Evidence-Calibrated Biomedical Decision Support Using Retrieval-Augmented Large Language Models for Clinical Literature Interpretation and Uncertainty-Aware Diagnosis Prediction

Bahuraksa Adriansyah
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
contactbahuraksa@colostate.edu

Milosz Galos
Department of Computer Science, University of Houston, Houston, TX, USA.
milosz1978@uh.edu

Aadi Bhandari
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
contactaadi@binghamton.edu

Eshana Sibal
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
helloeshana@missouri.edu

Abstract

Biomedical decision support requires both faithful interpretation of clinical literature and calibrated diagnostic prediction, yet retrieval-augmented language models often treat retrieved text as uniformly reliable and produce confidence scores that are poorly aligned with evidence strength. This manuscript develops an evidence-calibrated framework that couples retrieval-augmented large language models with uncertainty-aware diagnosis prediction. The framework represents clinical queries and patient features jointly, scores documents by semantic relevance, evidence level, recency, and risk of bias, and converts retrieval scores into a soft evidence distribution. Generated literature interpretations are decomposed into claims whose evidential support is measured against retrieved spans and study-quality metadata. Diagnostic probabilities are then obtained by integrating over retrieval subsets and model variants, allowing total uncertainty to be separated into aleatoric and epistemic components. A risk-sensitive decision rule combines expected misdiagnosis cost, epistemic uncertainty, and claim-support thresholds to recommend a diagnosis, request further evidence, or defer. The contribution is a connected methodological account of evidence weighting, claim attribution, uncertainty decomposition, calibration assessment, and deferral logic tailored to clinical literature interpretation. The framework does not assume that semantic similarity equals clinical validity, nor that verbal model confidence is sufficient for safety-critical use. It specifies where retrieval error, contradictory evidence, model overconfidence, and cost misspecification enter the decision pathway, and it proposes evaluation logic based on calibration curves, decision-curve analysis, and claim-level evidence auditing. The approach provides a technically defensible basis for evidence-calibrated biomedical decision support prior to clinical validation.

Keywords

retrieval-augmented generation; clinical literature interpretation; uncertainty quantification; calibration; diagnosis prediction; evidence synthesis

1. Introduction

Clinical decision support sits between two imperfect information sources: a patient-specific record that is incomplete and a biomedical literature base that is extensive, heterogeneous, and often contradictory. Large language models can synthesize clinical text and answer medical questions with substantial fluency, but fluency is not evidence. Retrieval-augmented generation reduces some grounding failures by conditioning generation on external documents [1], and transformer architectures provide the representational machinery for joint retrieval and generation [2]. Domain pretraining further improves biomedical language understanding when models are exposed to biomedical corpora [3], [4]. Evaluations of clinical knowledge in large language models suggest meaningful capability, but also inconsistent reasoning and sensitivity to prompt context [5]. Reviews of medical language-model use emphasize benefits in documentation, education, and decision support while cautioning against unsupported clinical deployment [6]. Hallucination remains a central failure mode because generated statements can be syntactically plausible, clinically fluent, and factually unsupported [7].

A second problem is uncertainty. Neural networks are often miscalibrated, assigning high confidence to incorrect predictions [8]. Bayesian approximations and dropout-based uncertainty estimates can expose model uncertainty [9], while retrieval-specific uncertainty controls have been proposed for trustworthy generation and risk management [10]. Ensembles provide a practical alternative for predictive uncertainty [11], and the distinction between aleatoric and epistemic uncertainty clarifies whether additional evidence or a better model could change a decision [12]. In biomedical settings, however, uncertainty is not only statistical. It is also evidential: a diagnosis may be uncertain because the literature is weak, because retrieved studies conflict, or because the patient features do not match the study population. Existing retrieval-augmented systems rarely encode this evidential hierarchy into the prediction layer.

This paper addresses that gap through an evidence-calibrated framework for clinical literature interpretation and uncertainty-aware diagnosis prediction. The central contribution is a coupled retrieval, attribution, and decision model in which document quality affects retrieval weights, generated claims carry explicit support scores, predictive uncertainty is decomposed across retrieval and model variation, and the final action is selected under clinical costs with an explicit deferral option. The framework is methodological and analytical rather than an empirical validation study. It specifies variables, assumptions, equations, failure modes, and evaluation logic needed before clinical use. It also treats safety concerns around large-scale language generation as design constraints rather than afterthoughts [13], [14].

2. Problem Formulation and Evidence Model

A clinical query is represented by q, containing the clinical question, and by a patient feature vector x, containing symptoms, laboratory values, history, and available diagnostic results. Candidate diagnoses form a finite set Y. The evidence corpus D contains biomedical documents d, each with text and metadata. Metadata include study design, sample size, population, outcome definition, publication year, and risk-of-bias assessment. The system

must produce two linked outputs: an evidence-grounded interpretation of the literature relevant to q and x, and a predictive distribution over Y that can support a clinical action.

