OrthoGraphRAG: Enhancing Clinical Decision Making with Multi-Level Knowledge Graphs

Venkatesh Tata [*1] **Zohra Bouchamaoui** [*1] **Nagavaishnavi V Bhaskara** [2]

Abstract

Large Language Models (LLMs) face accuracy and complex reasoning challenges in specialized medical domains like orthopedics. We introduce OrthoGraphRAG, a multi-level Graph Retrieval-Augmented Generation (GraphRAG) framework, to address these issues. OrthoGraphRAG constructs a novel multi-level knowledge graph linking private clinical knowledge with public UMLS data, building on recent medical GraphRAG advancements. The framework retrieves query-entity-based subgraphs, augments them with clinical note text, allowing an LLM to synthesize informed responses from combined graph and textual evidence. Evaluated on real-world orthopedic clinic letters with diverse query complexities, OrthoGraphRAG demonstrated effectiveness, particularly in contextual reasoning integrating private patient data with broader medical knowledge. This multi-level GraphRAG approach offers a promising path to safer, more capable, and contextually aware LLMs for specialized clinical applications. Our code is released at: https://github.com/venkateshtata/OrthoGraphRAG

1. Introduction

Large Language Models (LLMs), despite their natural language prowess, face reasoning limitations in specialized, information-dense domains like orthopedics due to restricted context windows and difficulty with extensive knowledge bases (Chen et al., 2024; Gao et al., 2023). While Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) and its counterpart, GraphRAG, provide external knowledge, broad medical applications can be overly complex. We introduce **OrthoGraphRAG**, specializing these techniques for orthopedic clinical decision-making. OrthoGraphRAG balances reduced query complexity with sophisticated knowledge representation, vital for orthopedics which demands factual accuracy and nuanced multi-source reasoning (Wu et al., 2024b).

OrthoGraphRAG's core is a multi-level GraphRAG framework with a novel multi-layer knowledge graph (Figure 1). This graph integrates private clinical knowledge from hospital letters with public UMLS data (Bodenreider, 2004) via cosine similarity-based entity linking (Wu et al., 2024b). Evaluation on real-world orthopedic clinic letters across diverse question categories (Information Retrieval, Explanatory Reasoning, Contextual Reasoning) showed high success rates, particularly excelling in complex contextual reasoning that synergistically integrates private patient data with broader medical knowledge.

Our contributions include: (1) the OrthoGraphRAG framework and its multi-layer graph integrating private records with public data (e.g., UMLS) for orthopedic retrieval; (2) its graph-based retrieval and response generation process; and (3) an expert-validated evaluation demonstrating superior capabilities in comparative analysis (Chen et al., 2024). This approach offers a path towards safer, more capable, and contextually accurate LLMs for specialized clinical applications.

2. Related Work

Integrating external knowledge is vital for applying Large Language Models (LLMs) in specialized domains like healthcare. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) initially addressed this by conditioning LLM outputs on retrieved text. However, the complexity of medical data often necessitates structured knowledge graphs (KGs) to capture richer relational information beyond flat text.

Consequently, Graph-based RAG (GraphRAG) techniques emerged, leveraging KGs to improve LLM reasoning over structured data, thereby enhancing contextual understanding, factual accuracy, and explainability in medicine (Wu et al.,

[*]Equal contribution [1]NHS, UK [2]NHS, UK. Correspondence to: Venkatesh Tata <venkatesh.tata@nhs.net>, Zohra Bouchamaoui <zohra.bouchamaoui@nhs.net>, Nagavaishnavi V Bhaskara <nagavaishnavi.bhaskara@nhs.net>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

2024a). This field has seen rapid advancements, including systems for evidence-based medical KG querying (e.g., MedGraphRAG (Wu et al., 2024a)) and methods integrating diverse data sources like multimodal EHRs or specialized guidelines (Kresevic et al., 2024; Ke et al., 2024; Neupane et al., 2024)).

