) — the most sensitive trust signal, indicating that adversarial conditions primarily compromise source traceability rather than generation fluency. BTI fell from 0.84 to 0.69 (17.9% reduction), providing a single interpretable signal capturing simultaneous degradation across all trust dimensions. Figure 5 presents the trust metric correlation heatmap: all five dimensions (EM, F1, attribution accuracy, ECE, BTI) are significantly intercorrelated, validating that the BTI components are complementary and non-redundant — each capturing a distinct, independently meaningful facet of trustworthiness. Attribution accuracy emerges as the most sensitive indicator of adversarial stress, suggesting AER should be the primary alert metric in deployed preservation pipelines.

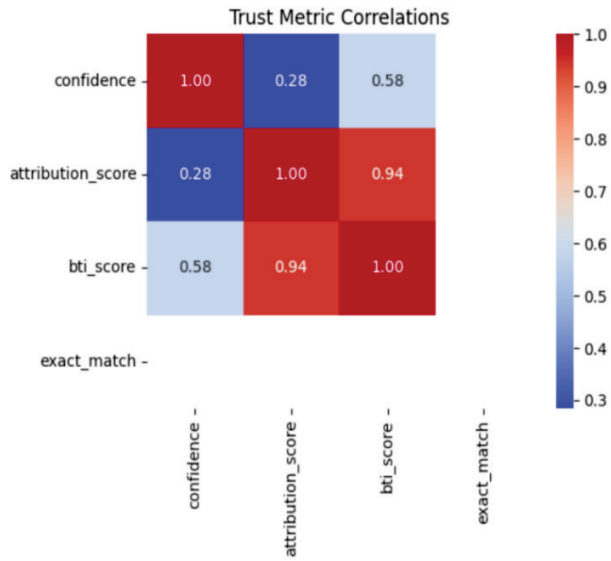

Fig. 5. Trust metric correlation heatmap.

Table 2 provides a comprehensive summary of all system performance metrics across clean and perturbed evaluation conditions, forming a reproducible benchmark for future comparative studies.

Table 2. Full system performance across all trust dimensions — clean vs. perturbed conditions (MedQuAD).

Metric	BM25	Dense (MiniLM)	Hybrid RRF (Ours)	Perturbed (Hybrid)
Recall@5	0.62	0.75	0.89	0.81
nDCG@5	0.58	0.70	0.81	—
MRR@5	0.47	0.61	0.71	—
Exact Match (EM)	—	—	61.4%	54.1%
F1 Score	—	—	75.2%	68.3%
Attribution Accuracy	—	—	85.6%	73.2%
Attr. Error Rate (AER)	—	—	12.4%	21.6%
ECE (Calibration)	—	—	0.092	0.128
BTI (Composite)	—	—	0.84	0.69

5 Discussion

Contributions and Comparison with Prior Work
Hybrid RRF retrieval significantly beats single-method retrieval, and ontology-aware expansion enhances this advantage by bridging vocabulary mismatches that keyword

search cannot resolve. Compared to BioMedRAG (Li et al., 2025), this system adds sentence-level traceability and adversarial evaluation; compared to TrumorGPT (Hang et al., 2025), it generalizes across all biomedical sub-domains without rigid graph construction. The BTI reduces the fragmented trust evaluation landscape identified by Albahri et al. (2023) to a single, context-configurable, institutionally deployable score. In terms of computational scalability, the attribution and ontology modules increased retrieval latency by about 22% over the basic BM25 pipeline. ChromaDB's approximate nearest-neighbour indexing scales sub-linearly with corpus size, allowing deployment on archives far bigger than MedQuAD; nevertheless, GPT-3.5-Turbo's 4,096-token context limit results in evidence truncation for long heritage queries. Asynchronous or batch retrieval over dispersed index shards is regarded as a critical engineering priority for large-scale archival deployment.

5.1 Research Questions Synthesis

RQ1: Eq. 3's sentence-level traceability achieved 85.6% accuracy and flagged 18.5% of outputs without manual review. RQ2: MeSH/UMLS expansion resulted in +9.4% Recall@5. RQ3: Average adversarial degradation was confined to $\Delta F1 = 6.9\%$, with syntactic noise having the lowest impact. RQ4: BTI = 0.84 (clean) vs. 0.69 (perturbed); significant EM correlation indicates validity of trust proxy. RQ5: Attribution and ontology modules increased retrieval latency by approximately 22%, a trade-off justified in heritage preservation contexts where accuracy trumps throughput.

