RURALDOC: EVIDENCE-BASED KNOWLEDGE GRAPH RETRIEVAL FOR SKIN DISEASE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Structured external knowledge is often introduced to improve large language model (LLM) reasoning, yet poorly integrated knowledge can instead override correct model beliefs. We study this failure mode in RuralDoc, a multimodal framework for skin-disease reasoning that integrates retrieval-augmented generation (RAG) with a hierarchical diagnostic knowledge graph (KG). An initial ranking-based design presented KG predictions directly to the LLM and degraded performance for most backbones we tested, with deference analysis showing that models frequently abandoned correct initial answers when the KG ranking was wrong. We therefore redesign the KG as an evidence provider rather than a diagnostic authority: the revised method uses IDF-weighted matching over diagnostic symptoms and visual features, withholds disease rankings and scores from the LLM, and presents retrieved facts as unordered reference evidence alongside dermatology context. Across 192 diagnostic cases and five language models, the combined redesign removes the degradation observed with the original KG integration and improves Top-1 accuracy from 70.3% to 78.6% for Llama-3.1-8B and from 88.1% to 93.8% for DeepSeek-V3.1 relative to model-only baselines. In cases where the KG's top prediction is wrong but the model's initial answer is correct, erroneous deference drops sharply, from 67% to 7% for Llama-3.1-8B, with similar reductions for the other four models. These results show that the way structured knowledge is retrieved and exposed to an LLM can materially affect downstream reasoning and motivate treating KG outputs as supporting evidence rather than authoritative predictions.

1 INTRODUCTION

Large language models (LLMs) increasingly rely on external information to improve factuality and domain-specific reasoning. Retrieval-augmented generation (RAG) augments model inputs with relevant textual context, while knowledge graphs (KGs) make structured relationships among entities explicit. Recent work has explored graph-based retrieval and KG-augmented prompting for question answering and reasoning, including methods that retrieve relevant subgraphs, encode graph structure into prompts, or combine graph evidence directly with LLM generation (Kim et al., 2023; He et al., 2024; Ao et al., 2025). In medical applications, graph-based RAG systems have similarly been proposed to improve the grounding and reliability of evidence (Wu et al., 2025; Zhao et al., 2025).

However, adding structured knowledge does not necessarily improve reasoning. A retrieval system may surface incomplete or noisy evidence, and the manner in which evidence is presented may also affect how strongly an LLM relies on it versus its own internal knowledge. This is especially consequential in diagnostic reasoning, where an externally produced disease ranking can appear authoritative even when the underlying retrieval heuristic is imperfect. In RuralDoc, we observe that a redesign combining improved retrieval scoring with non-authoritative evidence presentation substantially reduces this kind of erroneous deference; we discuss in Section 8 what would be needed to attribute this effect to presentation format alone.

We investigate this problem through **RuralDoc**, a multimodal framework of reasoning about skin-diseases motivated by limited access to dermatology expertise in rural and resource-constrained settings. Skin-disease diagnosis often depends jointly on the appearance of the lesion and contextual information such as itching, pain, body location, morphology, and progression of the disease.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

RuralDoc combines image retrieval, symptom descriptions, a structured skin-disease KG, dermatology reference retrieval, and LLM reasoning, with the long-term goal of a smartphone-oriented diagnostic support tool.

During development, we found that the original KG integration could reduce rather than improve diagnostic accuracy. The initial design scored candidate diseases via overlap in the symptom list and exposed the resulting ranking list directly to the LLM as diagnostic evidence. Although the graph contained clinically relevant relationships, the scoring heuristic was noisy, and models frequently abandoned an answer they had previously produced correctly in favor of the KG's incorrect top prediction. This motivated a redesign in which the KG retrieves potentially relevant facts using IDF-weighted matching, but its internal rankings and scores are withheld from the LLM; the retrieved facts are instead presented as unordered supporting evidence, and the model is instructed to reason independently.