The first modeling assumption is that retrieval relevance and clinical evidence quality are distinct. A semantically similar case report may be less probative than a lower-similarity systematic review. The second assumption is that generated literature summaries should be auditable at the claim level, not only at the document level. The third assumption is that diagnostic uncertainty should be represented before a decision rule is applied; collapsing uncertainty into a single confidence score obscures whether deferral would help. These assumptions follow from evidence-based medicine, where strength of recommendation depends on quality, consistency, directness, and precision of evidence [15], and from systematic-review practice, where transparent selection and reporting are essential for reproducibility [16].

Let sim(q,x,d) denote a normalized semantic similarity between the clinical query-patient representation and document d. Let e(d) be an evidence-quality score derived from study design and risk of bias, r(d) a recency or currency score, and b(d) a bias or applicability penalty. A retrieval score can be written as a weighted combination:

$$s(d|q,x) = \lambda_1 sim(q,x,d) + \lambda_2 e(d) + \lambda_3 r(d) - \lambda_4 b(d)$$

Here λ_1 to λ_4 are nonnegative weights that sum to one, and each component is normalized to a common interval. The weights are design parameters, not estimated quantities in this formulation. The equation makes explicit that semantic similarity is only one input. A high-quality guideline or systematic review can outrank a topically similar but methodologically weak report when λ_2 is appreciable. The penalty term b(d) addresses population mismatch, outdated evidence, or high risk of bias. This retrieval score is then converted into a soft distribution over documents:

$$p(d|q,x) = exp(s(d|q,x)/\tau) \; / \; \Sigma_\{d' \in D\} \, exp(s(d'|q,x)/\tau)$$

The temperature τ controls how sharply retrieval concentrates on the highest-scoring documents. Low τ produces nearly hard retrieval; high τ preserves a broader evidence set. In clinical use, a broad distribution is useful when the literature is contested, because it exposes the model to multiple positions and allows contradiction to be represented. A narrow distribution is useful for focused questions but increases the risk that retrieval error becomes invisible. The distribution p(d|q,x) is not a probability that document d is true. It is a normalized retrieval weight conditioned on the query, patient features, and metadata.

3. Evidence-Calibrated Retrieval-Augmented Generation

The generator receives the query, patient features, and a retrieved evidence set R. It produces a literature interpretation S composed of atomic claims c. Each claim is checked against retrieved spans g within documents d. A claim-support score can be defined as a maximum similarity between the claim and an evidence span, weighted by document quality:

$$A(c|R) = max_\{d \in R\} \, max_\{g \subset d\} \, sim(c,g) \, e(d)$$

The symbol g denotes a contiguous span or structured evidence statement such as a recommendation, effect estimate, or diagnostic criterion. The score A(c|R) is bounded after normalization. It is not a truth value; it measures whether a claim is attributable to high-quality retrieved evidence. A low score flags a claim for revision, removal, or explicit hedging. A high score does not eliminate the need for contradiction checking, because two high-quality sources may support opposing claims. For that reason, the framework separates

support from consistency: support measures attributability to at least one source, while consistency measures agreement across sources.

The generator is constrained to produce claims only when A(c|R) exceeds a minimum threshold, or to mark unsupported material as uncertain. This constraint changes the generation objective from fluent summarization to auditable synthesis. The model may still paraphrase, but each paraphrase must remain traceable to a span and a document quality profile. The interpretation output therefore includes the claim, its supporting document identifiers, its support score, and any detected contradiction. This design responds directly to hallucination risks in natural language generation [7] and to the broader concern that large language models can reproduce plausible but ungrounded content at scale [13].

For diagnosis prediction, retrieval and generation are not separate pipelines. The retrieved evidence informs the prior plausibility of diagnoses, the patient features inform likelihoods, and the language model can supply structured mappings from evidence statements to diagnostic criteria. A practical predictive distribution averages over K retrieval-model configurations:

$$\bar{p}(y|x,q) = 1/K\ \Sigma_\{k = 1\}\^K\ p(y|x,q,R_k,\theta_k)$$

Here R_k is a retrieval subset sampled from p(d|q,x), θ_k denotes a model variant or stochastic decoding configuration, and p(y|x,q,R_k,θ_k) is the diagnostic distribution produced under that configuration. The averaging operation is a design choice that reduces dependence on a single retrieval set and a single model path. It also creates a representation of uncertainty that can be decomposed. The distribution $\bar{p}$(y|x,q) is a predictive summary, not a validated clinical probability until calibration is demonstrated on appropriate data.