Existing GraphRAG systems are often broad or use single-layer graphs. To address these limitations, OrthoGraphRAG introduces a specialized multi-level GraphRAG architecture for orthopedics. Inspired by (Wu et al., 2024a), it features a distinct hierarchical KG integrating a Private Clinical Knowledge Graph (PKG) from clinic letters with the public UMLS (Bodenreider, 2004). This multi-level structure, combined with a retrieval process that augments subgraphs with textual evidence from original narratives, aims to enhance factual precision and multi-source information synthesis in this specialized domain.

3. Methodology

OrthoGraphRAG provides precise, context-aware orthopedic query responses using a multi-stage methodology: (1) data acquisition/preparation; (2) multi-level KG construction (private clinical narratives, public ontologies); (3) a KG-leveraging query processing and RAG pipeline; and (4) comprehensive evaluation. Figure 1 details the graph creation, while Figure 2 illustrates the query processing and response generation workflow.

3.1. Data Corpus and Pre-processing

OrthoGraphRAG utilizes a dual-source data strategy for clinical specificity and broad medical context:

- **Private Clinical Data:** Synthetic orthopedic clinic letters, rich in patient-specific details (diagnostic histories, treatments, outcomes), are pre-processed into coherent text chunks (e.g., sentences/paragraphs). The private clinical data comprises synthetic orthopedic clinic letters. This approach was chosen to navigate the ethical and privacy challenges associated with accessing real patient records, while still ensuring the data is inspired by and reflects real-world clinical scenarios. The scale of this dataset is therefore a direct consequence of this careful, privacy-preserving generation process. This segmentation aids granular retrieval, efficient downstream processing, and forms the response evidence base.
- **Public Medical Knowledge:** The Unified Medical Language System (UMLS) (Bodenreider, 2004) complements private data, serving as a standardized public medical knowledge repository. UMLS provides medical concepts, types, and relationships; its Metathesaurus is used for normalization/enrichment, and its Semantic Network for orthopedic-relevant broader relationships.

3.2. Multi-Level Knowledge Graph (KG) Construction

3.2.1. Private Clinical Knowledge Graph (PKG) Formulation

From the corpus of M pre-processed (chunked) orthopedic clinic letters, $D_{priv} = \{C_1, C_2, \ldots, C_M\}$, an automated knowledge extraction pipeline identifies entities and relationships for each chunk $C_k \in D_{priv}$:

- **Entity Extraction:** An LLM, denoted LLM_{ent_doc} (specifically Llama3.3 70B), performs medical entity recognition on each chunk C_k, identifying a set of medical entities:

  $$E_{pkg}^{(k)} = \{e_1^{(k)}, e_2^{(k)}, \ldots, e_{N_k}^{(k)}\}$$

  Each entity $e_i^{(k)} \in E_{pkg}^{(k)}$ is a tuple:

  $$e_i^{(k)} = (\text{name}_i, \text{type}_i, \text{span}_i)$$

  representing the recognized entity string (name_i), its assigned orthopedic-relevant semantic type (type_i, e.g., `Symptom`, `Diagnosis`, `Treatment`), and its textual position within C_k (span_i).
- **Relationship Extraction:** Subsequently, an LLM, LLM_{rel_doc}, extracts relationships $R_{pkg}^{(k)}$ between entity pairs $(e_i^{(k)}, e_j^{(k)})$ within the same chunk C_k. Each relationship $r \in R_{pkg}^{(k)}$ is represented as:

  $$r = (e_i^{(k)}, \text{rel\_type}_{ij}, e_j^{(k)})$$

  where rel_type_{ij} is a concise phrase describing their connection (e.g., `treated_by`, `caused_by`).