Our contributions are threefold:

1. We identify and quantify **erroneous deference to ranked KG predictions**, showing that KG augmentation can degrade LLM diagnostic reasoning even when the underlying graph contains clinically relevant information.
2. We propose an **evidence-based KG retrieval formulation** that retrieves diagnostic facts while withholding candidate rankings and scores, separating evidence retrieval from final decision-making.
3. We evaluate the approach across five LLM backbones and 192 diagnostic cases, showing that the redesigned evidence-based configuration removes most of the degradation produced by the original ranking-based system and yields the largest gains for smaller models.

2 Related Work

2.1 Multimodal AI for Medical and Dermatological Reasoning

Medical diagnosis often requires integrating information across modalities: in dermatology, lesion appearance must be interpreted together with symptoms, anatomical location, duration, and patient history. Multimodal medical models such as LLaVA-Med (Li et al., 2023) and Med-PaLM M (Tu et al., 2023) demonstrate the potential of joint image-text reasoning for clinical tasks, but multimodal models can still hallucinate findings or perform unevenly on underrepresented conditions. This motivates architectures that supplement a model's visual judgment with explicit external knowledge rather than relying on visual similarity alone.

RuralDoc differs from a standalone multimodal classifier in that it combines image-derived candidates with structured diagnostic knowledge and retrieved textual evidence, specifically to help distinguish diseases with overlapping visual presentation.

2.2 Knowledge Graphs and LLM Reasoning

Several recent approaches combine KGs with LLMs by retrieving relevant graph structures and exposing them to the model. KG-GPT (Kim et al., 2023) separates retrieval and reasoning over graph structures; G-Retriever (He et al., 2024) retrieves relevant portions of textual graphs before generation; LightPROF (Ao et al., 2025) follows a retrieve–embed–reason paradigm that converts relevant graph information into an LLM-compatible representation; and KnowGPT (Zhang et al., 2024) constructs structured prompts while reducing knowledge noise.

These methods primarily optimize *which* graph information is retrieved or encoded. RuralDoc addresses a complementary question: how should retrieved KG knowledge be presented when it conflicts with the LLM's own answer, and how does presentation format interact with diagnostic accuracy?

2.3 Medical Graph RAG

Medical KG and graph-RAG systems seek to ground LLM responses in structured domain knowledge. DrugRep-KG represents heterogeneous biomedical entities relationally (Ghorbanali et al.,

2023), while MedRAG uses a diagnostic KG to expose differences between clinically similar diseases (Zhao et al., 2025). MedGraphRAG extends graph-based retrieval to evidence-based medical generation by linking source documents to credible medical references and controlled vocabularies (Wu et al., 2025).

RuralDoc adopts a related principle but differs in two respects. First, it targets multimodal dermatology rather than EHR-based or general medical question answering. Second, our experiments directly compare exposing KG outputs as ranked diagnostic predictions versus unordered supporting evidence, treating presentation format as a variable of interest alongside retrieval quality.

3 RURALDOC FRAMEWORK

3.1 OVERVIEW

RuralDoc accepts a skin image together with a symptom description and optional body location. The framework has four components: (1) image retrieval, (2) DermNet knowledge retrieval, (3) diagnostic KG evidence retrieval, and (4) LLM reasoning over the combined evidence.

3.2 IMAGE RETRIEVAL

Skin images are encoded with a DINOv2 image encoder (Oquab et al., 2023) and indexed with FAISS (Johnson et al., 2019). The image database contains 27,195 images across the 12 target diseases, drawn from ISIC 2019 (Gessert et al., 2019), SD-198 (Sun et al., 2016), and a Mendeley scabies dataset (Mendeley Data). For an uploaded image, its embedding is compared against this index, and disease labels associated with the retrieved neighbors form the image-derived candidate set.

Comparative tests showed DINOv2-small and DINOv2-base achieve similar retrieval performance (approximately 50% Top-1 and 82% Top-3 accuracy on held-out images), suggesting that simply scaling the general-purpose encoder did not substantially improve retrieval in our setting.

