



UWL REPOSITORY

repository.uwl.ac.uk

Retrieval-Augmented Large Language Model for clinical decision support

Saidu, Fatima and Wall, Julie ORCID logoORCID: <https://orcid.org/0000-0001-6714-4867> (2026)
Retrieval-Augmented Large Language Model for clinical decision support. *Electronics*, 15 (3).

<https://doi.org/10.3390/electronics15030555>

This is the Published Version of the final output.

UWL repository link: <https://repository.uwl.ac.uk/id/eprint/14513/>

Alternative formats: If you require this document in an alternative format, please contact:
open.research@uwl.ac.uk

Copyright: Creative Commons: Attribution 4.0

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy: If you believe that this document breaches copyright, please contact us at open.research@uwl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Rights Retention Statement:

Article

Retrieval-Augmented Large Language Model for Clinical Decision Support with a Medical Knowledge Graph

Fatima Saidu and Julie Wall ^{*}

School of Computing and Engineering, University of West London, London W5 5RF, UK;
32147050@student.uwl.ac.uk

* Correspondence: julie.wall@uwl.ac.uk

Abstract

This study examines clinician interactions with a Knowledge Graph (KG)-enhanced Large Language Model (LLM) for diagnostic support, with an emphasis on the rare condition pseudohypoparathyroidism (PHP). Ten medical professionals engaged with simulated diagnostic scenarios, using the KG-enhanced LLM to support reasoning and validate differential diagnoses. Evaluation included basic model performance (RAGAS = 0.85; F1 = 0.79) and clinician-centered outcomes, such as diagnostic conclusions, confidence, adherence, and efficiency. Results show the tool was most valuable for rare or uncertain cases, increasing clinician confidence and supporting reasoning, while familiar cases elicited selective adoption with minimal AI engagement. Participant feedback indicated generally high usability, accuracy, and relevance, with most reporting improved efficiency and trust. Statistical analysis confirmed that AI assistance significantly reduced time-to-diagnosis ($t(8) = 4.99, p = 0.001$, Cohen's $d_z = 1.66$, 95% CI [73.8, 197.2]; Wilcoxon $W = 0.0, p = 0.0039$). These findings suggest that KG-enhanced LLMs can effectively augment clinician judgment in complex cases, serving as reasoning aids or educational tools while preserving clinician control over decision-making. The study emphasizes evaluating AI not only for accuracy, but also for practical utility and integration into real-world clinical workflows.

Keywords: knowledge graph; large language model; retrieval-augmented generation; clinical decision support system; diagnostic accuracy; explainable AI; pseudohypoparathyroidism

1. Introduction



Academic Editors: Domenico Rosaci and Yazhou Zhang

Received: 30 September 2025

Revised: 21 December 2025

Accepted: 19 January 2026

Published: 28 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\) license](#).

Healthcare systems continue to face persistent challenges due to physician shortages, increasing workloads, and high rates of stress and burnout [1]. Clinical decision support systems (CDSS), which are designed to improve healthcare outcomes, also help mitigate these pressures by providing clinicians with relevant, evidence-based guidance to inform their decisions [2]. Large Language Models (LLMs) show promise in supporting diagnostic reasoning and reducing errors [3,4]. However, LLMs often struggle with factual accuracy and long-term information retention [5,6]. They can also produce misleading or fabricated content, a limitation known as hallucination and their outputs may lack transparency, reliability, and alignment with domain-specific standards [7].

To address these limitations, a promising strategy is to enhance LLMs with Knowledge Graphs (KGs). KGs are structured representations that explicitly encode entities and the relationships between them, facilitating organised data management and supporting domain conceptualisation [8]. By grounding LLMs in this structured knowledge, KGs

improve factual accuracy, interpretability, and domain relevance [9]. Current studies [10,11] have demonstrated that integrating KGs can enhance LLM performance in diagnostic support. However, evaluations often focus primarily on performance metrics such as accuracy, without considering whether these improvements address clinician concerns or support their reasoning.

In practice, the benefits of CDSSs are frequently limited by low adoption and inconsistent use among clinicians [12]. Key factors influencing adoption include the perceived usefulness and relevance of the information, as well as the system's ease of use and efficiency [13]. Clinicians are more likely to engage with support tools in situations of high diagnostic uncertainty, such as rare or atypical cases, whereas familiar cases may evoke a low perceived need for assistance. This highlights the importance of evaluating not only model performance but also understanding how these models address clinician concerns and actively support clinical reasoning across varying diagnostic scenarios.

This study examines how clinicians interact with a KG-enhanced LLM for diagnostic support when presented with rare case presentations that can be easily misdiagnosed as common conditions, and compares this interaction with their approach when faced with a familiar case that may not require support. Contributions include the following:

1. Exploration of KG-enhanced LLM in Rare Case Support: Investigates how clinicians selectively use AI assistance for rare or atypical cases, highlighting contexts where decision support is most valuable.
2. Understanding Clinician–AI Interaction Patterns: Observes clinician behaviour across familiar and unfamiliar cases to identify patterns of selective adoption, reliance, and impact on diagnostic confidence.

2. Related Works

LLMs have advanced natural language processing (NLP), enabling capabilities in text generation, summarisation, and semantic interpretation, supporting CDSS, NLP processing for electronic health records, medical question-answering systems, and healthcare education [4]. However, the reliability of LLM-generated content in medical contexts remains a significant concern, as limited exposure to curated medical data during training increases the risk of factual inaccuracies [14], deviations from established guidelines [15], and the amplification of biases, including those related to ethnicity [16]. Such limitations pose risks, particularly in high-stakes applications, as over-reliance on training data can lead to diagnostic errors, which may cause inappropriate treatment, unnecessary interventions, and significant harm [17,18].

To address these limitations, recent research has explored combining LLMs with KGs, leading to two primary integration paradigms that focus on LLM enhancements: KG-enhanced LLMs and synergised LLMs + KGs [19]. KG-enhanced LLMs incorporate structured knowledge to improve the accuracy, consistency, and interpretability of model outputs during different stages of the LLM cycle, while in a bidirectional, synergised LLM–KG integration, both systems iteratively support each other [20].

KGs support LLMs in handling complex queries by providing explicit relationships between entities, which helps LLMs reason over multiple connected concepts, resolve ambiguities, and generate outputs grounded in verified knowledge. This helps LLMs reduce hallucinations and enhance reasoning accuracy, which is particularly valuable in clinical settings where accuracy and traceability are essential [20]. They also enhance data integration, contextualisation, and decision-making, improving adaptability to real-world clinical scenarios [21]. However, these improvements remain constrained by the coverage and correctness of the underlying graph; any incompleteness or bias limits the precision of the resulting model [22].

Many studies have built and evaluated the performance of LLMs when augmented with KGs, reporting promising improvements in reasoning, prediction, and classification tasks across different medical domains. Table 1 summarises recent representative works, highlighting their integration approach, medical application, methodology, models and datasets, evaluation metrics, and key results.

Table 1. Overview of representative studies integrating Knowledge Graphs with LLMs.

Paper	KG-LLM Integration Type	Medical Application	Evaluation Methods & Metrics	Key Outcomes
KG-Rank [15]	KG-enhanced LLM	Question & Answering system	Quantitative analysis of KG-based reranking framework using ROUGE-L, BERTScore, MoverScore & BLEURT	KG-based reranking improved QA performance; GPT-4 achieved +18% ROUGE-L improvement on ExpertQA-Bio dataset
DR. KNOWS [23]	Synergised Integration	Diagnostic Prediction	Quantitative analysis of KG path augmentation in prompt-based LLMs (ROUGE-2, ROUGE-L, CUI Precision, Recall, F1) & Human evaluation with 2 medical professionals	KG augmentation improved ChatGPT(GPT-3.5-Turbo) diagnostic F1 from 20.96 to 26.02 (5-shot). Human evaluation confirmed a 5% absolute improvement in correct diagnostic rationale agreement ($p < 0.001$).
XAI-Based CDSS [11]	Synergised Integration	Clinical Decision Support	Quantitative analysis of classification tasks using Precision, Recall, F1	KG integration led to modest F1 gains across mental health detection tasks, including a +0.1291 increase for Cause/Factor Detection with the Joint RoBERTa model.
MedIKAL [10]	Synergised Integration	Disease Diagnosis	Quantitative analysis of automatic diagnosis on datasets; comparison with LLMs, KG-enhanced LLMs, and synergised KG-LLMs using Precision, Recall, F1	Outperformed all other models including KG-Rank on CMD dataset on all metrics

While these studies demonstrate the potential of KG-enhanced LLMs, their evaluations largely emphasise correctness, factuality, or text similarity metrics such as ROUGE and BLEU. Although such metrics demonstrate measurable improvements on standard benchmarks, they provide limited insight into the practical utility of these models in supporting clinician reasoning. These metrics capture surface-level performance but fail to reflect critical aspects of clinical decision-making, including reasoning quality, reliability, and the ability to justify outputs in complex or uncertain scenarios. Consequently, a disconnect remains between prevailing evaluation frameworks and the real-world requirements of support systems, where nuanced reasoning and clinically meaningful guidance are essential.

Even when human evaluation is included (e.g., Gao et al., 2025 [23]), it often measures agreement with expert labels or retrospective performance on benchmark tests rather than

on clinician behavior, selective adoption, or decision-making under real-world diagnostic uncertainty [24]. As a result, there is little understanding about how structured knowledge integration affects clinician interaction, reliance, and the likelihood of clinical adoption in real-world workflows. As a result, little is known about how structured knowledge integration affects clinician reliance, interaction patterns, or adoption in real-world workflows. Moreover, the heavy reliance of LLMs on effective prompting means that model use depends not only on model capabilities but also on clinician habits, experience with prompting, and expectations of the tool, all of which can influence perceived usefulness, efficiency, and willingness to adopt. This has practical implications: verbose outputs or repeated prompting may reduce efficiency and lead to selective adoption [25].

To address this gap, our study shifts the focus from model-centric performance evaluation to clinician-centred assessment. Rather than aiming to establish performance superiority, this work adopts an exploratory, formative approach to understand how clinicians interact with a KG-enhanced LLM across different diagnostic scenarios. We observe how KG-enhanced LLMs are used differently depending on diagnostic uncertainty, capturing interactions, trust, reasoning, and confidence, emphasising practical utility.

