LARGE LANGUAGE MODELS IN CLINICAL ADVICE: DIRECT GENERATION AND RETRIEVAL AUGMENTED GENERATION VS EXPERT ADVICE

Iblal Rakha¹ and Noorhan Abbas²

¹Oxford University Hospitals NHS Foundation Trust, Oxford, OX3 9DU, UK

² University of Leeds, Woodhouse, Leeds, LS2 9JT, UK

ABSTRACT

The NHS faces mounting pressures, resulting in workforce attrition and growing care backlogs. Pharmacy services, critical for ensuring medication safety and effectiveness, are often overlooked in digital innovation efforts. This pilot study investigates the potential of Large Language Models (LLMs) to alleviate pharmacy pressures by answering clinical pharmaceutical queries. Two retrieval techniques were evaluated: Vanilla Retrieval Augmented Generation (RAG) and Graph RAG, supported by an external knowledge source designed specifically for this study. ChatGPT 40 without retrieval served as a control. Quantitative and qualitative evaluations were conducted, including expert human assessments for response accuracy, relevance, and safety. Results demonstrated that LLMs can generate high-quality responses. In expert evaluations, Vanilla RAG outperformed other models and even human reference answers for accuracy and risk. Graph RAG revealed challenges related to retrieval accuracy. Despite the promise of LLMs, hallucinations and the ambiguity around LLM evaluations in healthcare remain key barriers to clinical deployment. This pilot study underscores the importance of robust evaluation frameworks to ensure the safe integration of LLMs into clinical workflows. However, regulatory bodies have yet to catch up with the rapid pace of LLM development. Guidelines are urgently needed to address the issues of transparency, explainability, data protection, and validation, to facilitate the safe and effective deployment of LLMs in clinical practice.

Keywords

Large Language Model Evaluation, Retrieval Augmented Generation, Clinical Question Answering, Knowledge Graphs, Healthcare Artificial Intelligence.

1. INTRODUCTION

The NHS has faced mounting pressures in recent years, driven by workforce attrition, secondary care backlogs, and high vacancy rates, all of which have placed significant strain on services [1]. Despite their critical role in ensuring the safe and effective use of medications, hospital pharmacy services are often overlooked in discussions on healthcare innovation and resource allocation. Pharmacists are an integral part of the multidisciplinary teams involved in patient care, providing clinical advice and ensuring appropriate supply of medications. Hospital pharmacies also provide a medicines information (MI) service, offering expert advice to healthcare professionals (HCPs)

David C. Wyld et al. (Eds): Signal, CDKP, AdNLP, MLT, SAI – 2025 pp. 39-57, 2025. CS & IT - CSCP 2025

DOI: 10.5121/csit.2025.150904

and patients. Innes et al. [2] demonstrated the value of a MI service: HCPs found that 74% of cases referred to MI resulted in positive patient outcomes and risks to patients were averted in 71% of cases.

The NHS has committed to using innovative solutions to find efficiencies. Artificial Intelligence (AI) is thought to present opportunities to improve patient care and workforce productivity [3]. An LLM could answer clinical queries or be used for clinical decision support. Koman et al. [4] conducted a study on physicians' attitudes to using a chatbot for medication management. They found that physicians are willing but stressed the need for validation. Morath et al. [5] investigated the use of ChatGPT 3.5 to answer medication queries and found that the majority of answers were wrong, with a quarter having a high likelihood of patient harm.

Currently, there are several obstacles to overcome before the use of LLMs, in a clinical setting, is feasible. These include wrong and/or inconsistent answers, out of date references, and lack of trust in LLMs answers [4, 5]. Parry [6] explored these challenges in an article on the role of AI chatbots in pharmacy, highlighting how LLMs can 'hallucinate', producing coherent but inaccurate answers. The author also stressed the importance of human validation and the use of 'guardrails', such as prompt engineering, to improve the accuracy of responses.

1.1. Retrieval Augmented Generation

One method to reduce hallucinations is to combine an LLM with a retrieval model utilising an external knowledge source (EKS), known as retrieval augmented generation (RAG) [7]. Ke et al. [8] argued for the use of retrieval augmented LLMs (RA-LLM) in healthcare due to the evidencebased approach of clinical tasks. Specialty guidelines can be used to form a knowledge source for an LLM to leverage when generating answers. During retrieval, relevant texts in the EKS are returned using cosine similarity - this method is referred to throughout this study as 'Vanilla RAG'. The quality of a RA-LLM output relies on the results from the EKS [7, 9]. Therefore, pre-retrieval strategies and/or post-retrieval enhancements (PRE) methods are crucial for optimising performance.

Graph RAG is a novel approach that defines an EKS in a logical, hierarchical structure of entities (as nodes), their attributes, and their relationships (as edges) as a knowledge graph (KG) [10]. These KGs are used as the EKS for a RA-LLM system. Retrieval in Graph RAG returns relevant graph elements, leveraging relational knowledge to find connected information. This can result in more accurate and holistic retrieval, thereby generating more precise answers [11].