3.3 DIAGNOSTIC KNOWLEDGE GRAPH

The skin-disease KG covers 12 diseases organized into four top-level categories and eight subcategories, with 143 nodes and 170 edges spanning eight relation types: `has_category`, `has_subcategory`, `includes_disease`, `has_symptom`, `has_visual_feature`, `key_diagnostic_difference`, `similar_to`, and `requires_referral_if`. The 12 target diseases comprise seven conditions with strong visual overlap in the source datasets and five additional conditions selected for rural-setting relevance or clinical similarity.

The KG stores compact diagnostic relationships (symptoms, visual features, differentiating criteria, and referral triggers), while longer treatment, medication, and follow-up guidance is maintained separately in a DermNet-derived reference store used for RAG context (DermNet New Zealand Trust, 2026).

4 FROM RANKING-BASED KG INTEGRATION TO EVIDENCE RETRIEVAL

4.1 INITIAL RANKING-BASED DESIGN

The initial implementation scored diseases by lexical overlap between patient-reported symptoms and KG facts, ranked diseases by this score, and exposed the resulting list to the LLM as KG-derived diagnostic evidence. The same ordering was used to arrange associated DermNet reference material, so the KG's preferred diagnosis was communicated to the LLM both through the explicit ranking and through the ordering of retrieved context.

This design implicitly assumed a higher KG score reflected stronger diagnostic evidence. In practice the scorer was weak: generic terms (e.g., "red" and "skin") contributed disproportionately, longer disease descriptions had more matching opportunities regardless of specificity, and raw lexical over-

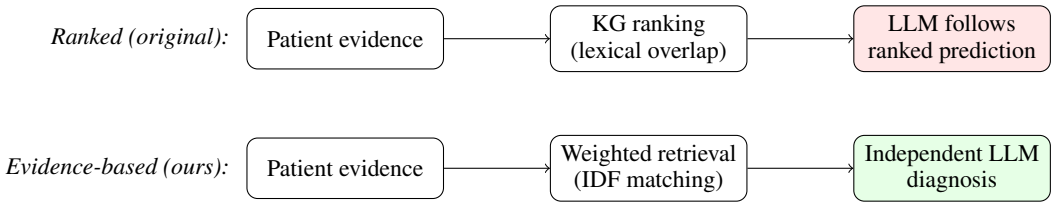


Figure 1: Contrast between the original ranking-based KG integration and the redesigned evidence-based configuration. The ranked design exposes an explicit disease ranking to the LLM; the evidence-based design retrieves relevant KG evidence but withholds rankings, presenting facts as unordered evidence instead.

lap did not reliably track diagnostic relevance. The initial KG scorer alone achieved approximately 45% Top-1 accuracy.

4.2 Erroneous Deference

The failure became apparent when comparing model-only predictions against KG-augmented predictions: in many cases, an LLM initially selected the correct disease but changed its answer after receiving an incorrect KG ranking. We define **erroneous deference** on the subset of cases where (1) the model-only prediction is correct and (2) the KG's top-ranked prediction is incorrect; an erroneous-deference event occurs when KG augmentation causes the model to abandon its correct initial diagnosis in favor of the KG's prediction.

This metric characterizes how models respond when an external KG prediction conflicts with an initially correct model answer. Because the two system variants also differ in retrieval scoring, it should not be interpreted as isolating presentation effects independently of retrieval quality.

4.3 Evidence-Based KG Retrieval

We redesign the pipeline so the KG acts as an evidence source rather than a classifier, changing three things simultaneously:

1. **Retrieval scoring.** Disease facts are weighted by inverse document frequency (IDF), up-weighting features specific to few diseases and downweighting generic symptoms shared across many. Retrieval is restricted to positive diagnostic relations such as `has_symptom` and `has_visual_feature`.

2. **Information hiding.** Rankings and scores are used only internally to select which facts to retrieve; they are never exposed to the LLM.

3. **Presentation.** Retrieved facts are serialized as unordered textual evidence, and DermNet reference text is likewise presented without KG-ranking-based ordering. The prompt explicitly tells the LLM that retrieved evidence may be incomplete or irrelevant and that it should independently determine the most likely diagnosis.