4. Uncertainty-Aware Diagnosis Prediction and Decision Rule

Total predictive uncertainty can be separated into aleatoric and epistemic components. Let H(·) denote Shannon entropy. The total entropy of the averaged predictive distribution is H($\bar{p}$). The expected conditional entropy across retrieval-model configurations is the aleatoric component. The difference is the epistemic component:

$$U_total = H(\bar{p});\ U_ale = 1/K\Sigma\ H(p_k);\ U_epi = U_total - U_ale$$

The aleatoric component reflects irreducible ambiguity in the patient presentation or in the outcome definition. The epistemic component reflects disagreement caused by retrieval choice, model variation, or limited evidence. This distinction matters clinically. High aleatoric uncertainty may call for additional diagnostic testing or a probabilistic discussion with the patient. High epistemic uncertainty may call for broader literature retrieval, expert review, or deferral. The decomposition is not a guarantee that model ensembles capture all relevant uncertainty; it is a structured diagnostic that makes the sources of disagreement inspectable.

Calibration is assessed separately from entropy. A model can be uncertain and calibrated, or confident and miscalibrated. For a set of predictions partitioned into M confidence bins B_m, expected calibration error is:

$$ECE = \Sigma_\{m = 1\}\^M\ |B_m|/N\ |acc(B_m) - conf(B_m)|$$

Here N is the number of predictions, acc(B_m) is the observed accuracy in bin m, and conf(B_m) is the mean predicted confidence in that bin. In a clinical evaluation, calibration should be reported with reliability diagrams and stratified by diagnosis prevalence, site, and

patient subgroup. The equation is an assessment definition, not an empirical result. It indicates what must be measured before a predictive distribution is allowed to influence care.

The final action is selected from the diagnosis set plus a deferral action, denoted $\perp$. Let C(a,y) be the cost of action a when the true diagnosis is y. Let A_min be the minimum claim-support score among claims used for the diagnosis, and let λ_U and λ_A be nonnegative penalties for epistemic uncertainty and weak evidence support. A risk-sensitive decision rule is:

$$a* = argmin_\{a \in Y \cup \{\perp\}\} [\Sigma_y\, C(a,y)\bar{p}(y|x,q) + 1\{a = \perp\}(\lambda_U\, U_epi + \lambda_A(1 - A_min))]$$

The indicator $1\{a=\perp\}$ activates the deferral penalty only when the deferral action is considered. The cost matrix C(a,y) encodes clinical consequences, including the harm of missed diagnosis, unnecessary treatment, and delayed care. Cost-sensitive decision theory provides the formal basis for such expected-cost selection [17], [18]. The additional terms prevent deferral from being treated as free and prevent diagnosis from being selected when evidence support is weak. In practice, the cost matrix and penalties require clinical governance and sensitivity analysis, because different specialties assign different harms to false negatives, false positives, and delay.

The predictive model and evidence-weighting parameters can be trained with a composite objective that combines classification loss, calibration loss, claim-support loss, and retrieval regularization:

$$L(\theta, \varphi) = L_CE(\theta) + \alpha L_Brier(\theta) + \beta L_support(\theta, \varphi) + \gamma L_ret(\varphi)$$

The symbol θ denotes generation and prediction parameters, φ denotes retrieval-scoring parameters, and α, β, and γ control the relative emphasis of calibration, evidence support, and retrieval regularization. L_CE is a standard cross-entropy term, L_Brier penalizes overconfident probability estimates, L_support penalizes unsupported claims, and L_ret regularizes the retrieval distribution toward evidence-quality features. This is a proposed training objective. It is not claimed that the objective has been optimized or clinically validated here; its purpose is to make explicit which behaviors the architecture is designed to reward.

5. Technical Discussion and Boundary Conditions

The framework's main technical advantage is that it refuses to collapse three distinct quantities into one: document relevance, claim support, and diagnostic probability. Retrieval-augmented generation often improves factuality by grounding generation in external text [1], but grounding is only as good as the retrieval distribution. If the retrieval score overweights semantic similarity, a clinically irrelevant case report can dominate a systematic review. If recency is overweighted, established evidence may be replaced by newer but weaker studies. The λ weights and temperature τ therefore require sensitivity analysis. A useful sensitivity report would vary λ_2 and τ, show how the retrieved evidence set changes, and describe how claim support and deferral rates respond. Such analysis is a design requirement even before empirical validation.

A second boundary condition is contradictory evidence. The claim-support score A(c|R) rewards attributability, but it does not by itself resolve conflict. Two claims may both be supported and mutually inconsistent. The generated interpretation must therefore represent contradiction explicitly, for example by reporting the source populations, effect directions,

and risk-of-bias differences. Diagnostic prediction should then inherit this conflict as epistemic uncertainty. This is preferable to forcing a single summary sentence that hides disagreement. The distinction between support and consistency is essential in clinical literature interpretation, where guidelines, randomized trials, observational studies, and mechanistic reasoning carry different inferential weight [15].