3. Materials and Methods

This section describes the design of a KG-enhanced LLM and an exploratory evaluation framework aimed at understanding how such systems may support clinical reasoning in rare disease diagnosis. Rather than optimising or benchmarking model performance, the focus of this study is on clinician interaction, perceived utility, and trust when engaging with a KG-enhanced diagnostic tool. The proposed framework embeds structured clinical context into LLM responses to promote grounded, interpretable outputs.

3.1. *Pseudohypoparathyroidism: Disease and Case Selection*

Pseudohypoparathyroidism (PHP) encompasses rare endocrine disorders characterised by end-organ resistance to parathyroid hormone (PTH), with subtypes including type 1A, type 1B, type 1C, pseudo-PHP, and type 2 [26]. Given the high level of clinical suspicion required to distinguish PHP from conditions such as idiopathic epilepsy or other causes of hypocalcaemia, an LLM augmented with a structured MKG could assist clinicians by systematically analysing symptom patterns and laboratory findings. Three case studies were selected for evaluation. The first two focus on PHP subtypes, and the third involves a common condition not included in the KG, serving as a control. This control ensures the KG does not provide information outside its scope and helps establish a baseline for clinician confidence when handling familiar conditions versus rare diseases. Quantitative and qualitative methods assess diagnostic accuracy, clinical relevance, and practical utility, offering insights into the potential of KG-enhanced LLMs to reduce misdiagnosis in complex cases.

3.1.1. Case Study 1 (Typical Presentation: PHP Type 1A)

Based on Najim et al. (2020), this scenario describes a 34-year-old woman who presented with symptomatic hypocalcaemia and was ultimately diagnosed with PHP type 1A [27]. Laboratory investigations revealed abnormal calcium, phosphate, and parathyroid hormone levels consistent with hormonal resistance. Additionally, the patient exhibited features of Albright hereditary osteodystrophy (AHO), consistent with the classical presentation of PHP type 1A, a rare but clinically important condition that is often underdiagnosed.

3.1.2. Case Study 2 (Atypical Presentation: Pseudo-PHP)

Adapted from Najim et al. (2020), this case involves a 9-year-old girl attending a routine check-up to monitor growth given her short stature [27]. She had no clinical complaints, and laboratory findings were normal. Despite the absence of biochemical abnormalities, the child exhibited features characteristic of AHO, including a round face, short stature, and brachydactyly. The constellation of findings suggested pseudo-PHP, an atypical variant in which phenotypic features are present without hormonal resistance.

3.1.3. Case Study 3 (Control: Severe Malaria, Out-of-Scope Condition)

As a control, this case centres on a 55-year-old woman who developed severe Plasmodium falciparum malaria following a trip to Ghana [28]. Upon returning to Florida, she was admitted with fever, confusion, and hypotension and was treated successfully with intravenous artesunate. Because malaria falls outside the KG's scope, this case was included to assess hallucination risk when the system encounters conditions without KG coverage.

3.2. KG Construction

The KG was manually constructed from scientific publications, including the peer-reviewed literature, textbooks, and clinical guidelines issued by authoritative institutions, publishers, and researchers. These sources are highly trustworthy, widely available, and provide a reliable foundation for creating a disease-specific KG. Manual construction was chosen to ensure careful curation of clinically relevant entities and relationships, minimising errors or omissions that automated methods might introduce. The focus was on PHP and one of its common misdiagnoses: epilepsy [29]. This diagnostic error can occur when patients present with seizure-like complications caused by chronic hypocalcaemia or when tetany is mistaken for seizures. This overlap highlights the need for careful curation of clinically relevant knowledge.

We manually extracted relevant entities and relationships from these sources. Extraction focused on key clinical features, diagnostic criteria, treatment options, and ways in which PHP is commonly confused with epilepsy. Entities and connections were organised into structured sets to identify critical relationships and construct a comprehensive knowledge base. A simple, well-defined graph schema was designed in Neo4j (Desktop 5.x) to capture both hierarchical and semantic relationships. Hierarchical relationships represent subtype structures, such as the IS_A link between PHP and its subtypes PHP type 1A and type 1B. Semantic relationships reflect clinical associations across entity types; for example, epilepsy has a HAS_SYMPTOM relationship with seizure and a DIAGNOSED_BY link to EEG (Figure 1). This schema supports meaningful clinical queries and enables complex reasoning across diagnostic pathways and differential diagnoses. Each node was populated with a detailed description of the associated entity. Nodes were also assigned properties to encapsulate relevant clinical information and salient characteristics. This enabled accurate, enriched representations of entities, making it easier to trace relationships and identify potential diagnostic patterns.

In total, the KG comprises 160 nodes, 252 edges, 69 node labels, and 71 relationship types. The KG's scope is small but focused, centering on PHP and selected related details, such as overlapping features with epilepsy, for example, differentiating seizures from tetany. While the size is limited, this was intentional for an exploratory study, allowing for the careful assessment of clinician interaction and system utility.

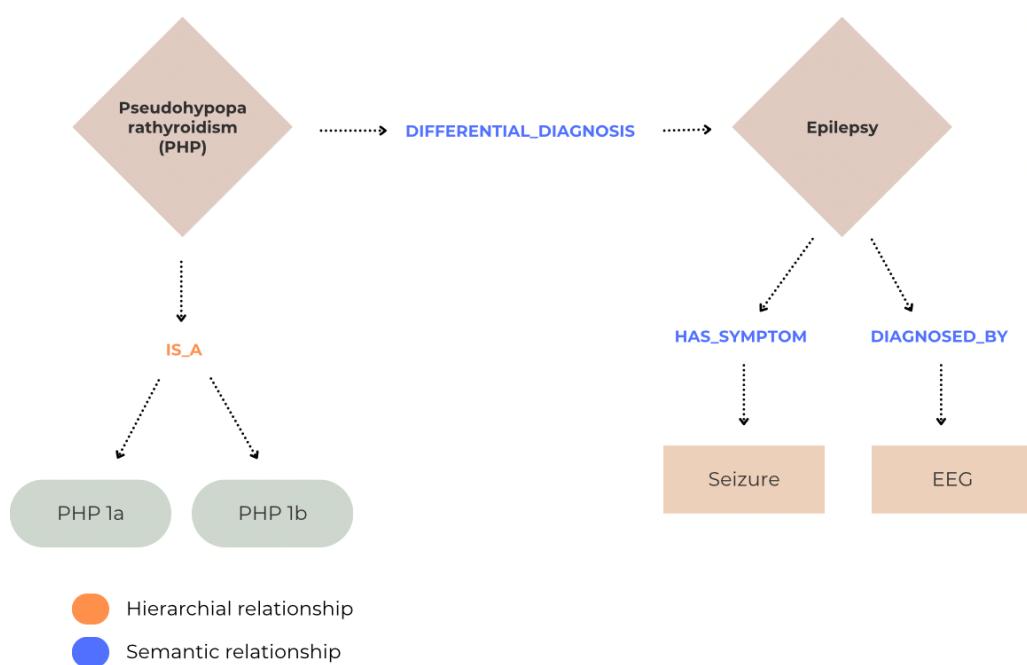


Figure 1. This schematic illustrates examples of hierarchical and semantic relationships in the Knowledge Graph. Hierarchical links represent subtype structures (e.g., PHP → PHP type 1A), while semantic links capture clinical associations (e.g., epilepsy HAS_SYMPTOM seizure).

3.3. LLM Integration

To support the KG-enhanced system, GPT-4o-mini (version date: 18 July 2024) was selected for its predictable reasoning and computational efficiency [30]. Compared with GPT-3.5-turbo, GPT-4o-mini demonstrates approximately four times the reasoning capacity and operates at roughly three times the processing speed, while supporting multimodal inputs and extended context lengths. It was chosen over other medical LLMs not for maximal diagnostic accuracy, but to enable smooth, low-latency interactions that allow clinicians to explore and evaluate the system's utility without introducing variability or unnecessary complexity.

The integration follows a straightforward, context-enriched framework. When a user query is received, relevant triples are retrieved directly from the Neo4j knowledge graph using Cypher queries. These triples are structured into a textual context and injected into a fixed prompt for GPT-4o-mini, which generates responses constrained to the KG content. This ensures that the LLM's reasoning is grounded in structured medical knowledge, avoiding hallucinations while providing context-aware guidance.

LangChain orchestrates the workflow, combining query handling, KG retrieval, and prompt construction into a seamless pipeline.

The fixed prompt template used to constrain the LLM's responses to the retrieved knowledge graph context is provided in Appendix A.

3.4. System Architecture

The KG-enhanced LLM system uses a lightweight, modular design tailored for exploratory evaluation. The backend is implemented in Python 3.12 with FastAPI, exposing a single synchronous API endpoint. Queries are processed sequentially, with responses generated only after full KG retrieval and LLM reasoning.

The KG is stored in Neo4j (Desktop 5.x), and a custom service layer retrieves subject-predicate-object triples, including node labels, descriptions, and properties. Retrieved triples are structured and serialized into a textual context for the LLM. No ranking, thresholds, similarity filtering, or re-ranking mechanisms are applied.

GPT-4o-mini handles reasoning via LangChain in a zero-shot configuration, using a fixed prompt that instructs the model to respond solely with the KG context and indicate when information is unavailable. Default model parameters are used, including a temperature of 0.7. Response times reflect the cumulative cost of KG retrieval, context preparation, and LLM generation; no latency measurements or optimization strategies were applied (see Table 2).

The frontend is minimal, featuring a query input box, system title, and response display area. This simple interface ensures that clinician focus remains on the system's reasoning support rather than the interaction design.

Table 2. Key Parameters and Design Choices of the KG-Enhanced LLM Framework.

Parameter	Setting/Description
Knowledge Graph Retrieval	
Knowledge retrieval strategy	Direct graph traversal over Neo4j using Cypher queries
Retrieval limit (Top- k)	Not explicitly parameterised; up to 1000 triples retrieved per query
Similarity metric	Not applicable (no embedding-based similarity search)
Re-ranking	Not enabled
Filtering thresholds	None applied
Context Handling	
Context formatting	Subject–relation–object triples rendered as text
Context injection method	Full retrieved context appended to the LLM prompt
LLM Configuration	
LLM model	GPT-4o-mini (via LangChain)
LLM temperature	0.7
System Behavior/Constraints	
Abstention behaviour	Implicit, enforced via prompt instruction when KG coverage is insufficient
Response grounding	Restricted to provided KG context via prompt constraint
Latency optimisation	Not explicitly optimised or measured

3.5. Evaluation Process

The evaluation adopts a formative, mixed-methods design intended to explore clinician interaction with a KG-enhanced LLM rather than to establish definitive performance gains. Evaluation is divided into two stages: (1) a limited technical assessment to ensure basic system reliability and faithfulness, and (2) a clinician-centered evaluation focused on usability, trust, and perceived support for diagnostic reasoning (Figure 2).