The resulting information flow is

$$\text{patient evidence} \rightarrow \text{weighted retrieval} \rightarrow \text{KG facts + DermNet context} \rightarrow \text{LLM diagnosis},$$

rather than an explicit KG disease ranking being passed to the model as a diagnostic recommendation.

Because these changes are introduced together, our experiments support the combined claim that the redesign improves performance; they do not isolate how much of the gain is attributable to improved retrieval versus withholding rankings.

Table 1: Top-1 diagnostic accuracy (%) across model-only, original ranking-based KG+RAG, and redesigned evidence-based KG+RAG configurations.

Model	LLM Only	Ranked KG+RAG	Evidence KG+RAG
Llama-3.1-8B	70.3	57.8	**78.6**
Qwen2.5-72B	84.9	78.1	**85.4**
GPT-5	91.1	87.0	**92.7**
DeepSeek-V3.1	88.1	88.0	**93.8**
Gemma-4-31B	93.8	92.2	**95.3**

5 Experimental Setup

5.1 Diagnostic Benchmark

The evaluation benchmark is built from 96 symptom vignettes, organized as a complete grid of 12 diseases $\times$ 4 difficulty tiers (textbook, lay, sparse, ambiguous) $\times$ 2 variants per tier. Vignettes were initially drafted with LLM assistance and then reviewed and edited by the two student authors. No dermatologist or other clinical domain expert reviewed the case content, which we treat as a limitation rather than as a clinically validated resource.

From these 96 vignettes, a test suite of 240 cases was assembled: 96 confusion cases (candidate options limited to diseases visually or clinically similar to the gold label), 96 open-12 cases (candidate options drawn from the full 12-disease set), and 48 unanswerable cases intended to probe abstention behavior rather than forced-choice accuracy. The experiments in this paper use the 192 cases with a determinate gold answer—the confusion and open-12 subsets combined. Answer options are shuffled per case so models cannot exploit a fixed answer position.

5.2 Models

We evaluate five LLM backbones: Llama-3.1-8B, Qwen2.5-72B, GPT-5, DeepSeek-V3.1, and Gemma-4-31B. Each model is evaluated under three conditions using the same 192 cases: (1) LLM only, (2) ranking-based KG+RAG, and (3) evidence-based KG+RAG. Model decoding settings are held fixed across conditions. The two KG configurations differ in both retrieval scoring and evidence presentation, as described above.

5.3 Metrics

The primary metric is Top-1 diagnostic accuracy. We additionally report erroneous-deference rate on the subset of cases where the model-only answer is correct but the KG's top prediction is wrong. Supporting experiments evaluate image retrieval and end-to-end multimodal performance.

6 Results

6.1 Ranking-Based KG Integration Reduces Accuracy for Most Backbones

Ranking-based KG augmentation reduces accuracy for four of the five models tested, most sharply for Llama-3.1-8B (70.3% $\rightarrow$ 57.8%) and GPT-5 (91.1% $\rightarrow$ 87.0%). DeepSeek-V3.1 is essentially unaffected (88.1% $\rightarrow$ 88.0%). These results show that adding structured knowledge does not by itself guarantee improved reasoning.

6.2 Evidence-Based Retrieval Removes the Degradation

Under the redesigned evidence-based configuration, which combines IDF-weighted retrieval with hidden rankings and unordered evidence presentation, performance recovers to or exceeds the model-only baseline for all five backbones. Llama-3.1-8B improves from 70.3% to 78.6% (+8.3 points); DeepSeek-V3.1 improves from 88.1% to 93.8% (+5.7 points); and Gemma-4-31B improves

Table 2: Erroneous-deference rate (%) under the original ranked and redesigned evidence-based configurations.

Model	Ranked KG	Evidence KG
Llama-3.1-8B	67	**7**
Qwen2.5-72B	25	**13**
GPT-5	12	**3**
DeepSeek-V3.1	11	**0**
Gemma-4-31B	7	**0**

from 93.8% to 95.3% (+1.5 points). Gains are largest for the smallest backbone, and no model shows the systematic degradation observed under the original ranking-based configuration.