A third boundary condition is overconfidence in the language model. Verbalized uncertainty can be informative but is not a substitute for calibrated probabilities [19], [20]. A model may state that it is uncertain while assigning a high probability to a diagnosis, or it may produce confident prose from weak evidence. The proposed decomposition uses ensemble and retrieval variation rather than relying solely on verbal hedges. Even so, ensembles can be underdispersed if all model variants share the same pretraining bias. External calibration on held-out clinical data, ideally with site-level and subgroup reporting, remains necessary. The evaluation logic should include reliability diagrams, expected calibration error, Brier scores, and decision-curve analysis, with pre-specified outcomes and transparent reporting consistent with systematic-review standards [16].

A fourth boundary condition is cost misspecification. The decision rule depends on C(a,y), and clinical costs are context-dependent. A high-sensitivity rule for a treatable emergency is not appropriate for a low-prevalence screening context. The deferral penalty terms λ_U and λ_A are also policy choices. They should be set through clinical review and stress-tested against plausible cost ranges. Cost-sensitive learning formalizes the optimization [18], but it does not remove the need for institutional judgment. The framework therefore treats the decision rule as a transparent policy layer rather than an autonomous diagnostic authority.

A fifth boundary condition is corpus quality and retrieval poisoning. If the evidence corpus contains retracted studies, predatory publications, or biased guidelines, retrieval calibration can amplify those weaknesses. Metadata audits, retraction checks, and provenance tracking are required. The retrieval distribution $p(d|q,x)$ should be stored with the generated interpretation so that auditors can inspect which documents influenced which claims. This provenance record also supports error analysis: a wrong diagnosis can be traced to retrieval omission, claim misattribution, model prior override, or cost misspecification.

Finally, the framework's evaluation must distinguish mechanistic plausibility from observed clinical benefit. A model can produce well-supported literature summaries and still fail to improve patient outcomes. Conversely, a model can improve a narrow prediction task while providing poor evidence explanations. The appropriate evaluation therefore has at least three layers: claim-level evidence auditing, predictive calibration and discrimination, and decision-level utility under clinical costs. Each layer answers a different question. Claim auditing asks whether statements are attributable and consistent; calibration asks whether probabilities match observed frequencies; decision analysis asks whether actions improve expected outcomes. No single metric substitutes for the others.

6. Conclusion

This manuscript has developed an evidence-calibrated framework for biomedical decision support that connects retrieval-augmented language modeling, claim-level evidence attribution, uncertainty decomposition, and risk-sensitive diagnosis prediction. The framework treats retrieval as a quality-weighted evidence distribution rather than a similarity ranking, requires generated claims to carry support scores, separates aleatoric from epistemic uncertainty, and selects among diagnosis and deferral actions under explicit clinical costs.

These design choices address known failure modes of language models in medicine: hallucinated evidence, overconfident prediction, hidden contradiction, and unexamined retrieval error. The contribution is methodological and analytical. It provides equations, assumptions, boundary conditions, and evaluation logic for a system that could be validated prospectively, but it does not claim clinical validation or empirical performance. The central implication is that trustworthy biomedical decision support requires evidence calibration at the level of documents, claims, probabilities, and actions. A model that interprets literature without representing evidence quality, or predicts diagnosis without representing uncertainty, leaves the most consequential sources of error outside the decision pathway.

References

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (Vol. 33, pp. 9459–9474).

2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5998–6008).

3. 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.

4. 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.

5. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Arcas, B. A., ... Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180.

6. Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29(8), 1930–1940.

7. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.

8. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 1321–1330). PMLR.

9. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (Vol. 48, pp. 1050–1059). PMLR.

10. Zhang, R. (2025, December). Uncertainty-Aware RAG: Trustworthy Generation and Risk Control. In 2025 5th International Symposium on Artificial Intelligence and Big Data (AIBDF) (pp. 526-531). IEEE.

11. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems (Vol. 30, pp. 6402–6413).

12. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems (Vol. 30, pp. 5574–5584).

13. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). ACM.

14. Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. Healthcare, 11(6), 887.

15. Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. BMJ, 336(7650), 924–926.

16. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71.

17. Sox, H. C., Higgins, M. C., & Owens, D. K. (2013). Medical decision making (2nd ed.). Wiley-Blackwell.

18. Elkan, C. (2001). The foundations of cost-sensitive learning. In Proceedings of the 17th International Joint Conference on Artificial Intelligence (pp. 973–978). Morgan Kaufmann.

19. Kadavath, S., Conerly, T., Askell, 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.

20. Lin, S., Hilton, J., & Evans, O. (2022). Teaching models to express their uncertainty in words. arXiv preprint arXiv:2205.14334.