3.5.1. Model Evaluation

Model evaluation follows a structured and automated approach using the Retrieval-Augmented Generation Assessment System (RAGAS). A dataset of 10 clinically relevant questions with expert-validated reference answers was prepared, balancing short-form and long-form queries. The RAGAS evaluation framework, designed for RAG systems [31], comprises five components: faithfulness, context precision, context recall, answer relevancy, and answer correctness. These metrics are computed automatically by the framework.

Faithfulness reflects factual accuracy, calculated as the number of correct facts divided by the total number of facts in the response, ensuring the system avoids introducing misleading or incorrect information. Answer relevance measures the proportion of relevant concepts in a response, indicating whether outputs address the clinical query meaningfully. Context precision captures the proportion of retrieved sentences that are relevant, highlighting retrieval efficiency, while context recall evaluates whether the system retrieves all relevant KG knowledge available for the query. Answer correctness combines semantic similarity and factual accuracy to assess alignment with validated ground truth.

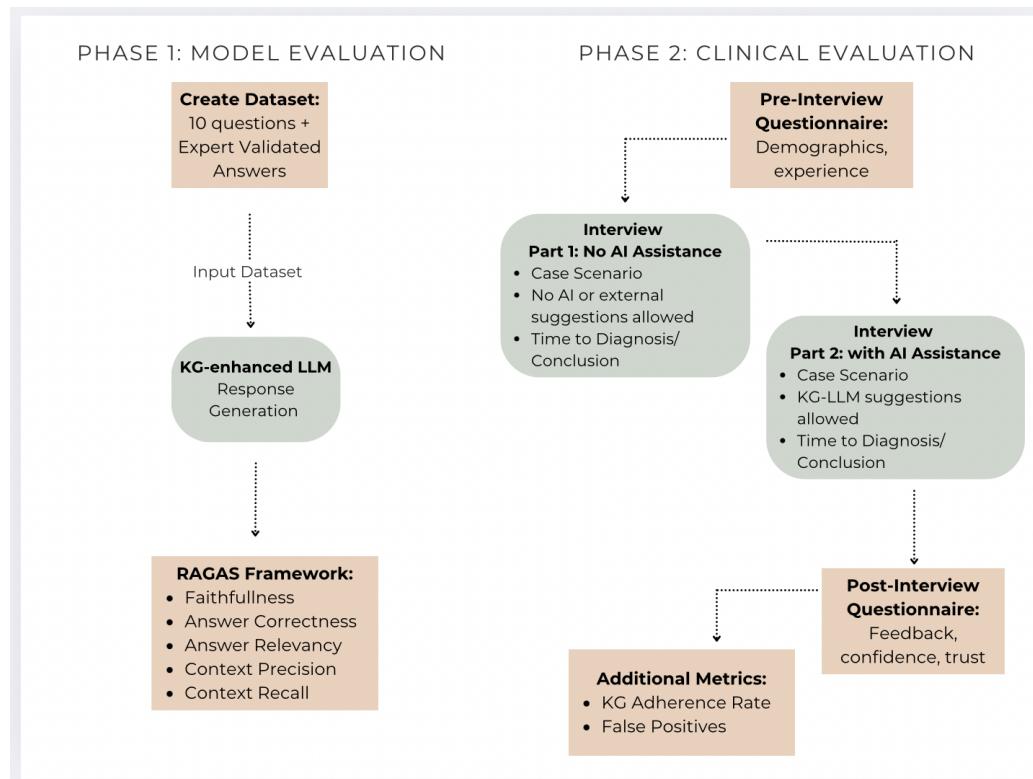


Figure 2. Overview of the evaluation process. The model evaluation uses a dataset of expert-validated questions assessed with RAGAS to measure faithfulness, relevance, and KG retrieval performance. The clinical evaluation involves a two-phase simulation where clinicians first diagnose cases unaided, then with KG-enhanced LLM support, capturing diagnostic accuracy, confidence, and adherence to AI recommendations.

While ROUGE metrics (Recall-Oriented Understudy for Gisting Evaluation) were also considered, their applicability was limited by the small sample size and the exploratory nature of this study. The evaluation prioritised contextual relevance, transparency, and faithfulness over surface-level text-overlap metrics such as ROUGE.

3.5.2. Clinical Evaluation

Clinical evaluation examines how clinicians interact with the KG-enhanced LLM and perceive its utility in diagnostic decision-making. The evaluation consists of a pre-interview survey, a two-phase clinical simulation, and a post-interview survey to capture both performance and perceived clinical utility. Before the simulation, participants complete a pre-interview survey to collect demographic data, including years of experience and specialty. During the interview, participants are presented with three test cases: two corresponding to diseases represented in the KG (Cases 1 and 2, differing in subtype and complexity) and one control case featuring a disease outside the KG (Case 3). This design allows for the assessment of the system's ability to manage both familiar and "out-of-scope" scenarios, including appropriate handling of uncertainty.

The simulation occurs in two phases. In Phase 1, participants complete the cases without AI assistance. They record their diagnostic conclusions and time-to-diagnosis while the interviewer acts as the patient, responding to history-taking questions. Participants may request physical examination findings, laboratory tests, and radiology results, which are provided according to the case and clinical judgment. The order of case presentation varied: the first participant completed Case 2 → Case 1 → Case 3; the next eight participants followed Case 1 → Case 2 → Case 3, starting with a straightforward case before progressing to a more atypical one; and the final participant, with more endocrinology experience,

completed Case 3 → Case 1 → Case 2 to explore the effect of starting with the most familiar condition.

In Phase 2, participants complete the same cases with access to the KG-enhanced LLM. They again record diagnostic conclusions, adherence to KG suggestions, and time-to-diagnosis. Time-to-diagnosis is treated as a secondary, descriptive outcome. Because participants have prior exposure to the cases and interactions with the AI include typing and prompting, these results do not allow for causal inference and are reported for illustrative purposes only.

The primary evaluation focuses on three key clinical outcomes: Diagnostic Assessment, confidence, and KG adherence. Diagnostic accuracy measures whether the participant reaches the correct diagnosis, including subtypes when applicable. KG adherence reflects the proportion of AI recommendations integrated into the final diagnosis, classified as full (AI suggestions fully incorporated and leading to the correct diagnosis), partial (some engagement with AI outputs but the case is not fully resolved), or none (AI disregarded or no diagnosis reached). Diagnostic confidence is measured using a 5-point Likert scale before and after AI interaction, recorded during post-interview feedback, to capture changes attributable to KG assistance. Post-interview feedback also captures participant perceptions of accuracy, relevance, usability, trust, and overall satisfaction with the AI-assisted workflow. Secondary metrics include time-to-diagnosis and observed instances where the model provided misleading or incorrect suggestions outside its scope. These instances were recorded descriptively, with a target threshold of <10%, but formal statistical false-positive rates were not calculated. Figure 3 shows the user interface used by clinicians to submit queries and view KG-enhanced LLM responses during the evaluation.

Knowledge Graph Query

What is PHP?

Answer:

PHP refers to Pseudohypoparathyroidism, which is a disease characterized by the body's resistance to parathyroid hormone (PTH), leading to biochemical abnormalities such as hypocalcemia and hyperphosphatemia. It has several subtypes, including Pseudohypoparathyroidism IA, Pseudohypoparathyroidism IB, Pseudohypoparathyroidism IC, Pseudohypoparathyroidism II, and Pseudopseudohypoparathyroidism. Treatment goals for PHP include maintaining normal calcium and phosphate levels, avoiding hypercalciuria, normalizing PTH levels, and preventing nephrocalcinosis.

Figure 3. User interface.

3.5.3. Participant Selection Rationale

For this pilot, the target group consists of 10–15 clinicians, including general practitioners (GP), residents, and junior doctors with 1–7 years of experience. This mix allows for diverse perspectives, ensuring the KG-enhanced LLM is evaluated by those most likely to benefit from decision support. Novice clinicians (1–3 years) are particularly likely to improve efficiency and confidence with KG assistance, as they tend to rely more on external support than experienced clinicians [32]. They may also exhibit the greatest gains in diagnostic confidence. Mid-level clinicians (4–7 years) provide valuable insight into KG usefulness for atypical cases, where clinical experience may be limited. Focusing on this group also avoids potential bias from expert clinicians (10+ years), who may dismiss the KG due to overconfidence in their diagnostic skills. This participant profile aligns with assessing the feasibility and early utility of the KG in real-world settings. This selection

supports the study's aim of generating early, clinician-informed insights to guide future system design and evaluation, rather than establishing generalisable effectiveness claims.

3.6. Ethical Considerations

Ethical principles were paramount, and we ensured compliance with key guidelines and regulations governing AI in healthcare. These steps ensured adherence to clinical ethical standards while addressing concerns related to generative AI in diagnostic decision-making.

3.6.1. Patient Data Privacy and Confidentiality

One critical ethical consideration was patient data privacy. To avoid risks related to sensitive data, participants were medical professionals, and cases were based entirely on published or edited materials with no personal information. This adheres to the General Data Protection Regulation (GDPR), which mandates explicit consent and transparency in the use of patient data [33]. Using simulated cases avoided real clinical settings and ensured no patient data were compromised. This process aligned with GDPR principles of protecting personal data and securing patient confidentiality.

3.6.2. Clinical Oversight and AI Limitations

The National Institute for Health and Care Excellence (NICE) guidelines stress the importance of healthcare professionals reviewing AI outputs before making clinical decisions [34]. We followed this principle by ensuring that the AI-driven, KG-enhanced LLM outputs were not solely relied upon for final diagnosis but used only to assist clinicians. A medical professional reviewed all diagnostic decisions based on AI recommendations, mitigating risks associated with over-reliance on AI. Moreover, NICE guidelines highlight that AI systems use fixed algorithms in clinical settings, limiting their ability to adapt to real-time data [35]. By conducting the study in a controlled environment where clinicians retained authority over final diagnoses, we ensured that AI complemented, rather than replaced, clinical judgement.

3.6.3. Risk Assessment and Transparency

In line with the G7 AI Code of Conduct, which advocates continuous risk assessment and transparency, the study prioritised transparency in its methodologies and results [36,37]. Detailed information about AI capabilities, limitations, and data used was made available to all participants. This ensured clinicians were well-informed about system operation, strengths, and constraints.