5.2 Limitations

MedQuAD is English-only and strictly structured, it does not reflect the linguistic diversity of real-world multilingual biomedical heritage archives. Future evaluations using corpora such as CLEF eHealth or EHR-Rel-Benchmark will be required to confirm cross-lingual generalizability in international heritage preservation contexts. The reliance on GPT-3.5-Turbo reduces repeatability due to opaque weights and API versioning; BioMedLM (2.7 billion parameters, openly licensed) is recognized as the key contender for reproducible offline replication of this framework. Equivalent BTI weights ($\alpha = \beta = \gamma = 1/3$) adhere to normal composite metric practice; further sensitivity analysis will identify context-optimal configurations for diagnostic versus archival deployment. There was no statistical significance testing performed; hence, the findings should be taken as proof-of-concept baselines pending larger-scale replication.

6 Conclusions

This research argues that trustworthiness in biomedical RAG is a multidimensional, structurally designable concept rather than an emergent attribute of model scale. The proposed methodology, which combines ontology-aware hybrid retrieval, sentence-level attribution, and the Biomedical Trust Index, achieves Recall@5 = 0.89, F1 = 75.2%, attribution accuracy = 85.6%, and BTI = 0.84 on MedQuAD. The metrics

degrade predictably under adversarial stress. The Trust Metric Correlation heatmap (Fig. 5) confirms the multi-component BTI architecture by proving that factual correctness, attribution fidelity, and calibration capture independent and non-redundant trust variation.

The BTI operationalizes trustworthiness as a measurable, context-configurable composite, allowing institutions to compare AI-mediated biomedical knowledge systems to a consistent standard. As biomedical heritage grows and digital preservation becomes more critical, frameworks that incorporate verifiability, provenance, and trust measurement at the architectural level provide a principled path to responsible, transparent, and auditable AI-mediated scientific heritage stewardship.

Acknowledgements

Heartfelt gratitude is extended to Soniya Murthy Rao and Keshav Murthy Balarama for their unwavering support and encouragement.

References

Aghaebrahimian, A., Anisimova, M., & Gil, M. (2022). Ontology-aware biomedical relation extraction. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, speech, and dialogue: 25th International Conference, TSD 2022, Brno, Czech Republic, September 6–9, 2022, proceedings* (Lecture Notes in Computer Science, Vol. 13502, pp. 160–171). Springer. https://doi.org/10.1007/978-3-031-16270-1_14

Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., Albahri, O. S., Alamoodi, A. H., Bai, J., Salhi, A., Santamaría, J., Ouyang, C., Gupta, A., Gu, Y., & Deveci, M. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, *96*, 156–191. https://doi.org/10.1016/j.inffus.2023.03.008

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2024). Chain-of-verification reduces hallucination in large language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 3563–3578). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.212

Fadeeva, E., Rubashevskii, A., Shelmanov, A., Petrakov, S., Li, H., Mubarak, H., Tsymbalov, E., Kuzmin, G., Panchenko, A., Baldwin, T., Nakov, P., & Panov, M. (2024). Fact-checking the output of large language models via token-level uncertainty quantification. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 9367–9385). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.558

Guan, J., Dodge, J., Wadden, D., Huang, M., & Peng, H. (2024). Language models hallucinate, but may excel at fact verification. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 1090–1111). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.62

Hang, C. N., Yu, P. D., & Tan, C. W. (2025). TrumorGPT: Graph-based retrieval-augmented large language model for fact-checking. *IEEE Transactions on Artificial Intelligence*. https://doi.org/10.1109/TAI.2025.3567369

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

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. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems 33* (pp. 9459–9474). Curran Associates. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Li, M., Zhan, Z., Yang, H., Xiao, Y., Zhou, H., Huang, J., & Zhang, R. (2025). Benchmarking retrieval-augmented large language models in biomedical NLP: Application, robustness, and self-awareness. *Science Advances*, *11*(47). https://doi.org/10.1126/sciadv.adr1443

Manakul, P., Liusie, A., & Gales, M. J. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 9004–9017). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.557

Sood, S., & Imran, H. (2024). Information retrieval and query expansion for biomedical data. In A. Sharan, N. Malik, H. Imran, & I. Ghosh, (Eds), *Text mining approaches for biomedical data* (pp. 193–235). Springer Nature Singapore. https://doi.org/10.1007/978-981-97-3962-2_11

Wang, S., Zhuang, S., & Zuccon, G. (2021). BERT-based dense retrievers require interpolation with BM25 for effective passage retrieval. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '21)* (pp. 317–324). Association for Computing Machinery. https://doi.org/10.1145/3471158.3472233

Zhou, Y., Liu, Y., Li, X., Jin, J., Qian, H., Liu, Z., Li, C., Dou, Z., Ho, T.-Y., & Yu, P. S. (2024). Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint*. https://doi.org/10.48550/arXiv.2409.10102

Received: March 15, 2026
Reviewed: May 12, 2026
Finally Accepted: June 04, 2026