1.2. Evaluating LLMs in Healthcare

Evaluating LLM outputs is critical for healthcare applications, especially when used as a clinical decision support tool [12]. Feldman et al. [13] outlined how the lack of trust in the use of AI is a significant barrier to the adoption of AI technologies in healthcare. Therefore, thorough evaluations are crucial in establishing that trust, and for safe and effective use in a clinical setting [4, 14]. That said, validating LLM outputs is challenging. Ambiguity or nuances in an LLM answer may lend itself to a few interpretations, hence, robust evaluations require specialist knowledge from subject matter experts (SME). Moreover, there are no industry standard metrics for evaluating LLM-generated outputs [15]. Adding to the task complexity, not all methods or metrics are suitable for use in all natural language generation (NLG) applications [16].

There have been several studies assessing the performance of LLMs in a clinical setting. Morath et al. [5] evaluated the answers from ChatGPT 3.5 for medication related questions. Six senior hospital pharmacists scored the answers by classifying them into categories based on the contents

of the answer (correct, incomplete or false) and the risk to patient safety (no risk, low risk or high risk). Benary et al. [17] evaluated LLMs' abilities in recommending personalised treatment plans for advanced cancer versus recommendations from a human expert. LLM precision, recall and F1 scores were recorded where reference treatment options existed.

Hager et al. [18] investigated LLM diagnostic capabilities when given real case studies and calculated the percentages of correct diagnoses. Ke et al. [8] evaluated RA-LLMs for preoperative instructions given realistic case studies. Researchers judged 'correctness' of RA-LLM outputs based on criteria from established guidelines; Cohen's H and Chi-squared were used to compare RA-LLM performance against junior doctors. Sblendorio et al. [19] investigated the feasibility of using LLMs in clinical practice by scoring LLMs on a Likert scale in several domains. The average Likert scores were computed per domain and used to navigate a decision tree to classify LLMs as 'recommended', 'usable with caution' or 'unusable'.

From the literature, there is no consensus on methods for evaluating LLM outputs in healthcare. Although, there are recurring themes: a panel of experts score LLM outputs against certain criteria [5, 19, 20] or the comparison of LLM answers against reference materials [8, 17, 18]. Lee et al. [21] proposed best practices for human evaluation of NLG, advocating for the use of Likert scales alongside well defined, domain specific criteria. Similarly, Tam et al. [20], in their framework for human evaluation of LLMs in healthcare, outlined principles for evaluation. These mirror the NHS UK MI service's guidance on assessing responses provided by medicines advice services [22]. The NHS UK MI guidance offers valuable insights into evaluating answers provided by pharmacy professionals to medication queries, making it particularly well-suited for assessing LLM-generated responses also.

1.3. Evaluating LLMs in Non-Healthcare Domains

Traditional metrics, such as BLEU and ROUGE, measure the n-gram overlap between generated outputs and reference texts. However, these metrics do not correlate well with human judgement and fail to capture semantic similarities between texts [15]. Embedding-based metrics, such as cosine similarity and BERTScore, have shown a stronger correlation with human judgement [15, 23, 24].

Cosine similarity provides a broad measure of semantic alignment between two texts by comparing their embedding vectors. While useful for identifying overall similarity, it does not take granular relationships into account, such as token-level alignment. BERTScore builds upon cosine similarity making it better suited for tasks requiring a more nuanced evaluation [24]. These metrics rely on embeddings, and the choice of embeddings plays a pivotal role in their effectiveness. In specialised domains like healthcare, where the language is often highly technical, context specific, and contains jargon [25], it is essential to use embeddings from models trained on specialist corpora [23]. Question answering natural language inference (QNLI) is another technique used to evaluate the quality of responses. QNLI is designed to assess whether the context provided (in this case, the generated answers) can be considered an entailment of the original question [26].

Study Design and Objectives. This pilot study evaluates LLM responses to clinical questions using two retrieval techniques: Vanilla RAG and Graph RAG. This includes developing a bespoke KG tailored to local protocols. Additionally, this study includes ChatGPT 40 without retrieval as a control. ChatGPT was selected as it ranks among the top models used by HCPs [27].

Both quantitative and qualitative evaluation methods are employed to compare the performance of RAG architectures and ChatGPT 40 against human reference answers. Expert human evaluation, serves as a critical validation step for all responses, including the human reference answers, ensuring mitigation of potential biases. The evaluation framework adopts best practices from Lee et al. [21] and Tam et al. [20] and aligns with NHS UK MI criteria [22], focusing on relevance, accuracy, and safety. This approach integrates domain specific standards with rigorous assessment methodologies, providing a comprehensive evaluation of LLM-generated responses in a clinical context.

2. METHODOLOGY

2.1. Data Collection and Preprocessing

Oxford University Hospitals (OUH) Pharmacy offers a MI service, answering medicines related questions from HCPs and patients. Questions to the service are recorded and assigned a level ranging from one to three for increasing complexity. Answers are provided by specialist pharmacy professionals and often written in retrospect where answers are given orally. This database was selected to extract 'gold standard' question-answer pairs. Analysis of the MI database guided the stratified random sampling during data collection. A small representative sample size of 53 questions were initially selected, due to the time constraints for SME evaluation. The data were screened and excluded against the following criteria:

- The question or answer contained any patient identifiable information
- The question or answer required referring to external sensitive information to answer the question (e.g. a discharge summary or patient notes).