3.6.4. Cultural Sensitivity and Inclusivity

Ethical guidelines also call for consideration of cultural factors in healthcare. Although cases varied in age, gender was not considered a differentiating factor, in line with real-world clinical cases. This choice reflected the need to represent a broad spectrum of patient demographics. Recognising that cultural factors can influence diagnosis and care, future studies should incorporate a broader range of cultural contexts to align with evolving standards for inclusivity and cultural sensitivity.

3.6.5. Medical Professional Involvement

Throughout the study, medical professionals played integral roles in development and evaluation phases, addressing concerns about over-reliance or potential misuse of technology in clinical practice. Active clinician involvement ensured appropriate oversight of AI use. Consistent with NICE and the G7 AI Code of Conduct, healthcare professionals retained control over AI-generated findings, with AI as a supportive tool rather than a

decision-making authority. This approach aligns with ethical principles that prioritise human expertise in healthcare. By ensuring AI enhanced, rather than replaced, clinical decision-making, the study emphasised the central role of clinicians in the diagnosis. This involvement also helped mitigate risks associated with over-reliance on AI systems, ensuring that final diagnoses remained with experienced medical professionals.

By following these ethical principles and complying with established guidelines, the study ensured responsible deployment of the AI model. These measures safeguarded patient privacy, maintained clinical oversight, and fostered clinicians' trust in AI technologies. Ultimately, by adhering to these ethical standards, the study aimed to establish a framework for responsible, transparent use of AI in healthcare to enhance diagnostic accuracy and support clinicians' decision-making.

4. Results

4.1. Model Evaluation Results

The primary aim of the model evaluation was to verify that a minimally viable, clinically coherent knowledge graph could be successfully queried and used by participants, rather than to benchmark model performance or establish generalisable accuracy claims. This section presents an exploratory evaluation of the KG-enhanced LLM using a small, curated question set ($n = 10$) with expert-validated answers (Table 3). The questions primarily address PHP and related endocrine features, with one question drawn from epilepsy. They span both fact-based knowledge (e.g., hormone function, genetic mutations, clinical features) and reasoning-oriented tasks (e.g., interpreting subtype characteristics and management considerations). Each question was submitted directly to the KG-enhanced LLM without additional context, instructions, or prompt engineering.

Table 3. Dataset of clinical questions with expert-validated answers and corresponding KG-enhanced LLM responses.

ID	Question	Expert Validated Answer	KG-Enhanced LLM Response
1	What is the primary function of PTH?	Regulation of calcium and phosphate concentrations	The primary function of PTH is to increase calcium levels in the blood.
2	Vitamin D deficiency is associated with which type of PHP?	PHP type 2	Vitamin D deficiency is associated with PHP type 2.
3	Which PHP subtypes have a preserved cAMP response?	Pseudo-PHP and PHP type 2	The PHP subtypes that have a preserved cAMP response are Pseudo-PHP and PHP type 2.
4	Which genetic mutation is associated with PHP?	GNAS mutation	The genetic mutation associated with PHP is the GNAS gene mutation.
5	AHO is typically associated with which types of PHP?	PHP types 1a, 1c, Pseudo-PHP, and sometimes type 2	AHO is typically associated with the following types of PHP: PHP type 1a, Pseudo-PHP, PHP type 1b, PHP type 1c, and PHP type 2.
6	What clinical features are characteristic of AHO?	Short stature, obesity, round face, brachydactyly, subcutaneous ossifications	The clinical features characteristic of AHO include: short stature, round facies, brachydactyly, obesity, and osteoma cutis.
7	How is Chvostek's sign elicited on physical exam?	Tap the cheek about 2 cm in front of the ear, over the facial nerve	Chvostek's sign is elicited on physical exam by performing facial nerve percussion.

Table 3. *Cont.*

ID	Question	Expert Validated Answer	KG-Enhanced LLM Response
8	When managing PHP, what should be prevented or avoided?	Hypercalciuria	When managing PHP, the following should be prevented or avoided: hypercalciuria and severe symptomatic hypocalcemia.
9	Which parameters should be monitored during long-term management of PHP?	Serum calcium, phosphate, PTH, renal function, urinary calcium	During long-term management of PHP, the following parameters should be monitored: serum PTH, serum calcium, serum phosphate, and 24 h urine calcium. Additionally, PHP treatment monitoring is recommended.
10	Which age groups are associated with febrile seizures?	Infants and children	Febrile seizures are seen in infants and children.

The curated questions and corresponding model responses were evaluated using the RAGAS framework, and the resulting scores (Table 4) confirm functional viability and surface obvious failure modes of the KG–LLM integration, rather than providing definitive measures of model performance or robustness.

Table 4. RAGAS Metric Results.

Question	Faithfulness	Context Precision	Context Recall	Answer Relevancy	Answer Correctness
Q1	1.00	1.00	1.00	1.00	1.00
Q2	1.00	0.70	1.00	0.88	1.00
Q3	1.00	0.92	1.00	0.99	0.85
Q4	0.00	0.20	0.00	1.00	1.00
Q5	0.67	0.33	0.80	0.89	1.00
Q6	0.60	1.00	0.67	0.97	1.00
Q7	0.50	1.00	1.00	0.99	1.00
Q8	0.67	1.00	1.00	0.99	1.00
Q9	0.67	0.70	0.60	1.00	1.00
Q10	1.00	0.75	1.00	0.97	1.00
Average	0.71	0.76	0.81	0.97	0.98

Across the dataset, answer relevancy (mean = 0.97) and answer correctness (mean = 0.98) were consistently high, indicating strong alignment between model outputs, clinical questions, and expert-validated answers. Context recall (0.81) was also relatively strong, suggesting that the system generally retrieved relevant entities from the KG. Although context precision (0.76) and faithfulness (0.71) exhibited greater variability, the overall mean RAGAS score (0.85) reflects acceptable retrieval and response alignment at the aggregate level.

To further characterise retrieval balance, an F1 score was calculated using context precision and recall:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Based on an average precision of 0.76 and recall of 0.81, the resulting F1 score was

$$F1 = 2 \times \frac{0.76 \times 0.81}{0.76 + 0.81} \approx 0.79.$$

This F1 score indicates a reasonably balanced level of retrieval accuracy and completeness.

However, aggregate performance masks important failure modes that become apparent at the question level (Table 4). Lower context precision indicates that, while most retrieved information was relevant, irrelevant details were occasionally included. Such “noise” can increase cognitive load, requiring clinicians to filter extraneous information to

extract clinically useful insights. Similarly, reduced faithfulness increases the risk of hallucinations, which can undermine clinical safety. This divergence highlights a fundamental limitation of automated RAG metrics: clinically correct answers may still be insufficiently grounded in retrieved evidence, while minor abstraction or paraphrasing may be penalised as faithfulness errors despite remaining clinically valid.

Several question-level examples illustrate these limitations. In Q4, the model achieved perfect answer correctness (1.00) despite scoring 0.00 for both faithfulness and context recall and only 0.20 for context precision. Although the response was factually correct, it was entirely ungrounded in the retrieved knowledge graph, representing a breakdown of the intended safeguard against unsupported answers. In Q5, reduced faithfulness (0.67) and low context precision (0.33) reflected partial hallucination, where correct information was combined with errors, including inappropriate inclusion of PHP type 1B and omission of subtype occurrence details. In contrast, Q7 demonstrated a different failure pattern: faithfulness declined to 0.50 despite strong performance across other metrics because a clinically acceptable paraphrase (“facial nerve percussion”) omitted specific procedural detail (“tap 2 cm anterior to the ear”). While less concerning than outright hallucination, this example highlights the sensitivity of faithfulness metrics to phrasing.

Overall, average performance indicates that the KG-enhanced LLM is sufficiently accurate and relevant for participants to meaningfully engage, although sensitivity to phrasing and occasional retrieval noise may pose minor challenges.

4.2. Clinical Evaluation Results

4.2.1. Pre-Interview Survey

Ten participants were included in the study. The information gathered from the pre-interview survey, including general demographic characteristics, is summarised in Table 5. Additional items assessed in the pre-interview survey included participants’ use of search or AI tools, perceived trust in these tools, perceived helpfulness of responses, and frequency of use, as summarised in Table 6. Furthermore, the types of tools used are shown in Figure 4, along with the primary purposes for which they were applied in Figure 5.

Table 5. Demographic Information of Participants. GOPD refers to General Outpatient Department. GP refers to General Practitioner. N/A refers to Not Applicable, used when participants are not currently working.

ID	Age	Practice	Department	Role	Years of Experience
DR01	20–29	Private	Internal Medicine	GP	1–3 years
DR02	30–39	Private	GOPD	GP	1–3 years
DR03	20–29	N/A	–	GP	1–3 years
DR04	20–29	Private	Paediatrics	GP	1–3 years
DR05	20–29	Private	GOPD	GP	1–3 years
DR06	20–29	Government	Emergency	GP	1–3 years
DR07	20–29	Private	Paediatrics	GP	1–3 years
DR08	20–29	N/A	–	GP	1–3 years
DR09	30–39	Government	GOPD	GP	3–5 years
DR10	20–29	Government	Internal Medicine	Junior Resident	1–3 years

Participants completed a survey to capture their prior experience with AI and search tools (Table 6). Responses were collected using structured scales:

- Use of Tools: Whether participants had used AI or search tools before (Yes/No).
- Helpfulness: How useful they perceived the tools to be (Not Helpful, Neutral, Helpful, Very Helpful).
- Trust: Level of trust in the tool’s outputs (Never, Rarely, Somewhat, Mostly, Completely).

- Frequency of Use: How often they used the tools (Never, Rarely, Monthly, Weekly, Daily).

Table 6. Pre-Interview Survey Responses on AI and Search Tool Usage.