6.3 Erroneous Deference

Deference is evaluated relative to the KG top prediction produced by the corresponding retrieval configuration in each condition—the ranked condition's lexical-overlap scorer and the evidence condition's IDF-weighted scorer, respectively—even though this top prediction is never shown to the model under evidence-based prompting. Because retrieval scoring also changes between the two systems, this analysis measures the behavioral effect of the complete redesign rather than isolating presentation format alone. Deference nonetheless drops sharply across all models, most dramatically for Llama-3.1-8B (67% $\rightarrow$ 7%).

6.4 Effects by Difficulty

For Llama-3.1-8B, evidence-based augmentation provides little improvement on canonical textbook-style cases but substantially improves sparse and ambiguous cases, with gains of approximately 14.5 and 12.5 percentage points respectively. This pattern is consistent with external knowledge being particularly useful when the input contains less explicit diagnostic evidence.

6.5 Multimodal Evaluation

Separate multimodal experiments assess whether image-derived candidates contribute useful information. Held-out image evaluations show approximately 50% image-only Top-1 accuracy and approximately 82% Top-3 retrieval accuracy using DINOv2. Combining image-derived candidates with symptoms, KG evidence, and RAG substantially increases end-to-end accuracy relative to image-only inference. These results support the broader RuralDoc architecture, but the primary KG-integration experiments above are evaluated on text-based diagnostic cases, and the redesigned evidence-based prompting scheme has not yet been evaluated in the full multimodal setting at the same scale.

7 Discussion

These findings suggest that the usefulness of external structured knowledge depends not only on its factual content but also on its interaction with the LLM. A ranked candidate list implicitly carries a decision signal; when generated by an imperfect retrieval heuristic, an LLM may treat it as more authoritative than the underlying evidence warrants.

Evidence-based retrieval separates two roles often conflated in KG-augmented systems: **retrieval** (which facts are potentially relevant) and **decision-making** (how those facts should be weighed against the model's own knowledge). The retrieval mechanism determines candidate evidence; the LLM remains responsible for integration. The larger gains on sparse and ambiguous cases further motivate future systems where KG retrieval is invoked adaptively rather than uniformly.

8 LIMITATIONS

- **Confounded redesign.** The transition from ranking-based to evidence-based KG integration changes both the retrieval scoring function (raw lexical overlap → IDF-weighted matching) and the presentation format (ranked → unordered, rankings hidden) simultaneously. Our results support the combined intervention but do not decompose how much of the gain is attributable to better retrieval versus withholding rankings.
- **Sample size and reporting.** The benchmark covers 192 cases across 12 diseases; several reported statistics are computed on smaller subsets, and no variance or significance testing is currently included.
- **Benchmark scope.** The benchmark covers 12 skin diseases and structured diagnostic descriptions rather than the full breadth of dermatology or a prospective clinical setting.
- **No clinical expert validation.** Benchmark vignettes were drafted with LLM assistance and edited by the two student authors, without review by a dermatologist or other clinical domain expert.
- **Multimodal evaluation gap.** The central KG-integration analysis is evaluated on text-based cases; a full image+text evaluation using the redesigned prompting scheme at comparable scale remains future work.
- **Retrieval mechanism.** The current retriever uses IDF-weighted lexical matching rather than a learned clinical retriever.
- **Not clinically validated.** RuralDoc is intended as decision support and has not been validated for autonomous diagnosis or treatment recommendation.

9 CONCLUSION

We find that the original ranking-based KG integration can harm LLM diagnostic reasoning, while a redesigned evidence-based configuration substantially reduces erroneous deference and improves accuracy across five LLM backbones. Recasting the KG as an evidence provider—retrieving via weighted matching while withholding rankings and scores—removes the degradation observed under the original ranking-based design, with the largest gains for smaller models. These findings motivate surfacing structured knowledge as evidence to reason over rather than as a prediction to follow, while highlighting the need for a cleaner ablation to isolate how much of this effect is attributable to presentation format alone.