The Private Clinical Knowledge Graph (PKG), denoted $G_{PKG} = (E_{PKG}, R_{PKG})$, aggregates these per-chunk entities and relationships. Its components are defined as:

$$E_{PKG} = \bigcup_k E_{pkg}^{(k)} \quad \text{and} \quad R_{PKG} = \bigcup_k R_{pkg}^{(k)}.$$

For instance, an example relationship is:

`Patient_ID_123` $\xrightarrow{\texttt{HAS\_DIAGNOSIS}}$ `Degenerative_OA_Right_Knee`

3.2.2. Public Medical Knowledge Integration (UMLS)

The public KG layer utilizes the Unified Medical Language System (UMLS) (Bodenreider, 2004). We define

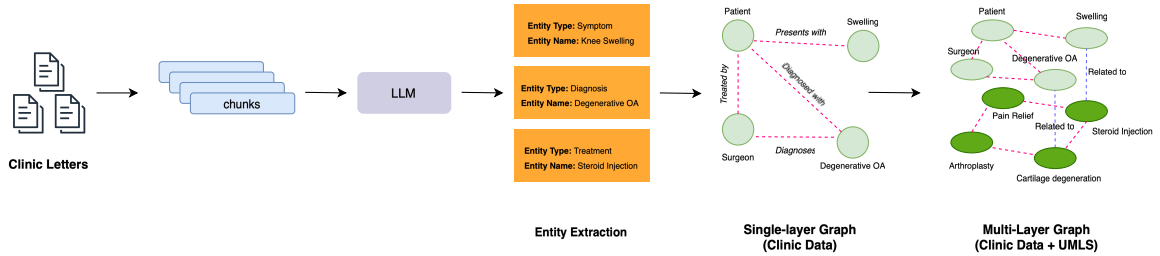

Figure 1. Diagram of the OrthoGraphRAG multi-level graph creation process, from clinic letters to an integrated graph with UMLS.

$G_{UMLS} = (E_{UMLS}, R_{UMLS})$ as a relevant UMLS subset, where E_{UMLS} includes Concept Unique Identifiers (CUIs) with associated information (names, semantic types), and R_{UMLS} represents inter-concept relationships (e.g., hierarchical, associative).

3.2.3. Inter-Graph Entity Linking and Enrichment

This phase creates the multi-level KG (G_{ML}) by bridging G_{PKG} and G_{UMLS}. Private entities $e_{pkg} \in E_{PKG}$ are linked to canonical UMLS concepts $e_{umls} \in E_{UMLS}$ if the cosine similarity of their contextual embeddings, $\phi(e)$, exceeds a threshold δ_{link}:

$$L(e_{pkg}, e_{umls}) = \mathbf{1}\left(\frac{\phi(e_{pkg}) \cdot \phi(e_{umls})}{\|\phi(e_{pkg})\|\|\phi(e_{umls})\|} \geq \delta_{link}\right)$$

3.3. Retrieval and Generation Process

The query processing pipeline (Figure 2) is OrthoGraphRAG's operational core, designed to interpret queries, retrieve relevant information, and synthesize clinically meaningful responses.

3.4. Query Processing and Contextual Subgraph Retrieval

Given a user's natural language query (Q), LLM_{ent_query} (Llama3.3 70B) extracts key medical entities (E_Q):

$$E_Q = \text{LLM}_{ent\_query}(Q)$$

These entities E_Q serve as entry points to the multi-level KG (G_{ML}). With LLM assistance (LLM_{graph_query}), a graph query (Q_G) is constructed from E_Q:

$$Q_G = \text{ConstructGraphQuery}(E_Q)$$

Executing Q_G against G_{ML} retrieves a contextual subgraph $G_S = (E_S, R_S)$:

$$G_S = \text{ExecuteGraphQuery}(Q_G, G_{ML})$$

G_S includes query entities, their k-hop neighbors, and connecting relationships, forming the initial structured query context.

