

Knowledge-graph anchoring enables safer large language model reasoning for rare-tumor management

Yuan Chu

University Medical Center Mannheim, Medical Faculty Mannheim, University of Heidelberg

Chengyi Hua

University of Mannheim

Qing He

Xiangya Hospital of Central South University

Franka Menge

University Medical Center Mannheim, Medical Faculty Mannheim, University of Heidelberg

Christoph Reißfelder

University Medical Center Mannheim, Medical Faculty Mannheim, University of Heidelberg

Bernd Kasper

Mannheim Cancer Center (MCC), Mannheim University Medical Center, University of Heidelberg

Yimiao Xu

University of Queensland

Jens Jakob

University Medical Center Mannheim, Medical Faculty Mannheim, University of Heidelberg

Chengpeng Li

Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing); Peking University Cancer Hospital & Institute

Cui Yang

Cui.Yang@umm.de

University Medical Center Mannheim, Medical Faculty Mannheim, University of Heidelberg

Research Article

Keywords: Soft tissue sarcoma, Knowledge graph, Retrieval-augmented generation, Large language models, Clinical decision support

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10232500/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Large language models (LLMs) hold promises for supporting rare-tumor multidisciplinary care but remain prone to hallucination and lack interpretable evidence anchors in specialty domains. We constructed a soft tissue sarcoma (STS) knowledge graph (KG) from 9,594 curated PubMed abstracts, comprising 14,333 nodes across 16 entity types and 23,022 edges across 21 relation types, and integrated it as a retrieval-anchoring layer within a guideline-grounded retrieval-augmented generation (RAG) pipeline. Across 20 blinded clinical scenarios spanning five task categories, three sarcoma surgeons independently rated three architectures: Zero-shot LLM (no retrieval), Guideline-only RAG (guideline retrieval), and Hybrid KG-RAG (Guideline-only RAG augmented with a KG-anchoring layer), all using Llama 3 70B Instruct as the generator on accuracy, safety, Miller's Pyramid clinical competence, intent capture, and helpfulness (Likert 1–5). Hybrid KG-RAG achieved the highest median composite score (23.83 vs. 21.50 vs. 12.50 out of 25; Friedman $\chi^2(2) = 22.13$, $p < 0.001$; Kendall's $W = 0.55$) and was top-ranked on 15 of 20 questions. It significantly outperformed Zero-shot LLM on all five dimensions after Bonferroni correction (paired effect size $r = 0.57$ – 0.81). KG anchoring improved literature retrieval precision (mean reciprocal rank $0.774 \rightarrow 0.929$) and answer faithfulness (SummaCConv $\Delta = +0.216$, $p < 0.001$, $d_z = 1.18$). Three distinct failure modes are identified, delineating the current scope and limitations of KG-anchored retrieval. Specialty-specific KGs offer a viable, interpretable anchoring layer for rare-disease clinical decision support.

Introduction

Soft tissue sarcomas (STS) comprise more than 70 histological subtypes with markedly distinct biological behavior, prognostic implications, and treatment requirements. Despite their rarity, STS remain a substantial disease burden. The Global Burden of Disease 2021 estimated 96,201 incident cases and 50,203 deaths worldwide for soft tissue and extraosseous sarcomas, while SEER (Surveillance, Epidemiology, and End Results; <https://seer.cancer.gov/statistics-network/explorer/application.html>) population-based five-year relative survival was 83% for localized disease, 60% for regional disease, and only 17% once disease had metastasized[1–6]. Optimal management of STS requires integrating evidence-based guidelines, including those from European Society for Medical Oncology (ESMO)[7], National Comprehensive Cancer Network (NCCN)[8], and the German Guideline Program in Oncology (GGPO)[9], with up-to-date primary literature, subtype-specific molecular and pathological information, and individualized clinical contexts. This information landscape is too heterogeneous for any single specialist to track exhaustively. The multidisciplinary approach has become the standard of care for STS[7, 10–12], and the multidisciplinary team (MDT) meetings depend on the rapid synthesis of disparate evidence sources, often within the limited timeframe of a tumor board meeting. Thus, tools that can reliably and transparently integrate guidelines, literature, and structured biomedical knowledge into a clinically actionable form would therefore address an unmet need in rare-tumor care.

Large language models (LLMs) have demonstrated substantial potential for synthesizing medical knowledge and supporting clinical reasoning. Recent benchmarks such as MultiMedQA show near-expert performance on standardized medical question-answering (QA) tasks[13]. However, several limitations constrain their direct deployments in supporting the decision-making for rare tumors. General-purpose LLMs sometimes are prone to confident but factually incorrect outputs, particularly on specialty topics with sparse representation in general training corpora[14, 15]. Retrieval-augmented generation (RAG) conditions LLM outputs on retrieved evidence and has been proposed as partial mitigation, showing measurable benefits on common medical conditions and in the STS settings [16–18]. Yet in rare-tumor contexts, generic RAG pipelines may surface tangentially relevant literature, fail to enforce subtype-specific distinctions, or struggle to integrate the heterogeneous source types that drive real MDT decisions. Equally important, the evidence trail underlying a generated answer is rarely fully interpretable, limiting clinical trust and downstream auditability in safety-sensitive oncology contexts[19, 20].

Domain-specific knowledge graphs (KGs) offer a complementary representation in which clinical entities and their relationships are structured for explicit traversal, filtering, and querying[21, 22]. Recent work has explored integrating KGs with LLMs either as retrieval substrates or symbolic-reasoning layers[23–25]. Most published medical KG–LLM systems, however, have been built on broad biomedical ontologies, including Unified Medical Language System (UMLS)[26] or pan-oncologic databases, with evaluation focused on technical retrieval metrics or generic medical-QA benchmarks. Whether a disease-specific KG, integrated as a precision-oriented retrieval-anchoring layer and evaluated under blinded expert review using clinically meaningful criteria, adds value beyond these generic approaches remains an open question in rare-tumor settings such as STS.

In this study, we constructed an STS-specific KG from 9,594 curated PubMed abstracts (2003–2024), comprising 14,333 nodes and 23,022 relations. The KG was integrated as a retrieval-anchoring and evidence-routing layer within a guideline-grounded hybrid RAG pipeline using the GGPO S3 guidelines for STS. We compared three architectures sharing a Llama 3 70B Instruct generator: a Zero-shot LLM baseline without retrieval, a Guideline-only RAG condition with retrieval restricted to guideline text, and the Hybrid KG-RAG in which Guideline-only RAG is augmented by a KG-anchored retrieval layer. To test whether KG anchoring contributes meaningful clinical value beyond Guideline-only retrieval, we applied a three-tier evaluation strategy across 20 expert-designed clinical scenarios: (i) KG-ablation of literature-ranking precision; (ii) automated answer-context consistency scoring; and (iii) blinded assessment by three independent sarcoma surgeons across five dimensions (accuracy, safety, clinical competence, intent capture, and helpfulness). The Hybrid KG-RAG outperformed both Zero-shot LLM and Guideline-only RAG under expert review, with complementary gains in literature retrieval ranking and answer faithfulness. Our findings highlight that specialty-specific KG anchoring translates interpretability and grounding benefits into expert-judged clinical utility, while also revealing distinct failure modes that delineate the current scope of KG-anchored retrieval for rare-tumor decision support.