The answers provided by the MI service were standardised for tone, grammar and perspective (third person) to facilitate a blinded and unbiased evaluation. ChatGPT 40 was utilised to rewrite the answers with instructions to preserve the original content without introducing additional information or altering clinical meaning. To mitigate the risk of unintended content expansion, the standardised outputs were reviewed and validated by an MI pharmacist, ensuring fidelity to the original responses. This process helped control for potential variability in documentation practices, making the comparisons between LLM-generated and human reference answers more reliable.

2.2. Knowledge Graph

OUH also provided medicines information leaflets (MILs), which are local guidelines to facilitate safe and cost-effective clinical practice and medication management. These MILs were used alongside summary of medicinal product characteristics (SMPCs) to form the content of the KG. Neo4j graph database [28] was chosen to store the data. LangChain [29] was also used as it offers tools to interface with the Neo4j database for storing, organising, querying and answer generation.

The LangChain LLMGraphTransformer class transforms documents to a graph based document, using an LLM to extract entities and relationships [29]. The granularity of the entity-relationship extraction differed depending on the size of the document. Recursive document splitting resulted in a more detailed knowledge graph at the expense of greater processing times and cost. Moreover, the lack of output reproducibility across extraction runs did not instil confidence in the quality of the KG. There are numerous methods to ameliorate this such as specifying allowed nodes and relationships, setting a temperature of 0, and prompt engineering. However, employing

these methods still produced inconsistent KGs. Therefore, the KG was manually defined as in a content-centric manner as opposed to entity-centric.

The texts from MILs were extracted and processed to extract all medications mentioned in the MIL, using a drug named entity recognition (DNER) library [30]. Using the list of extracted active ingredients obtained from the DNER exercise, all associated SMPCs were collected. One active ingredient could have multiple SMPCs associated with it due to differing strengths or formulations etc. SMPCs were chunked based on content, guided by the sections defined by the electronic medicines compendium.

The final KG contained 24 MILs, 107 active ingredients and 879 medicinal formulations (SMPCs). The architecture and content of the KG was validated by two pharmacists. Thereafter embeddings were calculated and stored for texts that were stored in MIL and SMPC section nodes. This was done to facilitate the document retrieval using cosine similarities from the same database, an overview of this process is shown in Fig. 1.

2.3. Knowledge Graph RAG

2.3.1. Pre-Retrieval Process

Cypher is the query language for the Neo4j graph database [28]. To optimise retrieval, the original question was generalised using ChatGPT 4o-mini. This process simplified the question to its core components ensuring alignment with the database schema. The refined query was then passed to another LLM chain, which utilised dynamic k-shot prompting to identify relevant examples. These examples guided another LLM, ChatGPT 4o, in formulating the Cypher query and retrieval. Due to the limited context window for answer generation, the number of retrieved results was limited to two for MIL texts and five for SMPC sections.



Fig. 1. Diagrammatic overview of content centric chunking method and subsequent node-relationship storage in the Neo4j graph database. The diagram on the left is the Neo4j representation of MIL node (blue), with an active ingredient (green), SMPC sections (pink) and the relationships between them.

2.3.2. Post-Retrieval Process

The raw results retrieved from the KG were subjected to a post-retrieval pruning process. This included summarisation of the context such that only the most relevant and accurate information was retained, utilising ChatGPT 40. Unprocessed results can otherwise contribute to hallucinations [7, 9]. Additionally, this process was essential for managing token usage. Without pruning, the volume of retrieved data could exceed token limits, particularly for complex queries requiring broader or deeper retrieval [31, 32].

The pruned results were integrated with the original question using carefully designed templates tailored to query complexity. Prompting strategies played a pivotal role in meeting the different requirements for question levels. Level one templates were designed for simple, fact-based questions, focusing on delivering concise responses. Level two templates addressed moderately complex queries by incorporating additional context. Level three templates were the most sophisticated, ensuring rigorous referencing. The model used to generate the final responses was ChatGPT 40.

2.4. Vanilla RAG

The MILs and SMPC sections were indexed using LangChain's vector indexing capabilities. The embeddings for both indexes were pre-computed and stored in the graph database. SMPC section nodes were indexed with a similarity score threshold of 0.7 and k=5; these parameters were selected to ensure high precision while maintaining sufficient recall. The MIL nodes were indexed with a lower similarity score threshold of 0.6 and k=2, reflecting the inherently larger and more diverse text structure of MILs. The retrieved results were then subject further processing. The overall process closely followed the approach outlined in 2.3, including pre-retrieval question refinement, post-retrieval pruning, and answer generation. The key difference was in the retrieval method. The same templates for generating the final answers were used as in Graph RAG.

2.5. Quantitative Evaluation

The dataset for evaluation was finalised, comprising of four answers for each corresponding question sourced from the MI database. These included responses generated using the knowledge Graph RAG approach (2.3), the Vanilla RAG approach (2.4), and ChatGPT 40 without any context retrieval, relying solely on the generation templates. Standardised human referenced answers were also included.