3.5. Textual Evidence Augmentation and Response Synthesis

Subgraph G_S provides structured data, while original clinical narratives (D_{priv}) offer detailed context. Source text chunks ($C(e_s)$) are retrieved for private entities in G_S ($E_S \cap E_{PKG}$). Additionally, other text chunks ($C_k \in D_{priv}$) semantically similar to query Q and subgraph G_S are identified. This selection uses embeddings (ψ) and a similarity threshold (δ_{text}), ensuring highly relevant retrieved text (D_{rel}):

$$D_{rel} = \{C_k \in D_{priv} \mid \text{sim}(\psi(C_k), \psi(Q, G_S)) \geq \delta_{text}\} \cup \{C(e_s) \mid e_s \in (E_S \cap E_{PKG})\}$$

The retrieved subgraph G_S and textual evidence D_{rel} form a comprehensive prompt P_{final} for a generator LLM (LLM_{gen}, Llama3.3 70B). LLM_{gen} synthesizes this to produce a factually grounded, contextually appropriate answer (A), connecting patient-specific details with broader medical knowledge:

$$A = \text{LLM}_{gen}(\text{FormatPrompt}(Q, G_S, D_{rel}))$$

4. Results

This section evaluates OrthoGraphRAG's enhancement of orthopedic clinical decision support, primarily its end-to-end performance against baselines and ablated configurations across diverse clinical question categories (Table 1). The selection of key components like the entity linking retriever is detailed in Section B.1.

4.1. OrthoGraphRAG System Performance

The core evaluation of OrthoGraphRAG assessed its ability to answer clinical queries across Information Retrieval, Explanatory Reasoning, and Contextual Reasoning categories, benchmarking against leading LLMs and two ablated system versions using Llama3.3 70B. Results are in Table 1.

Overall Performance: The full OrthoGraphRAG system (`Private Data and UMLS GraphRAG`) achieved a **97.62%** overall success rate, significantly outperforming

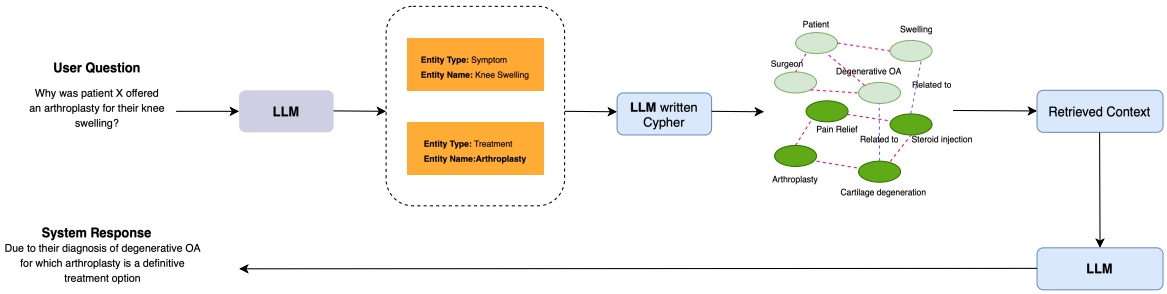

Figure 2. OrthoGraphRAG workflow: An LLM extracts query entities, guiding an LLM-generated Cypher query for subgraph retrieval from the multi-level KG. A final LLM then synthesizes this context into a natural language response.

Table 1. This table displays success rates, BLEU scores (n-gram overlap quality), and human feedback (expert contextual accuracy assessment for contextual reasoning) across reasoning categories, with question counts (n) provided for each..

Model	Information Retrieval (n=40)		Explanatory Reasoning (n=20)		Contextual Reasoning (n=10)	Overall (n=70)
	Success (%)	BLEU (%)	Success (%)	BLEU (%)	Domain Expert Score (%)	Success (%)
LLaMA2-70B	76.32	2.26	55.00	9.78	50.00	67.77
LLaMA3.1-70B	92.11	3.80	85.00	17.60	83.33	88.41
DeepSeek-R1-70B	89.47	0.60	85.00	3.43	83.33	89.86
LLaMA3.2-Vision-90B	84.21	3.37	70.00	14.89	66.67	78.26
Mistral-Latest	65.79	1.60	50.00	11.60	66.67	57.97
LLaMA3.1-8B	84.21	1.77	60.00	9.76	33.33	69.57
Private Data Graph (LLaMA3.3 70B)	90.53	3.50	15.00	3.00	5.00	56.73
Private Data GraphRAG (LLaMA3.3 70B)	100.00	4.20	97.00	19.50	17.00	87.29
Our Method	**100.00**	**4.50**	**100.00**	**20.50**	**83.33**	**97.62**

standalone LLMs like `DeepSeek-R1-70B` (89.86%) and demonstrating the benefit of its structured knowledge retrieval and augmentation. (Llama3.3 70B used for OrthoGraphRAG components).