Results

KG overview

The constructed STS KG comprised 16 node types containing 14,333 nodes in total (Fig. S1a) and 21 relation types encompassing 23,022 edges (Fig. S1b), with provenance to 9,485 source PMIDs at the node level and 5,611 at the edge level. Among the 20 most frequently mentioned nodes (Fig. S1c), histological subtypes accounted for 14 entries and covered the principal recognized STS entities, with the remainder comprising treatment and biomarkers.

The 20 highest-degree nodes (Fig. S1d) clustered along the four dimensions that structure most clinical sarcoma decisions: treatment modalities, histological subtypes, biomarkers, and the principal primary and metastatic anatomical locations of STS. The concentration of cross-concept connectivity along precisely these axes indicates that the graph's most densely linked information layer is aligned with the dimensions that drive multidisciplinary sarcoma care. Relation-layer composition (typed and untuned relations) is summarized in Supplementary Note S1.

KG retrieval evaluation

To evaluate whether the KG component was able to surface question-relevant literature more effectively than retrieval without KG support, we compared the full Hybrid KG-RAG pipeline against an ablation in which KG retrieval was disabled. Of the 20 evaluated questions, 14 (70%) were eligible for KG anchoring analysis because they contained at least one KG-supported PMID target. Within this evaluable subset, KG anchoring increased the proportion of questions for which the top-ranked retrieved abstract was a relevant reference from 64.3% to 92.9%, and mean reciprocal rank improved from 0.774 to 0.929. Improvements were observed across all early-rank retrieval metrics but did not reach conventional statistical significance under the small evaluable sample (paired permutation $p = 0.125$; exact McNemar $p = 0.125$ for hit@1). No corresponding differences were observed at deeper ranks ($k = 3, 5, 10$), indicating that KG anchoring specifically promoted question-relevant references to the top of the ranked list rather than altering deeper-list recall. Full retrieval metrics, bootstrap intervals, and secondary diagnostics are reported in Supplementary Note S2.

Answer grounding evaluation

We next examined whether the retrieval changes were accompanied by stronger context-grounded generation under the fixed downstream model. Across the 20 questions, Hybrid KG-RAG achieved a higher SummaCConv (sentence-level grounding to the retrieved context, significantly improved) [27] than Guideline-only RAG (0.474 vs. 0.259; paired mean difference = 0.216; effect size $d_z = 1.18$; $p < 0.001$). A similar trend was observed on the secondary metric (AlignScore[28]: 0.431 vs. 0.340), though this difference was smaller and did not reach significance ($p = 0.201$). These findings indicate that KG-enhanced retrieval changes were accompanied by stronger context-grounded answer support under the fixed downstream generation model.

Expert evaluation: reliability and consensus

The three sarcoma surgeons independently scored 300 scoring units blindly (20 clinical queries $\times$ 3 response variants $\times$ 5 evaluation dimensions). Pre-adjudication reliability was uniformly high across all five evaluation dimensions, with point estimates of 0.980 for Accuracy, 0.984 for Safety, 0.937 for Miller's Pyramid clinical competence, 0.941 for Intent capture, and 0.951 for Helpfulness (Fig. S2). After consensus adjudication, ICC estimates increased to 0.994, 0.993, 0.959, 0.951, and 0.968, respectively (Supplementary Table S1). All subsequent analyses used the post-adjudication consensus score (mean across the three raters) as the unit of analysis.

Clinical performance across evaluation dimensions

Under blinded expert evaluation, the three architectures produced markedly different multidimensional performance profiles (Fig. 1). Hybrid KG-RAG achieved the highest median consensus score on every evaluated dimension, with Zero-shot LLM consistently lowest and Guideline-only RAG intermediate. Per-dimension medians, interquartile ranges, and pairwise comparisons are reported in Supplementary Table S2.

To synthesize multidimensional performance into a single utility measure, we summed the five Likert-scale scores within each scoring unit to yield a composite expert score with a theoretical range 5 to 25. Composite scores demonstrated substantial differences across conditions: Zero-shot LLM, 12.50 (IQR, 5.33–20.33); Guideline-only RAG, 21.50 (IQR, 19.92–22.58); and Hybrid KG-RAG, 23.83 (IQR, 21.83–24.42).

Examination of the score distribution provided a complementary view of clinical utility. Across the five dimensions, the proportion of responses receiving the highest score on the Likert scale of 5 rose progressively from the Zero-shot LLM (0–33%) to Guideline-only RAG (15–77%) and to Hybrid KG-RAG (47–87%), with the largest gains in the more demanding dimensions of clinical competence, intent capture and helpfulness (Fig. 1c). Pooling all five dimensions confirmed this monotonic trend (Fig. 1d), with top scores increasing from 14% to 43% to 67% across the three systems. Conversely, the proportion of clinically unacceptable responses (score 1) declined from 37–45% for Zero-shot LLM to 12–18% for Guideline-only RAG and to 0–10% under Hybrid KG-RAG.

Figure 1 Multidimensional expert evaluation of the three architectures across five clinical dimensions. (a) Median expert consensus score on accuracy, safety, clinical competence, intent capture, and helpfulness for Zero-shot LLM (gray), Guideline-only RAG (blue), and Hybrid KG-RAG (orange). Bars show the median and error bars the interquartile range. Brackets denote Bonferroni-corrected Wilcoxon signed-rank P values within each dimension (* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$). (b) Radar plot of the same per-dimension medians, illustrating the multidimensional performance profile of each architecture. (c) Distribution of individual Likert-scale ratings stratified by evaluation dimension for each architecture (each bar combines 20 questions $\times$ 3 raters; $n = 60$ per architecture per dimension). (d) Same data pooled across the five dimensions ($n = 300$ ratings per architecture). Full per-dimension and composite statistics are reported in Supplementary Table S2.

Question-level dominance, task and difficulty category

While dimension-wise pairwise comparisons captured average effects, they could not directly address whether the observed dimensional advantages translated into consistent question-level preferences. We therefore examined per-question rankings based on composite expert scores (Fig. 2a). Hybrid KG-RAG achieved the highest composite score on 15 of 20 questions (75%) and tied for first on one further question, compared with 3 of 20 (15%) for Guideline-

only RAG and 1 of 20 (5%) for Zero-shot LLM (Fig. 2b). The magnitude of the KG benefit was further modulated by task category (Fig. 2c; Supplementary Note S4; Supplementary Table S3). The benefit was largest where retrieval-free generation failed outright. On diagnostic-biopsy, systemic/advanced-disease and GIST-specific questions, Hybrid KG-RAG reached median composites of 23.7 (IQR 20.7–23.8), 23.3 (21.8–24.4) and 24.0 (23.1–24.8), respectively, against 5.3, 5.5 and 10.5 for the Zero-shot LLM. Across these same categories it also delivered a further median paired increment of roughly 2–5 points over Guideline-only RAG. For surgical-management questions, Guideline-only RAG already performed close to the ceiling (22.3, IQR 20.8–23.1), leaving only a small incremental gain for Hybrid KG-RAG (24.3, IQR 23.8–24.6). Radiotherapy was the sole exception: Hybrid KG-RAG conferred no benefit over Guideline-only RAG and a marginally lower median (19.7, IQR 13.0–22.0 vs. 21.3, IQR 21.2–23.2), a decrement attributable entirely to a single failure case (Q17; Supplementary Note S5).