ID	Use of Tools	Helpfulness	Trust	Frequency of Use
DR01	Yes	Neutral	Somewhat	Weekly
DR02	Yes	Helpful	Somewhat	Weekly
DR03	Yes	Very Helpful	Mostly	Weekly
DR04	Yes	Helpful	Somewhat	Rarely
DR05	Yes	Very Helpful	Mostly	Weekly
DR06	Yes	Very Helpful	Mostly	Rarely
DR07	Yes	Helpful	Mostly	Daily
DR08	No	—	—	—
DR09	Yes	Helpful	Mostly	Weekly
DR10	Yes	Very Helpful	Mostly	Daily

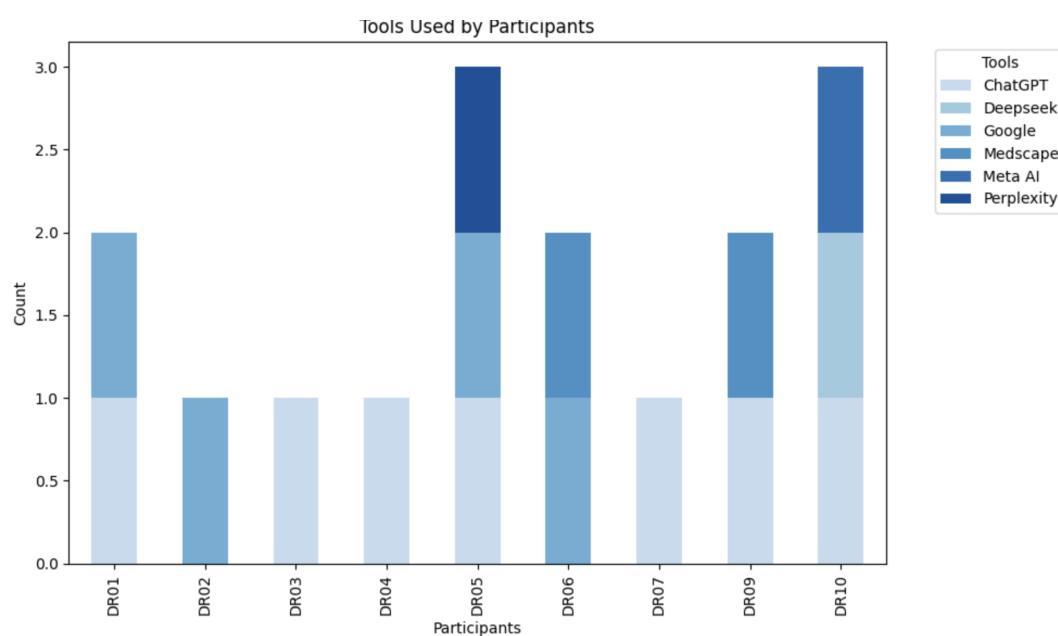


Figure 4. AI and search tools used by individual participants. Each bar represents a participant (n = 10), showing which tools they reported using, including ChatGPT (7 participants), Google (4), Medscape (2), Perplexity (1), and others.

4.2.2. Diagnostic Assessment & Adherence

Diagnostic performance was assessed descriptively, focusing on whether participants identified the relevant condition or included appropriate differentials, including subtypes where applicable. Participants were not required to reach a definitive diagnosis; instead, they reported their differential diagnoses and conclusions when faced with uncertainty. These actionable conclusions, recorded in Tables 7 and 8, provide insights into how the KG-enhanced LLM may support clinical reasoning.

Adherence patterns were analysed for Cases 1 and 2, in which all ten participants engaged with the KG-enhanced LLM. Case 3 was excluded, as only three participants used AI in this case; the remaining eight opted not to engage with the system, citing high diagnostic confidence and satisfaction with their own clinical reasoning. Across Cases 1 and 2, this resulted in 20 instances in which participants attempted to reach a diagnosis using the KG-enhanced LLM (Table 9).

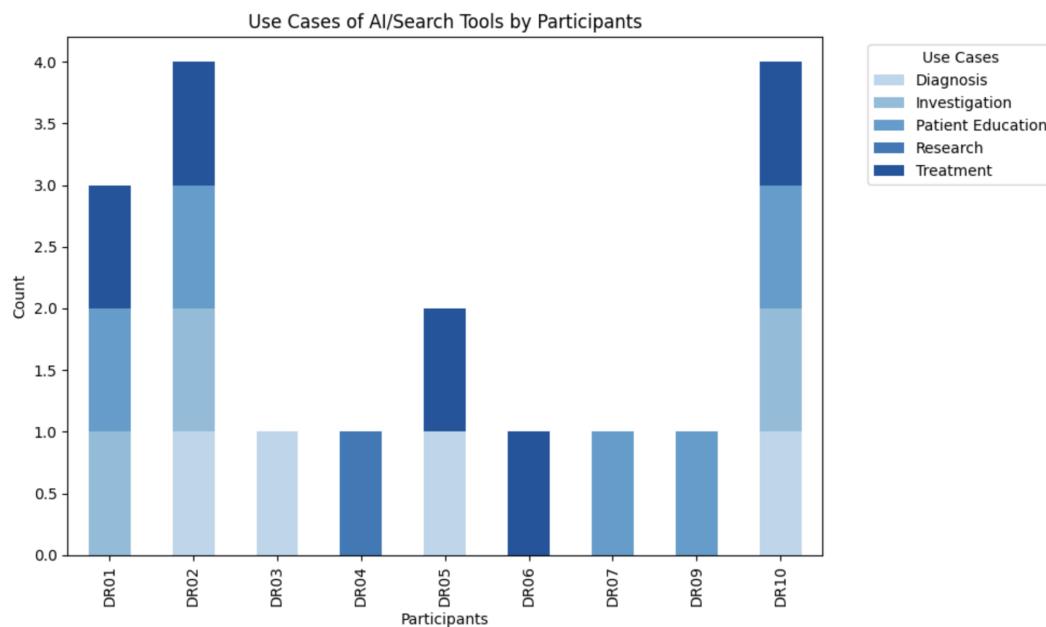


Figure 5. Use cases of AI and search tools reported by individual participants. Each bar represents a participant ($n = 10$), showing the clinical purposes for which they reported using AI or search tools, including diagnosis (4 participants), Patient education (4 participants) and others.

Table 7. Case 1: Diagnostic Assessment and KG Adherence (PHP Type 1A).

ID	No AI Conclusion	AI Conclusion	Adherence
DR01	Manage Symptoms	PHP, no subtype	Partial
DR02	Referral	PHP type 1A	Full
DR03	Referral	PHP, no subtype	Partial
DR04	Secondary Hypoparathyroidism	PHP, no subtype	Partial
DR05	Hypocalcaemia	Inconclusive	No
DR06	Referral	PHP, no subtype	Partial
DR07	Inconclusive	PHP, no subtype	Partial
DR08	Hypoparathyroidism	PHP, no subtype	Partial
DR09	Referral	PHP, no subtype	Partial
DR10	PHP, no subtype	PHP, no subtype	Partial

Table 8. Case 2: Diagnostic Assessment and KG Adherence (Pseudo-PHP).

ID	No AI Conclusion	AI Conclusion	Adherence
DR01	Inconclusive	Inconclusive	No
DR02	Referral	Pseudo-PHP	Full
DR03	Referral	PHP, no subtype	Partial
DR04	Cushing's Syndrome	Pseudo-PHP	Full
DR05	Referral	Inconclusive	No
DR06	Referral	PHP, no subtype	Partial
DR07	Referral	Pseudo-PHP	Full
DR08	Referral	Pseudo-PHP	Full
DR09	Referral	Pseudo-PHP	Full
DR10	Referral	PHP, no subtype	Partial

In three out of twenty instances (15%), the KG-enhanced LLM was unable to provide relevant or usable diagnostic support. In each of these cases, the system explicitly communicated its limitation with messages such as “This information is not contained in my knowledge base.” A similar outcome occurred when three participants prompted the KG-enhanced LLM in the control case. Importantly, in none of the twenty instances did the KG-enhanced LLM provide misleading, incorrect, or out-of-scope information.

Table 9. Summary of Participant adherence to KG-enhanced LLM suggestions.

Case	Full Adherence	Partial Adherence	No Adherence	Total Instances	Adherence Rate
Case 1	5	3	2	10	80%
Case 2	1	8	1	10	90%
Total	6	11	3	20	85%

4.2.3. Time/Efficiency Analysis

To examine whether the KG-enhanced LLM influenced diagnostic efficiency during the interview, time-to-diagnosis was analysed as a secondary outcome. Average and median completion times were calculated for Cases 1 and 2. Case 3 was excluded from this analysis, as it served as a control condition and involved minimal AI usage. Time differences were computed at the participant level by subtracting AI-assisted completion times from non-AI times (Table 10), where positive values indicate faster task completion with AI assistance.

Across nine participants, excluding missing data from DR05, the average time difference was 135.5 s, with a median of 141 s and a standard deviation (SD) of 81.5 s (Table 11). These values indicate that, on average, participants took longer to complete tasks without AI, suggesting that the use of AI improved efficiency in this context.

A paired-samples t-test was conducted to assess whether the observed improvement in time efficiency with AI was statistically significant. This test was appropriate because each participant completed tasks under both AI and non-AI conditions, allowing for direct within-subject comparisons. For each participant, the average completion times for Case 1 and Case 2 were calculated separately for the AI and non-AI conditions. The test statistic was calculated using the following formula:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

where \bar{d} is the mean of the paired differences, s_d is the standard deviation of these differences, and n is the number of valid paired observations. The degrees of freedom for the test were calculated as the number of valid paired comparisons minus one:

$$df = n - 1 = 9 - 1 = 8,$$

reflecting the nine participants with complete paired data (excluding the missing AI value for DR05). The mean difference was $\bar{d} = 135.5$ s with a standard deviation of $s_d = 81.5$ s. This resulted in a test statistic of

$$t(8) = \frac{135.5}{81.5 / \sqrt{9}} \approx 4.99,$$

with a corresponding probability under the null hypothesis of $p = 0.001$, indicating that the observed reduction in completion time with AI was very unlikely to have occurred by chance.

The effect size was calculated using Cohen's d_z for paired samples:

$$d_z = \frac{\bar{d}}{s_d} = \frac{135.5}{81.5} \approx 1.66,$$

indicating a very large effect of AI on time efficiency. A 95% confidence interval (CI) for the mean difference was computed as

$$CI_{95\%} = \bar{d} \pm t_{0.025, 8} \cdot \frac{s_d}{\sqrt{n}} = 135.5 \pm 2.306 \cdot 27.17 \approx [73.8, 197.2].$$

To check the robustness of these findings, we conducted a Wilcoxon signed-rank test including all valid paired observations (excluding the missing value for DR05). The Wilcoxon test statistic was calculated as:

$$W = 0.0, \quad p = 0.0039,$$

confirming that the reduction in completion time with AI was also statistically significant under this non-parametric sensitivity check.

Table 10. Participant-level Results (No AI vs. AI).