The human reference answers were used as the standard against which the LLM-generated answers were evaluated. Cosine similarities and BERTScores were calculated using Clinical BERT (CBERT) embeddings. QNLI scores were also calculated using the open-sourced Electra QNLI and PubMedBERT-QNLI (PQNLI) transformer models. The latter is fine-tuned for biomedical applications therefore was selected for its ability to handle the specific language and nuances of medical texts [33]. The human reference answers were also evaluated for entailment against the original questions to provide a benchmark, enabling direct comparison with the LLM-generated outputs.

2.6. Expert Human Evaluation

All answers were evaluated by a panel of four senior pharmacists. The pharmacists recruited for this study had specialisations across several clinical fields: surgery and nutrition, maternity, obstetrics and gynaecology, and cardiology. While their collective expertise was broad, it is

important to consider that the questions themselves span multiple disciplines, or none. The distribution of clinical topics could influence scoring consistency, especially for questions outside the direct expertise of the evaluators. However, in this study the evaluators aligned with the clinical subject matter in 17 out of 24 questions.

The same exercise was also performed with ChatGPT 40 to draw comparison between expert human and LLM evaluation. To streamline the evaluation process, the number of questions was reduced to 24, striking a balance between evaluation practicality and statistical representation. The selection included 12 level one queries, 10 level two queries, and two level three queries, reflecting the proportional complexity of queries typically encountered in clinical practice. This reduction resulted in 96 individual responses for assessment.

The evaluation criteria were guided by the NHS UK MI service [22], which supports NHS trusts in the safe use of medicines and provides evidence-based guidance. Their published criteria for grading responses to clinical enquiries served as the framework for developing the evaluation questionnaire. Pharmacists were asked to rate their agreement with specific statements using a 7-point Likert scale, ranging from "strongly agree" to "strongly disagree". The following formed the basis for the evaluation: To what extent do you agree or disagree with the following statements:

- 1. The answer provided adequately addresses the proposed question (assessing comprehension of the question).
- 2. The answer provided is accurate and correct.
- 3. The answer provided does not pose a risk to patient safety.
- 4. The answer provided was produced by a human.

3. RESULTS

3.1. Embeddings Based Results

Appendix A is a table outlining the summary of statistical results for embedding-based metrics. All LLM-generated answers demonstrated high cosine similarity with the human reference answer, with means exceeding 0.90 and medians close to or surpassing their respective means. CBERT scores, while lower than cosine similarity, were consistent across the three LLM-generated responses. The lower CBERT scores suggest a greater sensitivity to token-level nuances and minor discrepancies in phrasing. Notably, Vanilla RAG exhibited a marginally larger interquartile range for both cosine similarity and CBERT scores, indicating slightly greater variability.

Graph RAG and ChatGPT 40 responses exhibited near identical cosine and CBERT similarity with human reference answers, with differences being negligible. In contrast, Vanilla RAG responses showed the greatest variability in comparison to the human reference answers. The CBERT F1 scores revealed a notable difference for Vanilla RAG when compared to the human reference, particularly for more complex questions.



3.2. ONLI Results

Fig. 1. Violin plots illustrating the density and spread of Electra QNLI and PQNLI scores for all responses.

A QNLI score of 1 indicates a strong entailment between the answer and the original question. For Electra QNLI, the scores and distributions of the three LLM-generated responses were high, with ChatGPT 4o, Vanilla RAG and Graph RAG showing particularly strong performance. Interestingly, human reference answers exhibited lower scores with a wider spread, with many responses clustering near 0, shown in the first plot of Fig. 2. Whereas for PQNLI, all responses displayed medians very close to 1, reflecting strong entailment overall. Most scores were concentrated near 1, with no scores approaching 0, aligning with expectations. These results suggest that the fine-tuned PQNLI scores are particularly well suited for evaluating this task compared to Electra QNLI.

The human reference answers had the widest distribution of scores, reflecting variability in their entailment with the original questions. ChatGPT 40 followed closely, displaying slightly greater variability compared to the other models, as shown by the broader range in its violin plot (Fig. 2). In contrast, Graph RAG showed the least variability, with scores tightly clustered around the median, demonstrating consistency in entailment. Vanilla RAG showed a moderate spread but retained high median values similar to the other models.

3.3. Expert Human Evaluation Results

A boxplot of the results from the expert human evaluation are presented in Fig. 3. The median scores obtained from ChatGPT 40 performing the same exercise are marked as black diamonds. Each model was assessed based on the four evaluation criteria outlined in 2.6. Given the small sample size and the likelihood of variability in scoring due to subjective differences among evaluators, it was assumed that the results would not be normally distributed. Therefore, the Kruskal-Wallis test [34], a non-parametric alternative to ANOVA, was chosen to determine any statistically significant differences between the answers, without assuming normality or equal variances.



Fig. 2. Boxplots of expert human evaluation results. The top plot displays scores across three question complexity levels, grouped by evaluation criteria. The bottom plot shows overall scores across all complexity levels, grouped by the same criteria.