Figure 2 Per-question dominance and clinical task-category profile of the composite expert score. (a) Composite score (sum of the five consensus Likert-scale ratings; range 5–25) for each of the 20 evaluation scenarios, ordered by Hybrid KG-RAG composite. Within each row, the horizontal connector links the three architectures (Zero-shot LLM, gray; Guideline-only RAG, blue; Hybrid KG-RAG, orange). (b) Number of questions on which each architecture obtained the highest outright composite score ($n = 20$). The Tie category corresponds to the one question on which Guideline-only RAG and Hybrid KG-RAG were tied at the top. (c) Median composite score by clinical task category: diagnostic biopsy ($n = 3$), surgical management ($n = 6$), systemic and advanced-disease therapy ($n = 4$), radiotherapy ($n = 3$), and GIST-specific decisions ($n = 4$). Bars show the median and error bars the interquartile range. Colors as in (a). Per-category descriptive statistics and pairwise increments are reported in Supplementary Table S3.

Hybrid KG-RAG's advantage over the Zero-shot LLM scaled inversely with question difficulty. It was largest where the baseline performed near-floor and narrowed toward the easiest questions, on which a modest median gain nonetheless persisted (Fig. 3a). Across questions, this gap correlated strongly and negatively with the Zero-shot LLM composite (Spearman $\rho = -0.82$, $p < 0.001$; Fig. 3b), a relationship partly reflecting a ceiling effect at high baseline scores. The Hybrid-over-Guideline-only RAG gap, by contrast, showed no such difficulty dependence ($\rho = +0.13$, $p = 0.57$). The knowledge-graph layer thus added a consistent, modest improvement across all difficulty levels, whereas its large advantage over the Zero-shot LLM concentrated where general-purpose LLM knowledge was weakest. The sole question on which Hybrid KG-RAG underperformed the Zero-shot LLM was a radiotherapy-timing case (Q17; Fig. 3b).

Figure 3 Hybrid KG-RAG advantage over Zero-shot LLM scales inversely with question difficulty. (a) Median composite expert score within each Zero-shot LLM composite tertile (hard, $n = 7$; medium, $n = 6$; easy, $n = 7$), grouped by architecture. Bars show the median and error bars the interquartile range. The median paired increment Δ median (Hybrid KG-RAG – Zero-shot LLM) is reported for each stratum. (b) Per-question performance gap between Hybrid KG-RAG and Zero-shot LLM (orange) and between Hybrid KG-RAG and Guideline-only RAG (blue), plotted against the Zero-shot LLM composite score (difficulty proxy); each point represents one of the 20 evaluation questions. Dashed lines indicate linear fits. Spearman rank correlations are $\rho = -0.82$ ($P < 0.001$) for the Hybrid vs Zero-shot LLM gap and $\rho = +0.13$ ($P = 0.57$) for the Hybrid vs Guideline-only RAG gap. Q17 is labeled as a knowledge-graph anchoring failure on a radiotherapy-timing question.

Discussion

This study constructed and evaluated an STS-specific KG integrated as a retrieval-anchoring layer within a guideline-grounded RAG pipeline. We evaluated the system in three independent ways: literature-retrieval precision, automated answer-context consistency, and a blinded multidimensional expert review. The Hybrid KG-RAG outperformed both the Zero-shot LLM and Guideline-only RAG on all three. It achieved the highest composite score in 15 of 20 clinical questions (75%) and demonstrated large effect sizes compared with the Zero-shot LLM on all five clinical dimensions. These findings support specialty-specific KG anchoring as a practical and interpretable approach for enhancing clinical decision support in rare-tumor care.

Our work contributes to a growing literature on integrating structured biomedical knowledge with LLMs for clinical question answering, while extending this paradigm to the underexplored domain of rare tumors. Several aspects of our approach distinguish it from previous studies. Most medical-LLM evaluations to date have focused on common-condition benchmarks such as MultiMedQA [13], with technical retrieval metrics serving as the primary endpoint. Existing biomedical KG–LLM systems have largely been built on broad ontologies and general-purpose graph resources: UMLS[26] and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)[29] for terminological grounding, Hetionet[24] and PrimeKG[25] for disease–drug–phenotype reasoning, and BIOS[30] or similar large-scale automatically constructed graphs for general biomedical question answering. More recent Graph-RAG paradigms have emphasized autonomous graph-traversal reasoning over LLM-mediated subgraph synthesis. In contrast, our design departs from these approaches in three respects. First, the specialty-specific KG is constructed from the disease-specific PubMed literature corpus of a rare tumor entity. Rather than functioning as an autonomous reasoner, it serves as a precision retrieval-anchoring layer. Second, the framework is evaluated through a three-tier strategy links retrieval ranking, automatic-context consistency, and blinded clinical scoring on the same question set. This combination addresses a recognized gap in rare-disease decision support, where broad biomedical graphs typically lack subtype-specific resolution and where general-purpose benchmarks fail to capture the multidisciplinary-team context in which decisions are actually made. The convergent improvement across all three evaluation levels also addresses a common methodological criticism of medical-AI work that relies on a single proxy metric. Finally, by linking each KG retrieval to citation provenance in the generated answer, the system provides an interpretability property that clinical use demands.

The mechanistic role of the KG in this pipeline merits explicit comment because it differs from the framings most adopted in the literature. The KG did not function as an autonomous reasoner or as a substitute for retrieval over guideline text but as a retrieval-anchoring and evidence-routing layer that constrained which primary literature was admitted into the generation context. The score distributions under each condition illustrate the effect clearly. Under Zero-shot LLM, a well-separated cluster of questions sat near the floor of the scale, clearly set apart by a gap, while the remainder spread across the middle-to-upper range. This reflects a population of clinical scenarios for which generalist LLM knowledge was either adequate or essentially absent. The principal effect of retrieval grounding, and of KG anchoring in particular, was to rescue the lower-scoring cluster rather than to lift those the model already handled competently. Three observations are consistent with this interpretation. First, SummaConv[27] showed gains while AlignScore[28], an NLI-style holistic

semantic alignment metric, did not. This dissociation suggests that KG integration primarily operated through improved retrieval and grounding rather than through changes in the generator's semantic reasoning. Second, the incremental gain of Hybrid KG-RAG over Guideline-only RAG was largely flat across question difficulty, in contrast to the strongly difficulty-dependent improvement over the Zero-shot LLM. The KG thus added a constant lift across the full range of question difficulty. Third, the clinical-category subgroup analysis showed the largest KG benefit on knowledge intensive questions concerning diagnostic biopsy and systemic/advanced disease therapy, where Guideline-only RAG provided partial but insufficient grounding, and the smallest benefit appeared on questions in which Zero-shot LLM already approached ceiling. These patterns point to a role for the specialty KG. It functions as a high precision retrieval layer that complements guideline text rather than replacing it, and it helps most on knowledge-intensive questions whose answers depend on subtype-specific literature that a generalist LLM may not reliably retrieve or prioritize.