Performance by Question Category:

- **Information Retrieval (n=40):** Full OrthoGraphRAG and its `Private Data GraphRAG` variant both achieved **100.00%** success with top BLEU scores (4.50% and 4.20%, respectively). The non-RAG `Private Data Graph` version was less successful (90.53%), highlighting RAG's value for text segment retrieval.
- **Explanatory Reasoning (n=20):** Full OrthoGraphRAG again achieved **100.00%** success (BLEU **20.50%**), followed by `Private Data GraphRAG` (97.00% success, 19.50% BLEU). The graph-only version performed poorly (15.00% success), underscoring RAG's importance for detailed explanations.

Evaluation of Contextual Reasoning: For this category, Domain Expert Scores were prioritized over BLEU to better capture clinical accuracy and nuanced understanding in complex inferential tasks.

These results robustly demonstrate that OrthoGraphRAG's multi-level knowledge graph and GraphRAG pipeline significantly improve LLM performance on complex clinical queries, especially for contextual reasoning requiring integration of patient-specific and general medical knowledge.

5. Conclusion

This paper introduced OrthoGraphRAG, a novel multi-level Graph Retrieval-Augmented Generation framework, to enhance LLM utility in orthopedics by tackling accuracy and complex reasoning challenges. OrthoGraphRAG constructs a multi-layer knowledge graph from private clinic letters and public UMLS data, coupled with a robust retrieval and generation pipeline.

Our key contributions include the framework design, the retrieval process, and expert evaluation, demonstrating OrthoGraphRAG's superior ability to answer complex clinical queries, especially those requiring deep contextual reasoning, compared to baseline models. This work signifies a promising step towards developing safer, more capable, and contextually-aware LLMs for specialized clinical decision support, aiming to improve patient care through more sophisticated AI tools.

References

Bodenreider, O. The unified medical language system (umls): Integrating biomedical terminology. *Nucleic acids research*, 32:D267–70, 02 2004. doi: 10.1093/nar/gkh061.

Chen, J., Lin, H., Han, X., and Sun, L. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:17754–17762, 03 2024. doi: 10.1609/aaai.v38i16.29728.

Gallego, F., López-García, G., Gasco-Sánchez, L., Krallinger, M., and Veredas, F. J. Clinlinker: Medical entity linking of clinical concept mentions in spanish, 2024. URL https://arxiv.org/abs/2404.06367.

Gao, Y., Li, R., Croxford, E., Tesch, S., To, D., Caskey, J., Patterson, B. W., Churpek, M. M., Miller, T., Dligach, D., and Afshar, M. Large language models and medical knowledge grounding for diagnosis prediction. *medRxiv*, 2023. doi: 10.1101/2023.11.24.23298641. URL https://www.medrxiv.org/content/early/2023/11/27/2023.11.24.23298641.

Ke, Y., Jin, L., Elangovan, K., Abdullah, H. R., Liu, N., Sia, A. T. H., Soh, C. R., Tung, J. Y. M., Ong, J. C. L., and Ting, D. S. W. Development and testing of retrieval augmented generation in large language models – a case study report, 2024. URL https://arxiv.org/abs/2402.01733.

Kresevic, S., Giuffrè, M., Ajčević, M., Accardo, A., Croce, S., and Shung, D. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *npj Digital Medicine*, 7, 04 2024. doi: 10.1038/s41746-024-01091-y.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401, 2020. URL https://arxiv.org/abs/2005.11401.