ID	Case 1 No AI	Case 1 AI	Case 2 No AI	Case 2 AI	No AI Avg	AI Avg	Difference
DR01	470 s	354 s	480 s	478 s	475.0 s	416.0 s	59.0 s
DR02	600 s	507 s	900 s	502 s	750.0 s	504.5 s	245.5 s
DR03	300 s	130 s	392 s	557 s	346.0 s	343.5 s	2.5 s
DR04	375 s	215 s	552 s	290 s	463.5 s	252.5 s	211.0 s
DR05	214 s	—	355 s	—	284.5 s	—	—
DR06	206 s	244 s	476 s	218 s	341.0 s	231.0 s	110.0 s
DR07	80 s	180 s	620 s	180 s	350.0 s	180.0 s	170.0 s
DR08	80 s	97 s	487 s	330 s	283.5 s	213.5 s	70.0 s
DR09	278 s	269 s	568 s	295 s	423.0 s	282.0 s	141.0 s
DR10	180 s	225 s	572 s	106 s	376.0 s	165.5 s	210.5 s

Table 11. Summary Statistics Across Participants excluding missing data (DR05).

Statistic	Case 1 No AI	Case 1 AI	Case 2 No AI	Case 2 AI	Difference
Average	285.4 s	246.8 s	560.8 s	328.4 s	135.5 s
Median	278.0 s	225.0 s	552.0 s	295.0 s	141.0 s
SD	174.3 s	123.4 s	144.3 s	154.5 s	81.5 s

4.3. Post-Interview Results

The post-interview survey collected changes in confidence and participants' evaluations of the AI-assisted tool.

4.3.1. Diagnostic Confidence

Changes in diagnostic confidence before and after AI use were analysed to assess whether the KG-enhanced LLM provided meaningful support during clinical reasoning. Confidence was captured using a five-point likert-type scale and converted into numeric change (Table 12) scores to enable comparison across cases and participants (Table 13).

Case 1: Most participants began with moderate confidence, with scores around "Neutral" (3 on the Likert scale). After using the KG-enhanced LLM, 7 out of 10 participants (70%) reported an increase in confidence, 2 participants (20%) experienced no change, and 1 participant (10%) reported a slight decrease. This indicates that the AI tool generally supported clinicians' confidence when dealing with this case.

Case 2: Participants initially showed mixed confidence, ranging from "Low" to "High" (2–4). Following interaction with the KG-enhanced LLM, 6 participants (60%) reported an increase in confidence, 3 participants (30%) had no change, and 1 participant (10%) experienced a decrease. The tool appeared to provide moderate benefit, particularly for those who initially had lower confidence.

Case 3: Most participants were already confident before AI assistance, with initial ratings mostly "High" or "Very High" (4–5). Eight participants (80%) who opted not to engage with the KG-enhanced LLM reported no change in confidence, while 2 participants (20%), despite receiving no helpful information from the AI, experienced a slight increase as they felt validated in their own knowledge.

Table 12. Mapping of qualitative Likert-scale metrics to numeric values before and after KG-enhanced LLM interaction, allowing for easier interpretation of changes.

Qualitative Metric	Numeric Value	Qualitative Change	Numeric Value
Very Low	1	Decreased Significantly	−2
Low	2	Decreased Slightly	−1
Neutral	3	No Change	0
High	4	Increased Slightly	+1
Very High	5	Increased Significantly	+2

Table 13. Participant-level change across cases before and after KG-enhanced LLM interaction.

ID	Case 1 Before	Case 1 After	Case 2 Before	Case 2 After	Case 3 Before	Case 3 After
DR01	3	+1	3	+1	3	0
DR02	2	+2	2	+2	4	0
DR03	2	+2	2	+2	5	0
DR04	4	+1	4	+2	3	0
DR05	4	−2	4	−2	4	0
DR06	4	+2	4	+2	4	0
DR07	2	+1	2	+1	2	0
DR08	3	+1	3	+1	3	0
DR09	3	+1	3	+1	5	0
DR10	4	+1	2	+1	4	+1

4.3.2. Participant Feedback

Participant feedback was collected using 5-point Likert scales for five aspects of the KG-enhanced LLM (Table 14). The scales were defined as follows:

- Usability: Overall ease of using the KG-enhanced LLM, rated on a 5-point scale from Very difficult to Very easy (Very difficult, Difficult, Neutral, Easy, Very easy).
- Relevance of Responses: How often the AI responses provided relevant and useful information to aid clinical diagnosis, rated on a 5-point scale from Never to Always (Never, Rarely, Sometimes, Most of the time, Always).
- Accuracy: How accurate participants perceived the AI tool's suggestions in supporting diagnosis, rated on a 5-point scale from Very inaccurate to Very accurate (Very inaccurate, Somewhat inaccurate, Neutral, Accurate, Very accurate).
- Efficiency: Whether the AI tool improved the diagnostic process, rated on a 5-point scale from Slowed significantly to Improved significantly (Slowed significantly, Slowed slightly, No impact, Improved slightly, Improved significantly).
- Trust: The degree to which participants trusted the AI tool's suggestions and guidance for clinical decision-making, rated on a 5-point scale from Never to Completely (Never, Rarely, Somewhat, Mostly, Completely).

Summary:

- Usability: Most participants found the tool easy or very easy to use (7/10).
- Relevance of Responses: Six out of ten reported the responses as mostly or always relevant.
- Accuracy: Seven out of ten rated the tool as accurate or very accurate.
- Efficiency: Eight out of ten felt the tool slightly or significantly improved their diagnostic process.
- Trust: Six out of ten reported mostly or complete trust in the AI suggestions.

Table 14. Participant feedback on KG-enhanced LLM (Likert scale responses).

ID	Usability	Relevance of Responses	Accuracy	Efficiency	Trust
DR01	Easy	Most of the time	Neutral	Improved slightly	Mostly
DR02	Easy	Most of the time	Accurate	Improved significantly	Completely
DR03	Very easy	Always	Very accurate	Improved significantly	Completely
DR04	Neutral	Always	Very accurate	Slowed slightly	Mostly
DR05	Difficult	Rarely	Somewhat inaccurate	Slowed slightly	Somewhat
DR06	Very easy	Always	Very accurate	Improved slightly	Mostly
DR07	Neutral	Most of the time	Accurate	Improved slightly	Mostly
DR08	Neutral	Sometimes	Neutral	Improved slightly	Somewhat
DR09	Neutral	Sometimes	Neutral	Improved slightly	Somewhat
DR10	Difficult	Sometimes	Accurate	Improved slightly	Somewhat

5. Discussion

PHP is an endocrine disorder that is frequently misdiagnosed, particularly as epilepsy in some regions, because of overlapping neurological symptoms. Diagnostic complexity is compounded by the condition's rarity and the varied presentations of its subtypes.

5.1. Case Study 1 (Typical Presentation: PHP Type 1A)

In Case 1 (Table 7), which included clearly abnormal laboratory values, diagnoses in the non-AI round were broad and inconsistent. Suggestions included secondary hypoparathyroidism, Cushing's syndrome, adrenal insufficiency, and in some instances, no definitive diagnosis. Most participants ultimately opted to refer the case. Four participants initially suspected Cushing's syndrome due to truncal obesity associated with AHO; one also mentioned osteodystrophy. One participant referred the patient to a neurologist because of seizure-like features. Only one participant, a junior resident with recent endocrinology experience, correctly diagnosed pseudohypoparathyroidism, but did not specify a subtype.

With AI support, participants engaged more effectively with the case. Six reached a diagnosis of PHP without specifying the subtype, whereas one participant asked targeted questions and used the KG-enhanced LLM to identify the correct subtype. Three participants remained inconclusive despite AI assistance.

5.2. Case Study 2 (Atypical Presentation: Pseudo-PHP)

Without AI assistance, most participants struggled with Case 2 (Table 8). The first participant was unable to make a diagnosis, even with the KG-enhanced LLM, because the information provided was deemed unhelpful. Among the subsequent eight participants, several requested genetic testing but were unfamiliar with how to interpret the results. Four offered incorrect differentials, Down syndrome, DiGeorge syndrome, autism, or Cushing's syndrome, based on physical features and observed behavioral abnormalities. Most opted to refer to a specialist (neurologist or paediatrician). However, only the last participant, the junior resident, correctly referred to an endocrinologist.

With AI support, participants navigated the atypical presentation more effectively. Four participants correctly diagnosed pseudo-PHP, while two identified PHP without specifying a subtype. Four participants remained inconclusive or misdiagnosed the case. Among the nine out of ten participants who completed Case 1 before Case 2, those who identified PHP without specifying a subtype found Case 2 confusing because its presentation resembled Case 1 but with normal laboratory values. This prompted some participants to ask more targeted questions, which in some cases led to identifying the specific subtype, something they had not achieved in Case 1.

5.3. Case Study 3 (Control: Severe Malaria)

In Case 3, most participants relied on their own clinical judgement. Without AI, only four out of ten participants diagnosed malaria, and just one correctly specified severe malaria. The majority suspected alternative diagnoses such as sepsis, pneumonia, or metabolic acidosis. Given the clinical findings, including a high white blood cell count and an abnormal anion gap, sepsis or metabolic acidosis were not unreasonable. Similarly, pneumonia was suspected because of respiratory distress.

Seven out of ten participants expressed high confidence in their clinical assessment, noting relief at handling a familiar condition and choosing not to use the KG-enhanced LLM. The three who engaged with the AI found it unhelpful because the clinical features of this case were not represented in the KG.

5.4. Participant Reflections and Feedback

Most participants noted that the interview felt more like an exam, which made it easier to forget routine questions and omit standard investigations, e.g. failing to request a malaria parasite test in Case 3, which is routine in the region). Participants also emphasised that, in clinical practice in this region, making a precise diagnosis is not always the immediate priority. Instead, the focus is often on managing presenting complaints and clinical abnormalities before referring the patient to a specialist. Most participants appreciated that the system provided information only when asked, rather than offering unsolicited suggestions; this gave users a sense of control and reduced the risk of information overload. However, some participants were concerned that the KG-enhanced LLM often provided too much information to be practical in a clinical setting, reflecting the context precision score (0.76) and its implications for cognitive load. “Responses need to be more specific,” one participant emphasised. This design also placed an additional burden on clinicians, who had to know what to ask and how to ask it. Participants uncertain about next steps or terminology sometimes failed to uncover helpful leads, not because the AI lacked the answer, but because the prompt did not provide practical guidance.

Many participants emphasised the importance of transparency and trustworthiness in medical AI tools. Participants suggested validation mechanisms, such as tracking accuracy rates or implementing clinical trials, before full adoption. Others proposed domain-specific restrictions or safeguards, although some could not identify specific requirements, possibly due to unfamiliarity with AI regulations or limitations. Several participants expressed concern that clinicians might gradually trust AI tools more than their own diagnostic reasoning, potentially leading to a decline in critical thinking over time. This aligns with existing concerns in the literature about automation bias in medical decision-making.