Beyond aggregating accuracy, the expert scoring carries concrete implications for clinical utility. Across all five dimensions, Hybrid KG-RAG improved most over Zero-shot LLM on clinical competence (Miller's Pyramid), intent capture, and helpfulness, which is particularly relevant when an MDT needs a recommendation that is both anchored in specialty evidence and actionable in practice. The shape of the score distribution shifted in two directions. The proportion of responses rated clinically unacceptable fell from 37–45% with the Zero-shot LLM to 0–10% with Hybrid KG-RAG, while the proportion of top-rated answers rose from 0–33% to 47–87%. This combination matters for rare-tumor decision support, where individual specialists may not have all current literature at hand and an erroneous recommendation may carry potential clinical consequences. In practice, Hybrid KG-RAG made the typical answer better while substantially reducing the frequencies of responses that were potentially unsafe or clinically unusable. A further advantage is traceability. Each retrieved abstract remains linked through the KG to its original source, enabling clinicians to verify the evidentiary basis of recommendations rather than relying solely on model outputs.

Three failure cases mark where the current pipeline breaks down. On Q17, Hybrid KG-RAG anchored its answer on a paper that was topically close but did not address the question. Gating answers on retrieval confidence and matching each question more strictly to candidate publications before an anchor is fixed would prevent this anchoring miscalibration. On Q2, retrieval returned technique-level evidence for what was a question of indication. Explicit prerequisite-checking or constraint-aware re-ranking would close this gap. On Q9 the error ran the other way: Zero-shot LLM produced a clinically unsafe affirmative answer that retrieval grounding correctly intercepted. This was the clearest case for guideline-grounded generation in safety critical setting.

Our study has several clinical and technical limitations. Clinically, three sarcoma surgeons with prior practice at different centers contributed cross-institutional perspective, yet they remain a small expert panel. The 20-question benchmark spans representative decision points but is a modest sample for fully characterizing LLM behavior across sarcoma practice. The rubric was anchored on a single national guideline, the German GGPO S3, which has known coverage differences from NCCN and ESMO. Technically, the KG carries extraction noise. The relation-validation agent retained 62% of candidate triples and the entity stage retained 72% of candidate mentions, reflecting a precision-over-recall design that has not been benchmarked against alternative pipelines. The system itself was instantiated with a single open-weights generator (Llama 3 70B Instruct). The KG anchoring retrieval probe was further confined to 14 questions with at least one silver-target PMID; bootstrap 95% confidence intervals lay entirely above zero, but permutation tests did not reach significance ($P = 0.125$), and a larger benchmark is needed to settle the effect at the population level. Finally, we measured expert-rated answer quality, not patient outcomes; impact on diagnostic concordance, decision time, and downstream care requires a prospective integration study.

Taken together, a specialty-specific KG used as a retrieval-anchoring layer rather than an autonomous reasoner can meaningfully improve guideline-grounded RAG for soft-tissue sarcoma clinical questions. Convergent gains across literature retrieval, answer-context grounding, and blinded expert review on the same question set indicate that the benefit of Hybrid KG-RAG is not an artifact of any single proxy metric. Each generated answer carries citation provenance to its source literature, providing the traceability that clinical use demands. Hybrid KG-RAG thus offers a credible, interpretable foundation for AI-assisted decision support in rare-tumor settings.

Methods

Ethical considerations

All clinically relevant information used in this study does not contain any real patient data. Consequently, approval from an institutional ethics committee is not required.

Data sources, KG construction and retrieval pipeline

This study used two primary textual data sources, the English version of the GGPO Soft Tissue Sarcoma Guideline, version 1.1[9] and a corpus of PubMed (<https://pubmed.ncbi.nlm.nih.gov/>) abstracts related to STS, and one derived structured resource, a KG constructed from the abstract corpus as a structured anchoring and retrieval-guidance resource for abstract evidence selection.

Data sources

The guideline was converted into a structured, machine-readable corpus preserving its internal structure, including recommendations, background sections, tables, quality indicators, section labels, and citation links. This structure allowed the system to retrieve evidence at the level of specific guideline sections rather than treating the guideline as an undifferentiated document.

STS-related PubMed abstracts were collected using subtype-specific keyword queries restricted to English-language publications from 2003 to 2024; the full search strategy and the complete list of subtype synonyms are reported in Supplementary Note S6 and Supplementary Table S5. Abstracts cited in the guideline reference list were added regardless of keyword filtering. After reserving 400 abstracts for model development, the final corpus comprised 9,594 abstracts.

KG construction

The KG was derived from the abstract corpus to capture structured relationships between clinically relevant entities, including diseases, interventions, biomarkers, and anatomical sites. Candidate entities were extracted using a biomedical Named Entity Recognition (NER) model fine-tuned on 400 manually annotated STS abstracts; 72% of model-predicted entities were confirmed by human reviews and retained for graph construction. Candidate relations between entity pairs were generated by pairing validated entity mentions co-occurring within a two-sentence sliding window. Each pair was then evaluated by an LLM-based relation validation agent. The full procedure is given in Supplementary Note S7. 62% of candidate relations were retained as graph edges. The final KG has two layers: a semantically specified relation layer that anchors retrieval guidance in order to prioritize precision and interpretability, and an exploratory association layer retained for future refinement.

Retrieval pipeline

For each clinical query, evidence was retrieved through a four-stage pipeline (guideline retrieval, citation extraction, KG retrieval, and abstract retrieval) coordinated by a Python orchestrator that ran the stages in sequence and assembled the final context (Fig. 4). Rather than querying all resources independently, later retrieval stages were conditioned on evidence surfaced earlier in the pipeline. The guideline served as the primary evidence source, with retrieval guided by the document's content and section structure to surface the most relevant recommendations and supporting content. Citation markers and section-linked references from top-ranked guideline passages were extracted and semantically ranked against the query. The existing corresponding PMIDs were then used to guide both KG retrieval and abstract retrieval. KG evidence was not used for open-ended graph reasoning; instead, it provided anchored nodes and KG-associated PMIDs. Candidate nodes were identified through direct question-to-node matching and PMID-guided retrieval using literature references from the citation stage, then organized into clinical anchor categories such as disease or intervention. PubMed abstract retrieval prioritized direct citation-stage PMIDs. KG-associated PMIDs were used as optional additions only when their best abstract chunks exceeded the best direct-citation abstract, or as fallback when no citation-linked abstracts were available. Corpus-wide retrieval was used only as the final fallback. All retrieved evidence units were normalized, deduplicated, ranked, and assembled into a structured prompt context with guideline recommendations weighted most heavily. KG-associated PMIDs were carried forward as literature anchors: abstracts whose PMIDs overlapped with KG anchor-supported nodes or KG evidence edges received additional ranking weight. Guideline and abstract evidence formed the textual context, while KG anchors were retained as an auxiliary structured layer. For the Hybrid KG-RAG condition, a second-pass enrichment step was applied after initial answer generation to append only retrieved abstract and KG-derived evidence. Full technical details of each pipeline component are provided in the Supplementary Notes S6 and S7.

Figure 4 Guideline-grounded Hybrid KG-RAG pipeline. The dashed boundary indicates modules coordinated by a Python orchestrator. Full details in Methods and Supplementary Notes S6 and S7.