Neupane, S., Mitra, S., Mittal, S., Golilarz, N. A., Rahimi, S., and Amirlatifi, A. Medinsight: A multi-source context augmentation framework for generating patient-centric medical responses using large language models, 2024. URL https://arxiv.org/abs/2403.08607.

Wu, J., Zhu, J., Qi, Y., Chen, J., Xu, M., Menolascina, F., and Grau, V. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation, 2024a. URL https://arxiv.org/abs/2408.04187.

Wu, J., Zhu, J., Qi, Y., Chen, J., Xu, M., Menolascina, F., and Grau, V. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation, 2024b. URL https://arxiv.org/abs/2408.04187.

A. Detailed Analysis of Contextual Reasoning Example

The clinical query and subsequent detailed analysis presented in this appendix are based on a synthetic example. This example was carefully constructed in collaboration with a medical orthopedic expert to realistically represent the complexities encountered in clinical practice and to effectively illustrate the framework's capabilities.

This appendix provides a detailed step-by-step analysis of how the OrthoGraphRAG framework processes a complex clinical query requiring contextual reasoning and differential diagnosis. The example illustrates the synergy between the Private Clinical Knowledge Graph (PKG), Retrieval-Augmented Generation (RAG) from clinic letters, and the public UMLS knowledge graph.

A.1 Scenario: Input Query and System Output

The following input question was posed to OrthoGraphRAG, followed by the system's generated response.

INPUT QUESTION

```
Alex Riley with hospital number 80572391 presented with
paraesthesia predominantly in the distribution of the median
nerve, with a positive Tinel's sign at the wrist. Using
his symptom localisation and neurophysiological findings,
explain why Carpal Tunnel Syndrome was diagnosed over
Cubital Tunnel Syndrome.
```

ORTHOGRAPHRAG GENERATED ANSWER

```
Based on the provided context, Alex Riley's symptoms and
medical history suggest that he is experiencing numbness and
tingling sensations (paraesthesia) primarily in the thumb,
index, middle, and ring fingers. These are classic signs
of median nerve compression.

The Phalen's test was positive at both wrists indicating a
possible compression of the median nerve. Furthermore,
Tinel's sign was also positive when tapping over the
cubital tunnel on both arms suggesting irritation or
damage to the ulnar nerve as well.

However, given that Alex Riley's symptoms predominantly
involve the thumb, index, and middle fingers (which are
innervated by the median nerve) rather than the little
finger (innervated by the ulnar nerve), Carpal Tunnel
Syndrome is a more likely diagnosis.

Carpal Tunnel Syndrome occurs when the median nerve, which
runs from the forearm into the palm of the hand through a
small space in the wrist (carpal tunnel), is compressed or
pinched. This compression can cause numbness, tingling,
and weakness in the hand, particularly in the areas
served by the median nerve.

In contrast, Cubital Tunnel Syndrome involves compression
of the ulnar nerve at the elbow, leading to symptoms
primarily affecting the little finger and ring finger.
While Alex Riley does have a positive Tinel's sign over
the cubital tunnel, suggesting possible ulnar nerve
irritation, his primary symptom distribution aligns more
closely with median nerve involvement.

Therefore, based on the symptom localization (thumb, index,
middle fingers) and neurophysiological findings (positive
Phalen's test at both wrists), Carpal Tunnel Syndrome is
diagnosed over Cubital Tunnel Syndrome in this case.
```

```
Further diagnostic testing, such as Nerve Conduction
Studies mentioned in the plan, will help confirm the
diagnosis and guide treatment.
```

A.2 Step-by-Step Analysis of OrthoGraphRAG's Response Generation

The generation of the above answer involves several key stages:

A.2.1 Query Decomposition and Initial Entity Recognition

The system first processes the input query to identify key medical entities and the core clinical question.