The null hypothesis was: "responses to clinical questions, whether produced by MI specialist professionals or LLM models, are equally scored for relevance, accuracy, and risk to patient safety, with no significant difference in their median scores". When analysing the quality of responses across different levels of question complexity, the Kruskal-Wallis test demonstrated that the null hypothesis could be rejected (at p<0.05) for accuracy and risk to patient safety in level one questions, and for relevance in level two questions. The small sample size of level three questions limited the ability to draw meaningful conclusions.

When considering the null hypothesis irrespective of question complexity, the test revealed statistically significant differences between the answers for relevance, accuracy, and risk to patient safety. Thereafter a post-hoc Dunn's test was applied. Dunn's test is designed to identify pairwise differences following a significant Kruskal-Wallis test [34]. While Dunn's test does allow for multiple comparisons, no explicit correction method was applied. Instead, statistical significance was interpreted in the context of distributional trends observed in the Fig. 3 and interquartile ranges. The significant Kruskal-Wallis and post-hoc Dunn's test are shown in Tables 1-6, significant p values (<0.05) are marked with an asterisk.

The null hypothesis was not rejected for the criterion: "the answer was produced by a human", indicating that the evaluators were unable to consistently distinguish between human-generated

and LLM-generated responses. This suggests that the blinding process was successful, minimising the risk of evaluator bias.

Table 1. Question Level 1, Criteria: Accuracy (Kruskal-Wallis H statistic = 16.35, p = 0.001)

	Human	ChatGPT 4o	Graph RAG	Vanilla RAG
Human	-	1.000000	1.000000	0.001199*
ChatGPT 40	1.000000	-	1.000000	0.043776*
Graph RAG	1.000000	1.000000	-	0.009109*
Vanilla RAG	0.001199*	0.043776*	0.009109*	-

Table 2. Question Level 1, Criteria: Risk (Kruskal-Wallis H statistic = 15.34, p = 0.002)

	Human	ChatGPT 40	Graph RAG	Vanilla RAG	
Human	-	0.579631	1.000000	0.001114*	
ChatGPT 40	0.579631	-	1.000000	0.227248	
Graph RAG	1.000000	1.000000	-	0.024459*	
Vanilla RAG	0.001114*	0.227248	0.024459*	-	

Table 3. Question Level 2, Criteria: Relevance (Kruskal-Wallis H statistic = 10.72, p = 0.013)

	Human	ChatGPT 40	Graph RAG	Vanilla RAG	
Human	-	0.011069*).011069* 1.000000		
ChatGPT 40	0.011069*	-	0.132440	1.000000	
Graph RAG	1.000000	0.132440	-	1.000000	
Vanilla RAG	0.464342	1.000000	1.000000	-	

Table 4. All Questions, Criteria: Relevance (Kruskal-Wallis H statistic = 12.55, p = 0.006)

	Human	ChatGPT 40	Graph RAG	Vanilla RAG
Human	-	0.018524*	1.000000	0.114398
ChatGPT 40	lo 0.018524* -		0.063495	1.000000
Graph RAG	1.000000	0.063495	-	0.313174
Vanilla RAG	0.114398	1.000000	0.313174	-

Table 5. All Questions, Criteria: Accuracy (Kruskal-Wallis H stat	tistic = $10.80, p = 0.013$)
---	-------------------------------

	Human	ChatGPT 40	Graph RAG	Vanilla RAG
Human	-	0.395476	1.000000	0.092160
ChatGPT 40	0.395476	-	0.198422	1.000000
Graph RAG	1.000000	0.198422	-	0.039647*
Vanilla RAG	0.092160	1.000000	0.039647*	-

Table 6. All Questions, Criteria: Risk (Kruskal-Wallis H statistic = 13.80, p = 0.003)

	Human	ChatGPT 40	Graph RAG	Vanilla RAG
Human	-	0.297117	1.000000	0.027486*
ChatGPT 40	0.297117	-	0.144930	1.000000
Graph RAG	1.000000	0.144930	-	0.010644*
Vanilla RAG	0.027486*	1.000000	0.010644*	-

4. DISCUSSION

4.1. Exploratory Analysis of Answers

Qualitative analysis revealed that the majority of the LLM-generated answers were closely aligned in content with the human reference answers. However, there were instances of deviations. One third of ChatGPT 4o's answers contained a hallucination, but were confined to the references section of the response and did not affect the main content. In contrast, both Vanilla RAG and Graph RAG produced responses free from hallucinations. Two Graph RAG answers included inaccurate information due to inaccurate context retrieval.

The answers were further assessed for their ability to promote the safe and effective use of medications (such as including signs or symptoms for monitoring or medication counselling). Vanilla RAG delivered quality counselling advice in 20 out of 24 responses, compared to 15 out of 24 for ChatGPT 40, Graph RAG and human reference responses.