Prompting and answer generation

We used a set of 20 clinical queries of STS management developed in a previous study[18]; the full list of queries and their category assignments is provided in Supplementary Table S4. To ensure models' capacity for semantic understanding, the scenarios were intentionally phrased in language distinct from that used in the GGPO S3 guideline.

Answer generation was performed using a locally deployed Llama 3 70B Instruct model served through Ollama 0.24 (<https://ollama.com/>). To keep generation variability low while allowing limited flexibility for synthesizing guideline, abstract, and KG-derived evidence, decoding parameters were fixed across experimental conditions, with temperature set to 0.2 and top-p set to 0.9. Three prompting conditions were defined: a Zero-shot LLM baseline without retrieved evidence, a Guideline-only RAG condition using only retrieved guideline context, and a Hybrid KG-RAG using the final context assembled from guideline excerpts, abstract evidence, and auxiliary KG-derived cues.

Prompt templates were designed to enforce concise, evidence-grounded answers in continuous prose. In the retrieval-based settings, the model was instructed to avoid unsupported claims, not to introduce external medical knowledge, and to preserve explicit restrictions, contraindications, and subtype-specific distinctions when present in the evidence. Output instructions further suppressed explicit source-referencing language, citations, and intermediate reasoning traces.

In the Guideline-only RAG condition, the model received only guideline-derived evidence and was instructed to base every conclusion exclusively on the retrieved guideline excerpts. In the Hybrid KG-RAG condition, a second-pass evidence enrichment step was applied after initial answer generation. This step considered retrieved abstract evidence together with auxiliary KG-derived cues, explicitly excluding guideline evidence, and was used solely to append previously omitted factual details supported by the retrieved literature context. The enrichment pass was restricted to supportive additions, such as numerical study findings, intervention–outcome relations, and subtype-specific distinctions, and was instructed not to restate or alter the main conclusion generated in the first pass.

Evaluation methodology

System performance and reliability were evaluated through a two-phase framework, encompassing both automated phase and expert clinical assessment. The automated phase addressed two distinct questions: whether generated answers were grounded in the retrieved evidence context, and whether KG anchoring improved the promotion of question-relevant literature within the retrieval pipeline. The clinical phase involved blinded, multidimensional scoring of model responses by sarcoma surgeons across 20 STS clinical scenarios.

Automated system evaluation

The first automated evaluation assessed whether generated answers were supported by the evidence context assembled by the retrieval pipeline (Fig. 5b). For each question-answer pair, the generated answer was compared against the corresponding cleaned evaluation context using two automatic metrics, AlignScore and SummaCConv. AlignScore[28] estimates factual consistency through a unified information-alignment function between a context and a claim, whereas SummaCConv[27] was originally developed for summary inconsistency detection by evaluating whether generated text is supported by a source document through sentence-level NLI aggregation. These metrics were selected because they do not require manual reference annotations and are suited to evaluating whether model outputs remain grounded in the retrieved evidence rather than merely matching a predefined gold answer. SummaCConv was treated as the primary automatic answer-context consistency metric, while AlignScore was included as a complementary alignment-based measure. Full implementation details and paired statistical procedures are reported in Supplementary Note S3.

The second automated evaluation examined whether the KG component fulfilled its intended retrieval function: promoting question-relevant literature into the evidence pool surfaced to the model. To isolate this mechanism, the full Hybrid KG-RAG pipeline was compared against an ablation run in which KG retrieval was disabled while the remaining pipeline components, guideline retrieval, citation linking, abstract retrieval, and context construction, were kept unchanged. For each question, we assessed whether abstracts linked to KG-identified entities ranked more prominently in the retrieved evidence under the Hybrid KG-RAG condition than under the ablation, using standard information retrieval metrics including recall, hit rate, and reciprocal rank at multiple cutoffs. Full retrieval-metric definitions and statistical procedures are reported in Supplementary Note S2.

Figure 5 Two automated probes of the Hybrid KG-RAG system. (a) KG anchoring mechanism evaluation. (b) Answer-context consistency evaluation. Metric definitions and statistical procedures are reported in Supplementary Notes S2 and S3.

Clinical scenario and human expert evaluation

The same 20-scenario dataset was previously used to evaluate GPT-4o with a guideline-grounded retrieval-augmented generation pipeline by our group[18].

Evaluation was conducted by three sarcoma surgeons, who are familiar with the GGPO S3 guideline. Each rater independently evaluated every model response (Zero-shot LLM, Guideline-only RAG, Hybrid KG-RAG), blinded to architecture identity. All clinical evaluations were benchmarked against the GGPO S3 guideline.

We incorporated five evaluation criteria: accuracy, safety, the Miller Pyramid of Clinical Competence, intent capture, and helpfulness. Each response was scored on a 5-point Likert scale across five dimensions, with 1 indicating clinically unacceptable performance and 5 indicating optimal performance: Because all questions were derived from complex clinical scenarios, we adopted the Miller Pyramid evaluation framework, which is widely used to assess clinical competence in medical education and trainee proficiency[31], and added a “Failure” category to the original four levels of “Knows,” “Knows how,” “Shows how,” and “Does” (Table 1). This framework evaluates not only the accuracy and safety of knowledge retrieval but also the clinical reasoning and competence demonstrated in model responses. To more thoroughly assess model performance in clinical Q&A scenarios after knowledge graph (KG) integration, we further adopted the MultiMedQA framework [13] to evaluate two additional dimensions. The first, Capture of User Intent, measures the extent to which a response addresses the specific underlying clinical purpose of the query. The second, Helpfulness, measures the operational utility of the answer in enabling the clinician to draw a clear conclusion or identify the necessary next steps in an MDT discussion (Table 1).

Where the range across the three raters exceeded one point, a consensus-session was convened, in which the three raters jointly reviewed the response and agreed on a harmonized score. The post-adjudication consensus mean score served as the unit of analysis for all subsequent statistical comparisons.

Table 1
Scoring rubric for expert evaluation of model responses across five clinical dimensions.

Miller's Pyramid of Clinical Competence		
Score	Level	Clinical assessment criterion
1	Failure	Provides incorrect medical facts or uses flawed logic that would compromise patient safety in a clinical setting.
2	Knows (Knowledge)	Provides basic facts or quotes guidelines but fails to apply them to the specific patient.
3	Knows how (Competence)	Provides a rational treatment strategy for the case but lacks technical details for execution.
4	Shows how (Performance)	Provides the correct strategy and explicitly demonstrates the evidence-based logic used to reach it.
5	Does (Action)	Provides a definitive, granular action plan ready for immediate independent implementation.
Capture of User Intent		
<i>How accurately the model understood the clinical constraints and the specific query.</i>		
1	Irrelevant	Entirely fails to understand the question or provides off-topic information.
2	Poorly captured	Misinterprets the primary clinical focus despite mentioning related terms.
3	Partially captured	Understands the general intent but misses key constraints such as tumor size or mutation status.
4	Mostly captured	Addresses the main intent and all primary questions with only negligible omissions.
5	Perfectly captured	Addresses all parts of the user's intent perfectly, integrating every provided clinical variable.
Helpfulness		
<i>How helpful is the answer to the user (for example, does it enable them to draw a conclusion or help clarify next steps)?</i>		
1	Not helpful	Useless, misleading, or contains content that does not contribute to solving the clinical problem.
2	Slightly helpful	Provides generic information that is too vague to offer practical value to a medical professional.
3	Moderately helpful	Provides correct facts but is insufficient for the user to draw a final conclusion without further research.
4	Very helpful	Significantly clarifies the clinical situation and provides high-quality support for decision-making.
5	Extremely helpful	Provides total clarity, enabling the user to draw a definitive conclusion and establish clear next steps immediately.