- **Recognized Entities from Query**: `Alex Riley` (patient identifier), `hospital number 80572391` (patient identifier), `paraesthesia`, `median nerve distribution`, `positive Tinel's sign at the wrist`.
- **Semantic Interpretation**: The system identifies the task as a request for a differential diagnosis explanation, specifically to differentiate `Carpal Tunnel Syndrome` from `Cubital Tunnel Syndrome` based on the provided patient findings.

A.2.2 Retrieval from Private Clinical Data (PKG and RAG)

The identified entities serve as entry points into the multi-level knowledge graph, prioritizing the private clinical data related to the patient.

- **Private Clinical Knowledge Graph (PKG) Access**: Entities like `Alex Riley` and his hospital number are used to access the patient's structured data within the PKG. This graph might contain nodes for the patient, their diagnosed conditions, recorded symptoms, and procedures. The query entities `paraesthesia`, `median nerve`, and `Tinel's sign at wrist` are mapped to corresponding concepts linked to this patient in the PKG.
- **Retrieval-Augmented Generation (RAG) from Clinic Letters**: Concurrently, the RAG component retrieves relevant text chunks from Alex Riley's unstructured clinic letters. This is crucial for details not explicitly structured in the PKG or for elaborating on PKG nodes.
- **Key Patient-Specific Evidence Retrieved and Integrated**:
  - **Symptom Elaboration**: The phrase *'paraesthesia predominantly in the distribution of the median nerve'* from the query is enriched by RAG to include specific finger involvement: *'primarily in the thumb, index, middle, and ring fingers'*. This level of detail is often found in narrative text.
  - **Confirmation of Query-Mentioned Findings**: The *'positive Tinel's sign at the wrist'* is confirmed.
  - **Discovery of Additional Clinical Findings (not in query)**:
    - *'The Phalen's test was positive at both wrists...'*. This is a significant finding for median nerve compression, likely retrieved by RAG.
    - *'Tinel's sign was also positive when tapping over the cubital tunnel on both arms...'*. This finding suggests potential ulnar nerve involvement and is critical for a comprehensive differential diagnosis.
  - **Contextual Information for Future Management**: The mention of *'Nerve Conduction Studies mentioned in the plan'* is retrieved, adding context about the diagnostic pathway.

A.2.3 Integration of Public Medical Knowledge (UMLS Graph)

To provide broader medical context and standardized definitions, entities and concepts from the private data are linked to the UMLS graph.

- **Terminological Grounding and Definitions**:
  - `Paraesthesia` $\rightarrow$ UMLS provides the definition: *'numbness and tingling sensations'*.
- **Anatomical Knowledge**:

  - `Median nerve` $\xrightarrow{\text{UMLS: innervates}}$ *'thumb, index, middle fingers, and the lateral aspect of the ring finger'*.
  - `Ulnar nerve` $\xrightarrow{\text{UMLS: innervates}}$ *'little finger, and the medial aspect of the ring finger'*.

- **Disease Definitions and Pathophysiology**:
  - `Carpal Tunnel Syndrome` → UMLS describes it as *'compression of the median nerve... through a small space in the wrist (carpal tunnel)... caus[ing] numbness, tingling, and weakness in the hand, particularly in the areas served by the median nerve.'*
  - `Cubital Tunnel Syndrome` → UMLS describes it as *'compression of the ulnar nerve at the elbow, leading to symptoms primarily affecting the little finger and ring finger.'*

- **Clinical Test Significance**:
  - `Positive Phalen's test` → UMLS links this as indicative of *'median nerve compression'*.
  - `Positive Tinel's sign at wrist` → UMLS links this to *'median nerve irritation'* at the carpal tunnel.
  - `Positive Tinel's sign at cubital tunnel` → UMLS links this to *'ulnar nerve irritation'* at the elbow.