Fig. 4 presents an example of responses to address a query regarding the interaction between naproxen and rivaroxaban. All responses highlighted a theoretical interaction between naproxen and rivaroxaban, recommending against their concurrent use. However, Graph RAG's response captured a particular nuance, acknowledging that while the interaction is theoretical, there is a lack of supporting evidence. This variation in responses can be attributed to differences in context retrieval mechanisms (see section 2.3 and 2.4). Both retrieval methods successfully returned the appropriate interaction sections (and the lack of evidence thereof). However, Vanilla RAG's retrieved results mostly consisted of loosely related MIL texts, whereas Graph RAG's results contained only the relevant SMPC interaction sections. Subsequently, during the post-retrieval process, that nuance in the Vanilla RAG's response was lost.

4.1.1. Graph RAG Retrieval Inaccuracies

Having said that, the Graph RAG architecture struggled with ambiguous questions, as seen in one example: "...how to take their isosorbide mononitrate for their angina". The query does not specify the formulation, leading the Cypher query to retrieve dosage information for all formulations in the database but was restricted to five results (see 2.3). This arbitrary cut-off resulted in information being retrieved primarily for modified-release capsules, which was then included in the generated response. As a result, the answer addressed a formulation that was not explicitly stated in the question, introducing potential safety risks. In contrast, Vanilla RAG's predefined top-k retrieval mechanism delivered a smaller, more focused subset of results, avoiding this issue. A similar limitation in Graph RAG was observed in the second instance of an inaccurate response.

As LLMs continue to advance, their ability to extract structured entity-relationship mappings may improve, allowing for more replicable and entity centric KG construction. This would complement the graph constraints outlined 2.2, ultimately reducing the volume of retrieved text and mitigating the need for the arbitrary retrieval limit used in this study. Additionally, LLMs fine-tuned specifically for entity extraction could provide more reliable and domain-specific KG construction, further enhancing retrieval precision.

Question: "Patient called helpline - was discharged from hospital yesterday (Monday 30th Jan 23) on rivaroxaban but forgot to mention they take naproxen. Concerned about interactions so would like to know if there is an interaction and if the naproxen should be stopped (no dose taken today) and when the naproxen should be restarted. Patient takes naproxen 500mg when required for pain in her hip. Rivaroxaban was prescribed only for 7 days post COVID."

Human	Nonsteroidal anti-inflammatory drugs can increase the risk of bleeding when taken with rivaroxaban or any other anticoagulant (a drug that prevents blood clots). It is advisable to avoid taking these two medicines together. The patient may resume taking naproxen, if necessary for pain control, a few days after completing the course of rivaroxaban. Paracetamol may be used for pain control if needed while taking rivaroxaban.
ChatGPT 4o	There is an interaction between rivaroxaban and naproxen, as both increase the risk of bleeding. Since the patient is on rivaroxaban for only 7 days, it would be safest to avoid taking naproxen during this time. The patient can resume naproxen after completing the rivaroxaban course. If pain relief is needed, they should consider alternatives like paracetamol and consult a healthcare professional for further advice.
Vanilla RAG	There is a known interaction between rivaroxaban and naproxen. Both medications increase the risk of bleeding, particularly gastrointestinal bleeding, when taken together. Therefore, it is generally advised to avoid using naproxen while on rivaroxaban unless under direct medical supervision. In this case, since the patient is currently taking rivaroxaban for a short course (7 days post-COVID), it would be safest to stop naproxen for the duration of the rivaroxaban treatment. The patient should not take naproxen until they have completed the course of rivaroxaban and consulted their GP or healthcare provider about when it is safe to restart naproxen, especially considering their pain management needs. In the meantime, the patient may want to discuss alternative pain relief options with their GP, such as paracetamol, which does not carry the same bleeding risk as NSAIDs like naproxen. It is important that the patient monitors for any signs of bleeding (e.g., unusual bruising, blood in stools or urine) and seeks medical attention if these occur.
Graph RAG	Rivaroxaban and naproxen can be used together, but caution is advised due to the potential for increased bleeding risk. While no clinically relevant prolongation of bleeding time was observed in studies, individual responses can vary, and some patients may experience a more pronounced effect. Given that rivaroxaban was prescribed for a short duration of 7 days, it may be prudent to avoid taking naproxen during this period unless absolutely necessary. If pain management is required, consider discussing alternative pain relief options with your GP or healthcare provider. If naproxen is essential, it should be used with caution, and any signs of increased bleeding, such as unusual bruising or bleeding, should be reported to a healthcare professional immediately. Once the course of rivaroxaban is completed, naproxen can be resumed as needed, but always under the guidance of a healthcare professional.

Fig. 3. Example of a level one question and corresponding answers.

In safety critical domains, granular knowledge graphs would be preferable, as missing noderelationships may result in incomplete or even dangerous clinical guidance. Recursive document splitting, which enables greater granularity, could help mitigate this risk. However, a major drawback of this approach is the high cost associated with using closed source LLMs for entity extraction, in addition to the rigorous validation required by SMEs. In contrast, content centric KGs, like the one used in this study, are simpler to validate due to their fewer nodes and relationships. An alternative approach to mitigate retrieval failures is query disambiguation, incorporating user in the loop clarification mechanisms. This method could potentially

circumvent the resource intensive and highly laborious requirements of KG construction, offering a more dynamic and adaptable retrieval strategy.