ETHICS STATEMENT

RuralDoc is a research prototype intended to study how structured knowledge should be exposed to LLMs; it is not a validated diagnostic device and should not be used to make real clinical decisions. The diagnostic benchmark was authored by non-clinician student researchers and has not been reviewed by a dermatologist (Section 8), so absolute accuracy numbers should not be interpreted as clinical-grade performance. All image datasets used in this work (ISIC, SD-198, and the Mendeley scabies dataset) are publicly released research datasets. We did not collect any new patient data for this study and did not intentionally use personally identifying information.

AI USE STATEMENT

Large language models were used in the preparation of this work as follows:

- **Synthetic data generation.** The 96 symptom vignettes underlying the diagnostic benchmark were initially drafted with LLM assistance and then reviewed and edited by the two student authors. No professional clinical review of benchmark content was performed.
- **Methodology and experimental design.** LLM assistance was used during exploratory discussion of experimental framing, including interpretation of the ranked-versus-evidence comparison and identification of the retrieval-scoring/presentation-format confound. The authors designed and implemented the underlying system, retrieval methods, experiments, and evaluation pipeline, and independently verified all reported results.

- **Manuscript drafting and editing.** LLM assistance was used to help draft, restructure, and copy-edit portions of this manuscript based on author-provided experimental results, code, and prior draft material. All experimental claims, numbers, and figures were verified by the authors against the underlying code and results before submission.

The authors take full responsibility for the accuracy and integrity of the submitted content, including all AI-assisted portions.

Reproducibility Statement

The paper reports the benchmark construction procedure, evaluation conditions, graph structure, retrieval configuration, and model comparisons used in the experiments. Additional implementation details, prompts, and per-case outputs can be included in supplementary material where available.

References

Tu Ao, Yanhua Yu, Yuling Wang, Yang Deng, Zirui Guo, Liang Pang, Pinghui Wang, Tat-Seng Chua, Xiao Zhang, and Zhen Cai. LightPROF: A lightweight reasoning framework for large language model on knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23424–23432, 2025. doi: 10.1609/aaai.v39i22.34510.

DermNet New Zealand Trust. Dermnet. `https://dermnetnz.org`, 2026. Accessed 2026.

Nils Gessert, Mads Nielsen, Mohsin Shaikh, René Werner, and Alexander Schlaefer. Skin lesion classification using ensembles of multi-resolution efficientnets with meta data. *arXiv preprint arXiv:1910.03910*, 2019.

Zahra Ghorbanali, Fatemeh Zare-Mirakabad, Mohammad Akbari, Nasim Salehi, and Ali Masoudi-Nejad. Drugrep-kg: Toward learning a unified latent space for drug repurposing using knowledge graphs. *Journal of Chemical Information and Modeling*, 63(9):2532–2545, 2023. doi: 10.1021/acs.jcim.2c01291.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. G-Retriever: Retrieval-augmented generation for textual graph understanding and question answering. In *Advances in Neural Information Processing Systems*, volume 37, pp. 132876–132907, 2024.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Jiho Kim, Yeonsu Kwon, Yohan Jo, and Edward Choi. KG-GPT: A general framework for reasoning on knowledge graphs using large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9410–9421, 2023. doi: 10.18653/v1/2023.findings-emnlp.631.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, pp. 28541–28564, 2023.

Mendeley Data. A benchmark image dataset for scabies detection 2.0. Mendeley Data, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaa El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A benchmark for automatic visual classification of clinical skin disease images. In *European Conference on Computer Vision*, pp. 206–222, 2016. doi: 10.1007/978-3-319-46466-4_13.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, et al. Towards generalist biomedical AI. *arXiv preprint arXiv:2307.14334*, 2023.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. Medical graph RAG: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28443–28467, 2025. doi: 10.18653/v1/2025.acl-long.1381.

Qinggang Zhang, Junnan Dong, Hao Chen, Daochen Zha, Zailiang Yu, and Xiao Huang. KnowGPT: Knowledge graph based prompting for large language models. In *Advances in Neural Information Processing Systems*, 2024.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. *arXiv preprint arXiv:2502.04413*, 2025.