A.2.4 Information Synthesis and Deductive Reasoning by the Generator LLM

The generator LLM (Llama3.3 70B in this framework) receives the query, the retrieved subgraph from the multi-level knowledge graph (PKG + UMLS links), and the relevant text chunks from RAG. It then synthesizes this information:

- **Evidence Consolidation**: The LLM first consolidates all findings for Alex Riley: paraesthesia in median nerve distribution (thumb, index, middle, ring fingers), positive Tinel's at the wrist, positive Phalen's test, AND positive Tinel's at the cubital tunnel.
- **Weighing Conflicting or Complex Evidence**: The system acknowledges the Tinel's sign at the cubital tunnel which might suggest ulnar nerve issues. However, it correctly prioritizes the *'predominant'* nature of the symptoms aligning with median nerve distribution, as specified in the patient's detailed RAG-retrieved information.
- **Differential Reasoning Logic**:
  1. The primary symptoms (*'thumb, index, middle fingers'*) are mapped to median nerve innervation (via UMLS).
  2. Phalen's test and Tinel's at the wrist strongly support median nerve pathology at the wrist (Carpal Tunnel Syndrome, via UMLS).
  3. The symptoms are less consistent with primary ulnar nerve pathology, despite a positive Tinel's at the cubital tunnel, because the main sensory disturbance is in the median nerve territory.
  4. The LLM uses the definitions and characteristics of Carpal Tunnel Syndrome and Cubital Tunnel Syndrome from UMLS to frame the explanation.
- **Formulating the Explanation**: The LLM constructs an argument that explains why Carpal Tunnel Syndrome is the more likely diagnosis by contrasting the evidence for both conditions and highlighting the stronger support for median nerve compression at the wrist.

A.2.5 Grounding and Generation of the Final Answer Structure

The final textual answer is generated with clear grounding to the synthesized information.

- **Introduction**: Acknowledges the patient and the nature of symptoms.
- **Presentation of Findings**: Lists all relevant positive findings, including those from RAG not present in the initial query.
- **Core Reasoning**: Explicitly states the importance of symptom localization (median vs. ulnar nerve distribution).
- **Explanation of Conditions**: Defines both Carpal Tunnel Syndrome and Cubital Tunnel Syndrome, drawing on UMLS knowledge.

- **Comparative Analysis**: Directly addresses why Carpal Tunnel Syndrome is favored despite some indication of potential ulnar nerve irritation.
- **Conclusion and Outlook**: Summarizes the diagnostic reasoning and mentions further confirmatory tests (Nerve Conduction Studies, as retrieved by RAG).

B. Supplementary Results

B.1. Entity Linking Retriever Selection

Effective inter-graph entity linking is crucial for OrthoGraphRAG (Gallego et al., 2024). We selected SapBERT as our embedding model after benchmarking biomedical language models on public datasets (MedMentions, NCBI Disease), where SapBERT demonstrated superior performance. This choice provides a strong foundation for our multi-level KG construction. Detailed benchmarking results are in Appendix B.1, Table 2.

The following table presents the detailed performance comparison for the entity linking retriever selection, as discussed in the main paper.

Table 2. Performance comparison of different model configurations on biomedical entity linking tasks. MRR = Mean Reciprocal Rank; Acc@k = Accuracy at rank k. Best results are highlighted in **bold**. Datasets: MedMentions (352,496 mentions) and NCBI Disease (713 mentions).

Model Configuration	MedMentions			NCBI Disease		
	Acc@1	Acc@5	MRR	Acc@1	Acc@5	MRR
Base Model (PubMedBERT_FT)	0.4304	0.5939	0.5000	0.5806	0.6732	0.6114
SapBERT Fine-tuning (SapBERT)	**0.5047**	**0.7525**	**0.6076**	**0.8597**	**0.9790**	**0.9101**
Improvement	*+0.0743*	*+0.1586*	*+0.1076*	*+0.2791*	*+0.3058*	*+0.2987*
Baseline (BERT-base)	0.4291	0.5829	0.4934	0.5540	0.6157	0.5828
Alt. Domain Adpt. (BioBERT)	0.3938	0.5667	0.4676	0.5891	0.6606	0.6171