Beyond managing token sizes, it is also important to consider whether the inherent complexity of the task contributes to increased token consumption. Interestingly, ChatGPT 4o's context window was exceeded even when the token count was below the maximum allowable threshold. More intricate questions, which involve multiple facets or layers of reasoning, may require a greater token allocation for both context and response generation. While direct references discussing this phenomenon are scarce, it aligns with broader observations about LLM behaviour in handling complex inputs and outputs [31, 32].

4.2. Comparing Human Reference Answers and LLM Generated Answers

4.2.1. Embedding Metrics

All models achieved consistently high scores, with no significant differences in performance between them (Appendix A). This may be due to the fact that all answers, including the human reference, were generated or standardised using ChatGPT 40 (see 2.1). The semantic, tonal, and structural similarities resulting from this standardisation likely contributed to the uniformly high scores.

However, cosine similarity, while effective for measuring general semantic alignment, does not capture small nuances or differences in the answers which may hold clinical significance [23]. CBERT was employed as it is specifically fine-tuned on medical corpora, making it particularly effective at capturing these nuances and semantic subtleties in clinical texts [35]. Pairwise cosine similarity is computed for every token embedding in the human reference and LLM-generated text, creating a similarity matrix of size M * N where M and N are the number of tokens in the reference and generated sentences. Each token from the generated answer is greedily matched to the most similar token in the reference answer to compute recall, precision and F1 score [24].

In a clinical setting, this method of analysis holds potential. A heatmap provides a visual representation of which parts of the LLM-generated text align closely with the reference, allowing experts to pinpoint areas of strong semantic similarity. Conversely, it can also reveal gaps or areas of divergence where critical information is missing or inadequately addressed. For instance, if a human reference emphasises specific adverse effects or monitoring parameters that are absent from the LLM-generated response, the heatmap would highlight these as regions of low similarity.

4.2.2. QNLI

Interestingly, human responses show a broader distribution than expected for Electra QNLI entailment scores (Fig. 2), with a significant number of scores closer to 0. This disparity can partly be attributed to the original method of recording the gold standard human answers. Human responses were mostly written retrospectively, focusing on providing the necessary information without explicitly referencing the original question. As a result, the explicit connection between the question and the answer may be less apparent, even where the content is clinically accurate, leading to lower entailment scores. In contrast, LLM-generated responses often included reiterations of the original question, which may have inflated their entailment scores compared to human answers.

4.2.3. Expert Human Evaluation

The evaluation of four responses to each of the 24 questions across four criteria resulted in a total of 384 individual evaluations. This presented a significant challenge in recruiting participants, as the scale of the task required considerable time and expertise. The relevance of the answer refers to scores from the assessment "To what extent do you agree the answer provided adequately addresses the proposed question?" Analysis of the post-hoc Dunn's test for relevance at level two questions (Table 3) alongside Fig. 3 demonstrated that ChatGPT 40 statistically outperformed human reference answers. This trend was consistent across all question levels, as reflected in Table 4. Overall, relevance scores were consistently high across all models, which aligns with expectations since all responses were explicitly designed to address the given questions.

For the assessment of the criterion "The answer provided is accurate and correct", Vanilla RAG outscored all other responses at level one (Table 1), while human reference answers scored the lowest. From Fig. 3, both human responses and Graph RAG responses displayed the greatest variability in scores. The lower scores and wider distribution observed for Graph RAG were due to the two instances of inaccurate context retrieval described in 4.1.

When examining the largest disparities between human and Vanilla RAG responses, key differences emerged in depth of explanations, clarity of structure, risk management, and scope of advice. Vanilla RAG consistently offered clear reasoning behind recommendations, actionable steps, and highlighted red-flag symptoms for escalation. In contrast, while human responses were technically accurate, they often lacked explanations, provided minimal counselling, and remained narrowly focused on the immediate question. This stems from the retrospective nature of human response documentation, where contextual details have been omitted.

Across all question levels, median scores for human reference answers were not statistically different from those observed for ChatGPT 40 and Graph RAG (Table 5). However, there was a difference between Graph RAG and Vanilla RAG. This could be because the interquartile range for Graph RAG was greater than that of human responses, the range was the same (Fig. 3). Notably, Vanilla RAG consistently outperformed Graph RAG across levels two and three questions, albeit marginally. This suggests that Vanilla RAG's retrieval was more effective at maintaining accuracy across increasing question complexity compared to the implemented Graph RAG architecture.

Patient safety risk can almost be interpreted as an inverse of the accuracy criterion. If an answer is accurate and correct, it is generally more likely to be safe. In this assessment, Vanilla RAG statistically outperformed human reference answers and Graph RAG, but not ChatGPT 40 (Table 6). The trends observed for accuracy were mirrored in the safety scores, with lower accuracy scoring responses also being deemed to pose a higher risk to patient safety.

Although, patient safety risk and accuracy are not perfectly aligned. Accuracy focuses on the factual correctness of information, while safety extends to how that information is interpreted and communicated to minimise potential harm. For example, an answer might be factually correct (accurate) but still present a safety risk if it lacks clear warnings or considerations relevant to patient care. To mitigate such risks, implementations could integrate automated flags for high-risk questions, or secondary validation by SMEs.