Statistical analysis

Expert-rating comparative statistics and their figures were generated in R 4.4.2 (base-R `friedman.test`, `wilcox.test` and `cor.test`; `ggplot2` 3.5.1 with `patchwork`, <https://www.r-project.org/>). Inter-rater reliability, retrieval and answer-faithfulness analyses were performed in Python 3.9 (`scipy` 1.11; `pingouin` 0.5; `matplotlib` 3.8; `seaborn` 0.13, <https://www.python.org/downloads/release/python-390/>); full procedures are given in Supplementary Notes S2 and S3. All tests were two-sided and $P < 0.05$ was considered significant after correction for multiple comparisons where applicable. Inter-rater reliability was quantified for each dimension using the intraclass correlation coefficient ICC(3,k) [32, 33], with bootstrap 95% confidence intervals (Fig. S2).

Non-parametric Friedman tests compared Zero-shot LLM, Guideline-only RAG, and Hybrid KG-RAG on each dimension and on the composite score (per-question sum of the five ratings; range 5 to 25). The proportion of questions won outright by each architecture was summarized as a win count, and the dependence of the Hybrid KG-RAG advantage over Zero-shot LLM on question difficulty (Zero-shot composite tertiles) was assessed by Spearman's rank correlation.

Retrieval-side and answer-context-consistency analyses, comparing the pipeline against KG-ablated pipeline against Guideline-only RAG conditions at the matched-question level, are described in full in Supplementary Notes S2 and S3.

Declarations

Author contributions

Y.C., C.Y., C.P.L. and C.Y.H. conceived the study and developed the methodology. Y.C. and C.Y.H. performed the data preparation and model construction. C.Y.H. carried out the initial evaluation of the knowledge graph and, as a non-clinician, blinded the questions and model-generated answers and compiled the scoring results. Y.C., C.Y. and C.P.L. devised the clinical evaluation criteria. Y.C., C.Y. and C.P.L. independently performed the blinded scoring of the model responses. Y.C., Q.H. and Y.M.X. analyzed the clinical-evaluation data and prepared the figures. F.M., C.R., B.K. and J.J. provided guidance. C.P.L. and C.Y. provided overall supervision and critically revised the manuscript. Y.C. wrote the original draft. All authors read and approved the final manuscript.

Acknowledgements

Funding

This project was supported by funding from the National Natural Science Foundation of China (Grant No. 91959120).

Declarations

Clinical trial number: not applicable.

Ethics and Consent to Participate declarations: not applicable.

Competing interests: The authors declare no competing interests

Data availability

All source materials are publicly accessible. The 9,594 PubMed abstracts (2003–2024) used were retrieved using the search strategy, date filters, and inclusion criteria detailed in Supplementary Note S6. Raw records remain retrievable from PubMed (<https://pubmed.ncbi.nlm.nih.gov/>). The German S3 guideline on adult soft-tissue sarcoma (GGPO) is publicly available at <https://www.leitlinienprogramm-onkologie.de/leitlinien/adulte-weichgewebesarkome>. The 20 clinical scenarios with the corresponding 60 model-generated answers across the three architectures, and the de-identified blinded expert rating sheet have been deposited in Zenodo (<https://doi.org/10.5281/zenodo.20415353>) under an embargo until publication. During peer review the archive is shared with editors and reviewers via a restricted-access link: <https://zenodo.org/records/20415353?preview=1&token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjlyYzI5ZjY0LWQ0NGQtNDg2OS1iMDFkLTU4ZDFiZTYyMjBINCiSlmRhdGEiOnt9LCJyYW5kb20iOiJhZjZjZzYTxs1fSPGgS0Ab7vrfktMDzEmjSE0HmcV9cJC2d2FCQqXuzWJMmpFFS4A>. The archive will be released publicly upon acceptance.

Code availability

The custom scripts used to construct the knowledge graph, run the retrieval pipeline and automated answer consistency evaluation have been deposited together with the data in the same Zenodo archive (<https://doi.org/10.5281/zenodo.20415353>). During peer review the archive is shared with editors and reviewers via a restricted-access link provided in the Data availability statement. The code will be available publicly under the above DOI upon acceptance. Large language model inference was performed using Llama 3 70B Instruct (Meta), served locally with Ollama v0.24.

References

1. Sbaraglia, M., E. Bellan, and A.P. Dei Tos, *The 2020 WHO Classification of Soft Tissue Tumours: news and perspectives*. Pathologica, 2021. 113(2): p. 70–84.
2. Stiller, C.A., et al., *Descriptive epidemiology of sarcomas in Europe: report from the RARECARE project*. Eur J Cancer, 2013. 49(3): p. 684–95.
3. Gatta, G., et al., *Burden and centralised treatment in Europe of rare tumours: results of RARECAREnet-a population-based study*. Lancet Oncol, 2017. 18(8): p. 1022–1039.
4. Chen, C., et al., *Global, regional, and national burden of soft tissue and extraosseous sarcomas from 1990 to 2021*. Prev Med Rep, 2024. 47: p. 102903.
5. Gamboa, A.C., A. Gronchi, and K. Cardona, *Soft-tissue sarcoma in adults: An update on the current state of histotype-specific management in an era of personalized medicine*. CA Cancer J Clin, 2020. 70(3): p. 200–229.
6. Society, A.C. *Survival Rates for Soft Tissue Sarcomas*. 2025; Available from: <https://www.cancer.org/cancer/types/soft-tissue-sarcoma/detection-diagnosis-staging/survival-rates.html>.
7. Gronchi, A., et al., *Soft tissue and visceral sarcomas: ESMO-EURACAN-GENTURIS Clinical Practice Guidelines for diagnosis, treatment and follow-up(®)*. Ann Oncol, 2021. 32(11): p. 1348–1365.
8. Guidelines, N. *NCCN Clinical Practice Guidelines in Oncology Soft Tissue Sarcoma*. 2026; Available from: <https://www.nccn.org/guidelines/guidelines-detail?category=1&id=1464>.
9. Wilson, P.A.H.P.I.J.P.C.M.C.S.R., *Evidence-based Guidelines: Soft Tissue Sarcoma*. 2022, German Guideline Program in Oncology.
10. Blay, J.Y., et al., *Improved survival using specialized multidisciplinary board in sarcoma patients*. Ann Oncol, 2017. 28(11): p. 2852–2859.
11. Blay, J.Y., et al., *Surgery in reference centers improves survival of sarcoma patients: a nationwide study*. Ann Oncol, 2019. 30(7): p. 1143–1153.
12. Blay, J.Y., et al., *Improved nationwide survival of sarcoma patients with a network of reference centers*. Ann Oncol, 2024. 35(4): p. 351–363.
13. Singhal, K., et al., *Large language models encode clinical knowledge*. Nature, 2023. 620(7972): p. 172–180.
14. Roustan, D. and F. Bastardot, *The Clinicians' Guide to Large Language Models: A General Perspective With a Focus on Hallucinations*. Interact J Med Res, 2025. 14: p. e59823.
15. Azamfirei, R., S.R. Kudchadkar, and J. Fackler, *Large language models and the perils of their hallucinations*. Crit Care, 2023. 27(1): p. 120.
16. Xiong, G., et al. *Benchmarking retrieval-augmented generation for medicine*. in *Findings of the Association for Computational Linguistics: ACL 2024*. 2024.
17. Zakka, C., et al., *Almanac - Retrieval-Augmented Language Models for Clinical Medicine*. NEJM AI, 2024. 1(2).
18. Li, C.P., et al., *Improving Accuracy and Source Transparency in Responses to Soft Tissue Sarcoma Queries Using GPT-4o Enhanced with German Evidence-Based Guidelines*. Oncol Res Treat, 2025. 48(6): p. 351–359.
19. Topol, E.J., *High-performance medicine: the convergence of human and artificial intelligence*. Nat Med, 2019. 25(1): p. 44–56.
20. Ennab, M. and H. McHeick, *Designing an Interpretability-Based Model to Explain the Artificial Intelligence Algorithms in Healthcare*. Diagnostics (Basel), 2022. 12(7).
21. Ying, H., et al., *CoRTEX: contrastive learning for representing terms via explanations with applications on constructing biomedical knowledge graphs*. J Am Med Inform Assoc, 2024. 31(9): p. 1912–1920.