Participant feedback indicated a generally positive reception toward AI-assisted tools, particularly as supportive instruments rather than primary diagnostic tools. Most participants would consider incorporating such a tool into their workflow, especially for rare or complex cases, but not in emergencies or routine scenarios where clinical judgment is more straightforward. Some participants feel the tool had greater value as an educational aid than as a primary diagnostic tool.

A relevant question that emerged, although not explicitly asked, seems crucial: Do clinicians fear misdiagnosing patients more because of their own judgment or because of over-reliance on an AI tool? This distinction could provide deeper insight into how responsibility, confidence, and trust interact in clinical decision-making. Future evaluations should incorporate such a question.

5.5. Limitations

This study has several limitations, the most significant being the small number of participants and the limited evaluation dataset. Only ten clinical questions were used, and the study involved ten participants. These constraints limit generalisability, and the results should be interpreted as exploratory rather than conclusive. While the qualitative insights were rich and meaningful, they do not support statistically significant conclusions.

Several limitations relate to the KG-enhanced LLM framework and its supporting knowledge graph. The KG was manually curated from trusted sources such as textbooks and clinical guidelines. While reliable, this introduced biases affecting completeness and scope. Selection bias occurred because only well-documented information about PHP and epilepsy was included, whereas newer or less established findings were excluded, limiting representation of the full clinical picture. Expert bias also influenced content, reflecting the perspectives and priorities of its creator. For example, the KG may emphasise certain causes, such as low calcium in PHP-related seizures, while overlooking alternative explanations, including neurological conditions or atypical presentations. This narrowing of diagnostic paths could reduce the likelihood of surfacing rare but important differentials. The restricted scope of the KG further limited its utility: with only two diseases represented, it could not provide detailed differentials or address broader diagnostic queries, and symptoms were represented simply as present or absent, without considering severity, frequency, or triggers. This simplification could reinforce textbook-style reasoning and, in complex cases, bias participants toward familiar presentations.

Other limitations relate to the clinical evaluation itself. Participant experience levels were skewed: nine had 1–3 years of clinical experience, and one had 3–5 years. All were general practitioners except for one junior resident. While this offered consistency in perspective, it reduced diversity in clinical backgrounds and may have influenced how the KG-enhanced LLM was used and evaluated. The simulated diagnostic setting was also artificial; several participants noted that sessions felt more like exams than natural clinical interactions, which could have affected prompting style, communication confidence, and willingness to explore the system. Additionally, the order of case presentation was not fully balanced, potentially introducing fatigue, priming effects, or familiarity bias. Because participants completed the same cases in both the AI-assisted and non-AI phases, observed differences may reflect learning effects rather than the AI's impact. Longer-term impacts, such as whether repeated use would influence diagnostic confidence, accuracy, or cognitive bias, were not assessed due to time constraints and study design.

Usability was also a constraint. The system relied entirely on participants to frame questions, offering no guidance when queries were vague or unclear. Consequently, participants sometimes failed to obtain useful responses even when the relevant information was present in the KG.

Despite these limitations, the study provides valuable exploratory insights into how clinicians interact with KG-enhanced LLMs and highlights practical challenges and considerations for future clinical evaluations, including participant diversity, naturalistic settings, and system usability.

5.6. Future Improvements

Future development of the KG-enhanced LLM should begin with targeted technical improvements. Expanding the KG's size and scope would enable coverage of a broader range of diseases, rare presentations, and atypical symptom patterns, supporting cases characterised by high diagnostic uncertainty. However, such expansion must be accompanied by careful information prioritisation and relevance filtering to prevent excessive or poorly structured information from overwhelming clinicians. Metric inconsistencies observed in

the current evaluation highlight the need for additional safeguards. Future work could incorporate abstention or fallback mechanisms to explicitly signal uncertainty when context is insufficient, alongside systematic failure-case analyses. Introducing a re-ranking stage to prioritise clinically relevant entities and relationships could further improve retrieval quality, interpretability, and overall clinical reliability.

An expanded KG would also enable larger, clinician-led trials to more rigorously examine diagnostic reasoning in atypical and rare cases under increased information volume. Rather than treating AI responses as standalone answers, future studies should explicitly examine how retrieved information supports different stages of clinical reasoning, such as hypothesis generation, confirmation, exclusion of alternatives, and confidence calibration. Capturing these interaction patterns would clarify when KG-enhanced LLMs provide meaningful support and when they are bypassed due to low perceived need for assistance.

Such trials would additionally allow for the assessment of whether KG-enhanced LLMs meaningfully reduce diagnostic effort or instead introduce additional cognitive load. Iterative refinement of system prompts and interaction design could then be used to optimise response length, tone, usability, and cognitive load reduction, supporting real-world clinical adoption. Future evaluations should also systematically assess interaction efficiency under realistic time and workload constraints, including prompting behaviour such as the number, length, and specificity of prompts, as well as the proportion of clinically useful information returned. Prompt-tuning strategies, context-aware interactions, and proactive detection of vague or incomplete inputs—with suggested clarifying follow-up questions—may further reduce interaction friction and minimise trial-and-error during clinical reasoning.

To reliably isolate AI-specific effects in future studies, a cross-over design with randomised case order would be required. Where cases are repeated, inclusion of a washout period would help minimise learning effects from prior exposure, ensuring that observed improvements can be more confidently attributed to AI assistance. Future evaluations should also expand the number and diversity of out-of-KG cases and explicitly require participant interaction with the system to reliably assess hallucination rates, abstention behaviour, and handling of unsupported or unfamiliar queries. In parallel, bias mitigation strategies, including adversarial datasets designed to expose overfitting or spurious correlations, will be essential to prevent misleading or overly narrow diagnostic suggestions. Evaluation frameworks should further account for cultural and clinical practice variations, as well as ethical considerations surrounding responsibility, trust, and accountability in clinical decision-making.

Long-term integration goals focus on embedding the KG-enhanced LLM in ways that align with clinicians' workflows and decision-making practices. Multi-centre validation across diverse clinical settings will be important to understand how clinicians adopt the system selectively, identify workflow-specific constraints, and evaluate usability in real-world contexts. Robust APIs compliant with interoperability standards such as HL7 FHIR would support smooth data exchange and context-aware decision support, reducing friction for clinicians. Piloting the system in targeted clinical settings will provide insight into clinician interaction patterns, including frequency of use, number and type of prompts issued per case, and how AI input is balanced with professional judgment. Collecting continuous feedback from clinicians will guide iterative refinements, ensuring that the system supports efficient decision-making, maintains trust, and integrates safely into routine practice without adding undue cognitive burden.

6. Conclusions

This study demonstrates that a KG-enhanced LLM can effectively support clinicians in complex or rare cases, particularly those with atypical presentations, while offering limited benefit in routine or familiar scenarios. Rather than replacing clinical judgment, the system functioned as an assistive tool, supporting reasoning, providing second-opinion insights, and acting as an educational aid. Clinicians were more likely to engage with AI support when diagnostic confidence was low, especially in rare endocrine cases such as PHP. Notably, AI responses that explicitly acknowledged uncertainty increased clinician trust, suggesting that transparency and humility are important design features for medical AI.

The findings also show that the system's usefulness depends heavily on clinician interaction, requiring users to recognise uncertainty and articulate effective queries. Improving the model's ability to detect ambiguity and proactively guide users through prompt suggestions may enhance its clinical value. Participants consistently preferred to rely on their own judgment in familiar cases, indicating that KG-enhanced LLMs are most beneficial in situations characterised by diagnostic uncertainty rather than routine decision-making.

Safeguards remain essential to prevent overreliance, preserve clinical reasoning, and reduce the risk of misdiagnosis, reinforcing the need for AI to remain an assistive, not authoritative, component of clinical workflows. Strengthening the underlying KG and validating performance across larger and more diverse datasets will be critical for ensuring reliability. Transparent feedback and performance indicators may further support trust and responsible adoption.

With careful design and validation, KG-enhanced LLMs can serve as effective collaborative tools in clinical decision-making, enhancing diagnostic confidence while keeping final responsibility with clinicians.

Author Contributions: Conceptualization, F.S. and J.W.; methodology, F.S.; software, F.S.; validation, F.S.; formal analysis, F.S.; investigation, F.S.; data curation, F.S.; writing—original draft preparation, F.S.; writing—review and editing, F.S. and J.W.; visualization, F.S.; supervision, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was approved by the Ethical Committee of the University of West London.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original contributions presented in this study are included in the article. The backend, frontend, and knowledge graph dump are publicly available on GitHub at <https://github.com/fateeS88/MKG-Pseudohypoparathyroidism.git> (accessed on 20 December 2025). Further enquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AHO	Albright Hereditary Osteodystrophy
CDSS	Clinical Decision Support Systems
EHRs	Electronic Health Records
EMRs	Electronic Medical Records
GDPR	General Data Protection & Regulation
GOPD	General Outpatient Departments
GP	General Practitioner
KG	Knowledge Graph

LLM	Large Language Model
MKG	Medical Knowledge Graphs
MMR	Maximal Marginal Ranking
NICE	National Institute for Health and Care Excellence
PHP	Pseudohypoparathyroidism
PLM	Pre-trained Language Model
PTH	Parathyroid Hormone
Q&A	Question-Answering
RAG	Retrieval-Augmented Generation
RAGAS	Retrieval-Augmented Generation Assessment System
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RR	Re-Ranking
SD	Standard Deviation
SFT	Supervised Fine-Tuning
SQuAD	Stanford Question Answering Dataset
UHC	Universal Health Coverage
UMLS	Unified Medical Language System
USMLE	United States Medical Licensing Examination

Appendix A. Prompt Template

The following prompt template was used to generate responses from the KG-enhanced LLM. Retrieved knowledge graph context was injected verbatim, and the model was instructed to ground its response strictly in the provided information.