22. Fei, Y., et al., *Application of knowledge graphs in rare disease research*. Front Public Health, 2026. 14: p. 1757612.
23. Edge, D., et al., *From local to global: A graph rag approach to query-focused summarization*. arXiv preprint arXiv:2404.16130, 2024.
24. Himmelstein, D.S., et al., *Systematic integration of biomedical knowledge prioritizes drugs for repurposing*. Elife, 2017. 6.
25. Chandak, P., K. Huang, and M. Zitnik, *Building a knowledge graph to enable precision medicine*. Sci Data, 2023. 10(1): p. 67.
26. Bodenreider, O., *The Unified Medical Language System (UMLS): integrating biomedical terminology*. Nucleic Acids Res, 2004. 32(Database issue): p. D267-70.
27. Laban, P., et al., *SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization*. Transactions of the Association for Computational Linguistics, 2022. 10: p. 163–177.
28. Li, Y.Z.Y.Y.R. and Z. Hu, *ALIGNSCORE: Evaluating Factual Consistency with A Unified Alignment Function*.
29. Gaudet-Blavignac, C., et al., *Use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for Processing Free Text in Health Care: Systematic Scoping Review*. J Med Internet Res, 2021. 23(1): p. e24594.
30. Yu, S., et al., *Bios: An algorithmically generated biomedical knowledge graph*. arXiv preprint arXiv:2203.09975, 2022.
31. Miller, G.E., *The assessment of clinical skills/competence/performance*. Acad Med, 1990. 65(9 Suppl): p. S63-7.
32. Shrout, P.E. and J.L. Fleiss, *Intraclass correlations: uses in assessing rater reliability*. Psychological bulletin, 1979. 86(2): p. 420.
33. Koo, T.K. and M.Y. Li, *A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research*. J Chiropr Med, 2016. 15(2): p. 155–63.

Figures

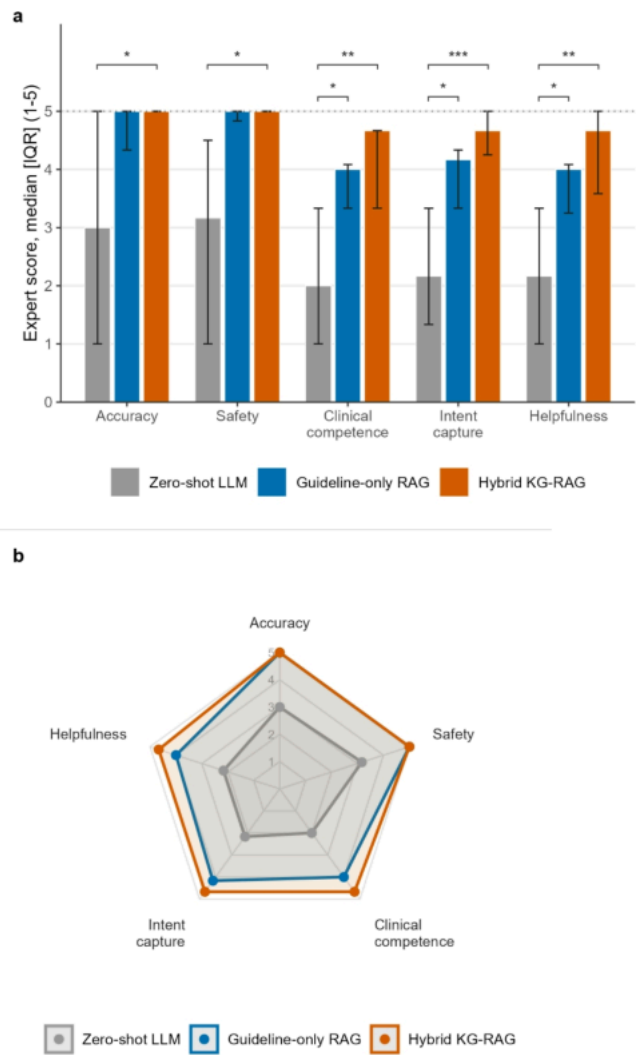


Figure 1

Multidimensional expert evaluation of the three architectures across five clinical dimensions. (a) Median expert consensus score on accuracy, safety, clinical competence, intent capture, and helpfulness for Zero-shot LLM (gray), Guideline-only RAG (blue), and Hybrid KG-RAG (orange). Bars show the median and error bars the interquartile range. Brackets denote Bonferroni-corrected Wilcoxon signed-rank P values within each dimension (*P < 0.05; **P < 0.01; ***P < 0.001). (b) Radar plot of the same per-dimension medians, illustrating the multidimensional performance profile of each architecture. (c) Distribution of

individual Likert-scale ratings stratified by evaluation dimension for each architecture (each bar combines 20 questions × 3 raters; n = 60 per architecture per dimension). (d) Same data pooled across the five dimensions (n = 300 ratings per architecture). Full per-dimension and composite statistics are reported in Supplementary Table S2.