Knowledge Graph Context:

{Retrieved triples formatted as subject–relation–object statements}

User Question:

{User-provided clinical query}

Instruction:

Please answer using only the information from the knowledge graph context above. If the information required to answer the question is not available, respond with “*I couldn’t find this information in my knowledge base.*”

References

1. Medford-Davis, L.; Malani, R. The Physician Shortage Isn’t Going Anywhere. 2024. Available online: <https://www.mckinsey.com/industries/healthcare/our-insights/the-physician-shortage-isnt-going-anywhere> (accessed on 17 December 2025).
2. Sutton, R.; Pincock, D.; Baumgart, D.; Sadowski, D.; Fedorak, R.; Kroeker, K. An overview of clinical decision support systems: Benefits, risks, and strategies for success. *npj Digit. Med.* **2020**, *3*, 17. [[CrossRef](#)]
3. Nazi, Z.A.; Peng, W. Large language models in healthcare and medical domain: A review. *Informatics* **2024**, *11*, 57. [[CrossRef](#)]
4. Zou, S.; He, J. Large Language Models in Healthcare: A Review. In Proceedings of the 2023 7th International Symposium on Computer Science and Intelligent Control (ISCSIC), Nanjing, China, 27–29 October 2023. [[CrossRef](#)]
5. Jovanović, M.; Campbell, M. Connecting AI: Merging Large Language Models and Knowledge Graph. *Computer* **2023**, *56*, 103–108. [[CrossRef](#)]
6. Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; Wu, X. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 3580–3599. [[CrossRef](#)]
7. Kim, E.; Shrestha, M.; Foty, R.; DeLay, T.; Seyfert-Margolis, V. Structured Extraction of Real World Medical Knowledge using LLMs for Summarization and Search. *arXiv* **2024**, arXiv:2412.15256. [[CrossRef](#)]
8. Abu-Salih, B. Domain-specific knowledge graphs: A survey. *J. Netw. Comput. Appl.* **2021**, *185*, 103076. [[CrossRef](#)]
9. Cai, L.; Yu, C.; Kang, Y.; Fu, Y.; Zhang, H.; Zhao, Y. Practices, opportunities and challenges in the fusion of knowledge graphs and large language models. *Front. Comput. Sci.* **2025**, *7*, 1590632. [[CrossRef](#)]

10. Jia, M.; Duan, J.; Song, Y.; Wang, J. mediKAL: Integrating Knowledge Graphs as Assistants of LLMs for Enhanced Clinical Diagnosis on EMRs. In Proceedings of the 31st International Conference on Computational Linguistics, Abu Dhabi, United Arab Emirates, 19–24 January 2025; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 9278–9298. Available online: <https://aclanthology.org/2025.coling-main.624/> (accessed on 28 November 2025).

11. Park, C.; Lee, H.; Lee, S.; Jeong, O. Synergistic Joint Model of Knowledge Graph and LLM for Enhancing XAI-Based Clinical Decision Support Systems. *Mathematics* **2025**, *13*, 949. [CrossRef]

12. Newton, N.; Bamgboje-Ayodele, A.; Forsyth, R.; Tariq, A.; Baysari, M.T. How Are Clinicians’ Acceptance and Use of Clinical Decision Support Systems Evaluated Over Time? A Systematic Review. In *MEDINFO 2023—The Future Is Accessible*; Studies in Health Technology and Informatics; IOS Press: Amsterdam, The Netherlands, 2024. [CrossRef]

13. Westerbeek, L.; Ploegmakers, K.J.; de Bruijn, G.-J.; Linn, A.J.; van Weert, J.C.M.; Daams, J.G.; van der Velde, N.; van Weert, H.C.; Abu-Hanna, A.; Medlock, S. Barriers and facilitators influencing medication-related CDSS acceptance according to clinicians: A systematic review. *Int. J. Med. Inform.* **2021**, *152*, 104506. [CrossRef]

14. Xie, Q.; Schenck, E.J.; Yang, H.S.; Chen, Y.; Peng, Y.; Wang, F. Faithful AI in Medicine: A Systematic Review with Large Language Models and Beyond. *medRxiv* **2023**. [CrossRef]

15. Yang, R.; Liu, H.; Marrese-Taylor, E.; Zeng, Q.; Ke, Y.; Li, W.; Cheng, L.; Chen, Q.; Caverlee, J.; Matsuo, Y.; et al. KG-Rank: Enhancing Large Language Models for Medical QA with Knowledge Graphs and Ranking Techniques. In Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, Bangkok, Thailand, 16 August 2024; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024. [CrossRef]

16. Yang, R.; Tan, T.R.; Lu, W.; Thirunavukarasu, A.J.; Shu, D.; Liu, N. Large Language Models in Health Care: Development, Applications, and Challenges. *Health Care Sci.* **2023**, *2*, 255–263. [CrossRef]

17. Balogh, E.P.; Miller, B.T.; Ball, J.R. Overview of Diagnostic Error in Health Care. In *Improving Diagnosis in Health Care*; National Academies Press: Washington, DC, USA, 2019. Available online: <https://www.ncbi.nlm.nih.gov/books/NBK338594/> (accessed on 17 December 2025).

18. Sadat, M.; Zhou, Z.; Lange, L.; Araki, J.; Gundroo, A.; Wang, B.; Menon, R.; Parvez, M.R.; Feng, Z. DELUCIONQA: Detecting Hallucinations in Domain-specific Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 822–835. Available online: <https://aclanthology.org/anthology-files/anthology-files/pdf/findings/2023.findings-emnlp.59.pdf> (accessed on 19 September 2025).

19. Pan, S.; Zheng, Y.; Liu, Y. Integrating Graphs With Large Language Models: Methods and Prospects. *IEEE Intell. Syst.* **2024**, *39*, 64–68. [CrossRef]

20. Ibrahim, N.; Aboulela, S.; Ibrahim, A.; Kashef, R. A survey on augmenting knowledge graphs (KGs) with large language models (LLMs): Models, evaluation metrics, benchmarks, and challenges. *Discover AI* **2024**, *4*, 76. [CrossRef]

21. Yadav, D.; Para, H.; Selvakumar, P. Unleashing the Power of Large Language Model, Textual Embeddings, and Knowledge Graphs for Advanced Information Retrieval. In Proceedings of the 2023 International Conference on Electrical, Computer and Energy Technologies (ICECET), Cape Town, South Africa, 16–17 November 2023. [CrossRef]

22. Zhang, W.; Tian, Y.; Meng, X.; Wang, M.; Du, J. Knowledge Graph-Infused Fine-Tuning for Structured Reasoning in Large Language Models. *arXiv* **2025**, arXiv:2508.14427. [CrossRef]

23. Gao, Y.; Li, R.; Croxford, E.; Caskey, J.; Patterson, B.W.; Churpek, M.; Miller, T.; Dligach, D.; Afshar, M. Leveraging Medical Knowledge Graphs Into Large Language Models for Diagnosis Prediction: Design and Application Study. *JMIR AI* **2025**, *4*, e58670. [CrossRef] [PubMed]

24. Croxford, E.; Gao, Y.; Patterson, B.; To, D.; Tesch, S.; Dligach, D.; Mayampurath, A.; Churpek, M.M.; Afshar, M. Development of a human evaluation framework and correlation with automated metrics for natural language generation of medical diagnoses (Preprint). *medRxiv* **2024**. [CrossRef]

25. Shamszare, H.; Choudhury, A. Clinicians’ Perceptions of Artificial Intelligence: Focus on Workload, Risk, Trust, Clinical Decision Making, and Clinical Integration. *Healthcare* **2023**, *11*, 2308. [CrossRef]

26. Bastepe, M.; Gensure, R.C. Hypoparathyroidism and Pseudohypoparathyroidism. [Updated 2024 May 8]. In *Endotext [Internet]*; Feingold, K.R., Adler, R.A., Ahmed, S.F., Anawalt, B., Blackman, M.R., Chrousos, G., Corpas, E., de Herder, W.W., Dhatariya, K., Dungan, K., et al., Eds.; MDText.com, Inc.: South Dartmouth, MA, USA 2000. Available online: <https://www.ncbi.nlm.nih.gov/sites/books/NBK279165/> (accessed on 17 December 2025).

27. Najim, M.S.; Mohammed, A.; Awad, M.; Ahmed, O. Pseudohypoparathyroidism presenting with seizures: A case report and literature review. *Intractable Rare Dis. Res.* **2020**, *9*, 166–170. [CrossRef]

28. Rodriguez, J.A.; Roa, A.A.; Leonso-Bravo, A.-A.; Khatiwada, P.; Eckardt, P.; Lemos-Ramirez, J. A Case of Plasmodium falciparum Malaria Treated with Artesunate in a 55-Year-Old Woman on Return to Florida from a Visit to Ghana. *Am. J. Case Rep.* **2020**, *21*, e926097. [CrossRef] [PubMed]

29. Dai, L.-Z.; Lin, C.; Lei, R.; Zhang, Y.; Ma, H. A Case of Pseudohypoparathyroidism Misdiagnosed as Idiopathic Epilepsy for 5 Years: Clinical Analysis and Follow-up Outcomes. *J. Int. Med. Res.* **2023**, *51*. [CrossRef]

30. OpenAI. OpenAI Platform. 2025. Available online: <https://platform.openai.com/docs/models/compare> (accessed on 17 December 2025).
31. Dhanakotti, K. RAGAS for RAG in LLMs: A Comprehensive Guide to Evaluation Metrics. Medium. 2024. Available online: <https://dkaarthick.medium.com/ragas-for-rag-in-llms-a-comprehensive-guide-to-evaluation-metrics-3aca142d6e38> (accessed on 17 December 2025).
32. Ghaffar, F.; Furtado, N.M.; Ali, I.; Burns, C. Diagnostic Decision-Making Variability Between Novice and Expert Optometrists for Glaucoma: Comparative Analysis to Inform AI System Design. *JMIR Med. Inform.* **2025**, *13*, e63109. [CrossRef] [PubMed]
33. Intersoft Consulting. General Data Protection Regulation (GDPR). 2018. Available online: <https://gdpr-info.eu/issues/consent/> (accessed on 17 December 2025).
34. NICE. 1 Recommendations. In *Artificial Intelligence (AI)-Derived Software to Help Clinical Decision Making in Stroke*; NICE: London, UK, 2024. Available online: <https://www.nice.org.uk/guidance/dg57/chapter/1-Recommendations> (accessed on 17 December 2025).
35. NICE. 3 Committee Discussion. In *Artificial Intelligence (AI)-Derived Software to Help Clinical Decision Making in Stroke*; NICE: London, UK, 2024. Available online: <https://www.nice.org.uk/guidance/dg57/chapter/3-Committee-discussion> (accessed on 17 December 2025).
36. EY Global. G7 AI Principles and Code of Conduct. 2023. Available online: https://www.ey.com/en_gl/insights/ai/g7-ai-principles-and-code-of-conduct (accessed on 17 December 2025).
37. Neha, F.; Bhati, D.; Shukla, D.K. Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review. *AI* **2025**, *6*, 226. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.