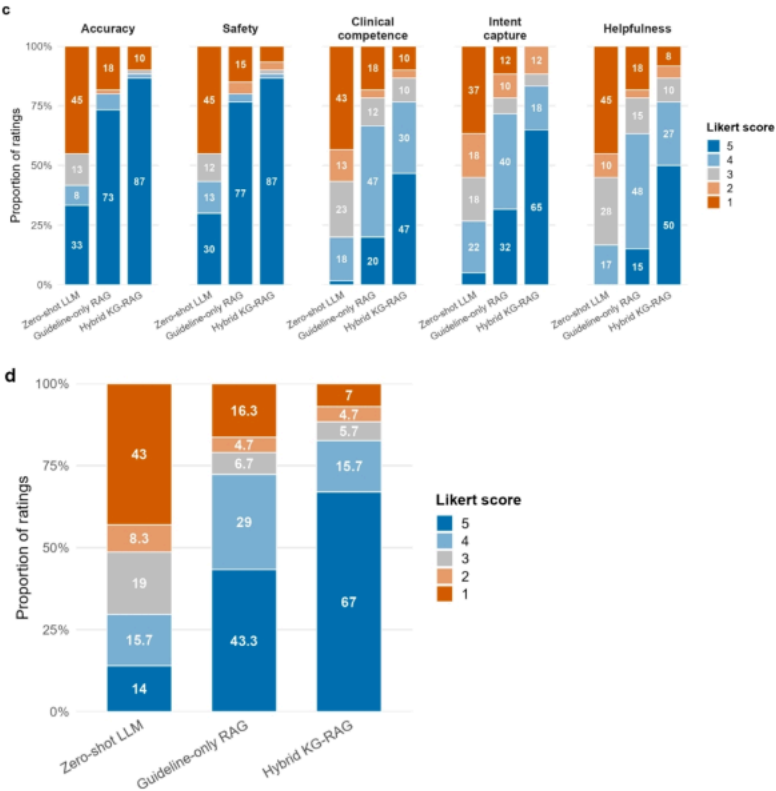


Figure 2

Per-question dominance and clinical task-category profile of the composite expert score. (a) Composite score (sum of the five consensus Likert-scale ratings; range 5–25) for each of the 20 evaluation scenarios, ordered by Hybrid KG-RAG composite. Within each row, the horizontal connector links the three architectures (Zero-shot LLM, gray; Guideline-only RAG, blue; Hybrid KG-RAG, orange). (b) Number of questions on which each architecture obtained the highest outright composite score ($n = 20$). The Tie category corresponds to the one question on which Guideline-only RAG and Hybrid KG-RAG were tied at the top. (c) Median composite score by clinical task category: diagnostic biopsy ($n = 3$), surgical management ($n = 6$), systemic and advanced-disease therapy ($n = 4$), radiotherapy ($n = 3$), and GIST-specific decisions ($n = 4$). Bars show the median and error bars the interquartile range. Colors as in (a). Per-category descriptive statistics and pairwise increments are reported in Supplementary Table S3.

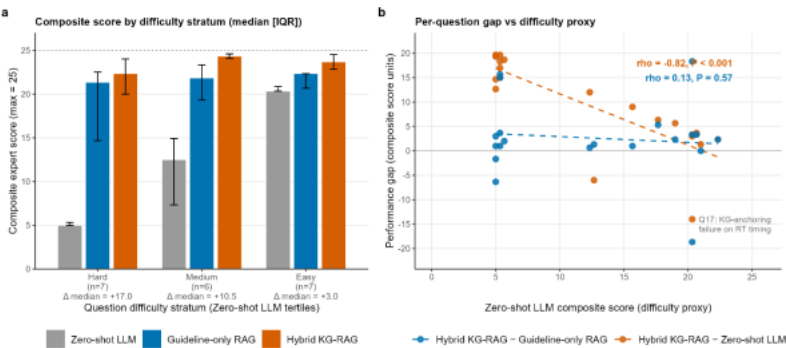


Figure 3

Hybrid KG-RAG advantage over Zero-shot LLM scales inversely with question difficulty. (a) Median composite expert score within each Zero-shot LLM composite tertile (hard, $n = 7$; medium, $n = 6$; easy, $n = 7$), grouped by architecture. Bars show the median and error bars the interquartile range. The median paired increment Δ median (Hybrid KG-RAG – Zero-shot LLM) is reported for each stratum. (b) Per-question performance gap between Hybrid KG-RAG and Zero-shot LLM (orange) and between Hybrid KG-RAG and Guideline-only RAG (blue), plotted against the Zero-shot LLM composite score (difficulty proxy); each point represents one of the 20 evaluation questions. Dashed lines indicate linear fits. Spearman rank correlations are $\rho = -0.82$ ($P < 0.001$) for the Hybrid vs Zero-shot LLM gap and $\rho = +0.13$ ($P = 0.57$) for the Hybrid vs Guideline-only RAG gap. Q17 is labeled as a knowledge-graph anchoring failure on a radiotherapy-timing question.

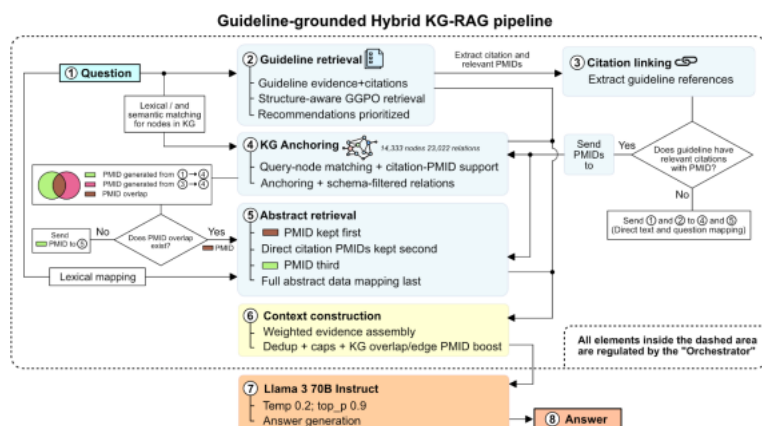


Figure 4

Guideline-grounded Hybrid KG-RAG pipeline. The dashed boundary indicates modules coordinated by a Python orchestrator. Full details in Methods and Supplementary Notes S6 and S7.

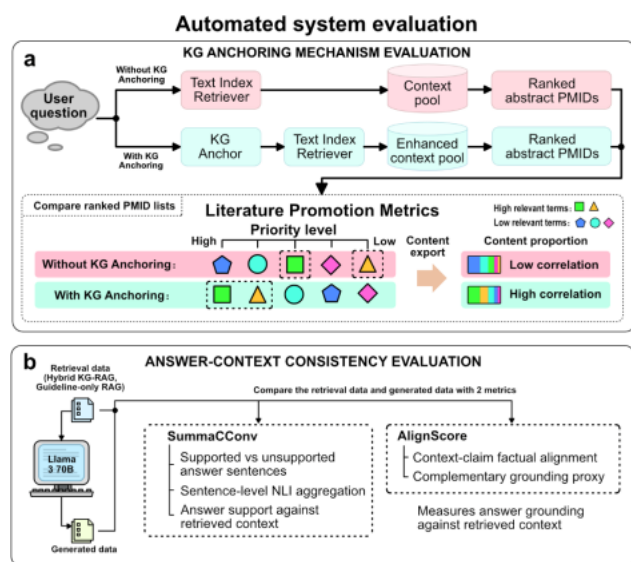


Figure 5

Two automated probes of the Hybrid KG-RAG system. (a) KG anchoring mechanism evaluation. (b) Answer-context consistency evaluation. Metric definitions and statistical procedures are reported in Supplementary Notes S2 and S3.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [STSKGSupplementary.docx](#)