The use of ChatGPT 40 in the same evaluation exercise revealed a tendency to score towards the extremes (strongly agree or strongly disagree), suggesting it struggles to decipher subtleties in the evaluation process. In contrast, human evaluators provided more distributed ratings, reflecting a more nuanced assessment approach. The results also indicate that expert human evaluation and

LLM evaluation are poorly correlated. Interestingly, ChatGPT 40 consistently scored 'neutral' for the criterion assessing whether an answer was produced by a human. This suggests that ChatGPT 40 was also unable to distinguish between human and LLM-generated responses, supporting the conclusion that the blinding process was effective.

5. CONCLUSION AND FUTURE WORK

This introductory study demonstrated that LLMs achieved strong performance, showing substantial alignment with human reference answers in quantitative evaluations and outperforming them in expert qualitative assessments for relevance, accuracy, and patient safety. Interestingly, the expert human evaluation revealed limitations in the human reference answers themselves, which often prioritised brevity over detailed reasoning and context. These limitations have a downstream impact on the reliability of embedding-based metrics, as their effectiveness is inherently tied to the quality of the reference answers.

The embedding-based metrics proved valuable in assessing semantic alignment between answers. However, these methods have limitations. While CBERT F1 scores excel at capturing nuances due to fine tuning on medical corpora, it struggles to identify higher-level structural inconsistencies or logical gaps in responses. The CBERT pairwise similarity matrix shows potential for deeper analysis, offering insights into token level relationships that could be leveraged for more granular evaluation. Similarly, QNLI scores, despite their precision, cannot fully evaluate appropriateness or contextual subtleties. These limitations suggest that embedding-based and QNLI evaluation techniques should be interpreted alongside qualitative insights for a thorough, holistic assessment.

When comparing retrieval architectures, Vanilla RAG emerged as the most consistent performer across relevance, accuracy, and patient safety criteria, suggesting that its controlled retrieval process is suited for answering clinical pharmaceutical queries. Graph RAG, while capable of flexible retrieval, was hindered by inaccurate context retrieval in certain cases. KG construction and Cypher query optimisation are resource intensive tasks, both of which require meticulous engineering to ensure reliable performance.

It is also crucial to stress that this study was conducted under controlled conditions, employing refined prompt engineering strategies and templates, validated via iterative improvements. These controls are essential in guiding LLM responses and mitigating variability. However, such 'guardrails' would not exist in real world settings, where time pressured users may pose ambiguous queries. Nonetheless, the results of the expert human evaluation indicated that ChatGPT 40 demonstrated strong overall performance, suggesting its potential utility as an 'out of the box' solution in a clinical setting. To enable safe adoption, clear governance frameworks and user training are essential. These must address model limitations, accountability, and ongoing human oversight to ensure that LLMs are integrated responsibly.

At present, there is uncertainty surrounding the direct use of LLMs in patient care. While adopted in operational tasks to improve efficiency, their clinical potential is increasingly recognised. In the UK, AI tools used in patient care are considered medical devices [12]. Therefore, requiring compliance with data protection, transparency, and explainability standards. Data protection remains an ethical concern, especially for closed source LLMs, where sensitive information may be shared with third party proprietary models. Additionally, limited explainability and transparency make auditing outputs and reasoning difficult – however RAG approaches do provide insights into the generation of answers.

A further challenge is the requirement for medical devices to have a defined intended use statement. LLMs' versatility, ranging from clinical note summarisation to differential diagnoses, creates a regulatory grey area. As LLMs evolve with multimodal capabilities, regulatory frameworks must adapt. Clear standards for defining acceptability, quantifying performance, and validating outputs remain underdeveloped and require urgent attention [6].

5.1. Regulatory Considerations and Safe Deployment

To ensure the safe and effective deployment of LLMs within the NHS, several critical considerations must be addressed. LLMs, particularly those supporting clinical decision-making, require well defined intended use statements to establish their scope and limitations. Engineers developing LLM architectures for question-answering systems should prioritise Retrieval-Augmented Generation (RAG) techniques, as they enhance interpretability and transparency, mitigating concerns related to the 'black box' nature of AI systems.

During the design and development phase, adherence to established regulatory and quality standards is essential. Developers should align their processes with the following international standards:

- 1. ISO 13485 Quality management systems for medical devices
- 2. ISO 14155 Clinical investigation of medical devices
- 3. ISO 14971 Risk management for medical devices
- 4. IEC 62304 Software lifecycle processes for medical devices

Compliance with these standards ensures that LLM systems are designed with rigorous attention to safety, quality, and risk management. Additionally, developers should complete DCB 0129, a framework that assists health IT manufacturers in demonstrating the clinical safety of their product. Given the overlap between these international standards and DCB 0129, this process is relatively straightforward. Following this, healthcare organisations implementing these systems should co-produce DCB 0160, a standard focusing on clinical risk management in deployment, typically led by the institution's Clinical Safety Officer.

Beyond regulatory compliance, a comprehensive data protection impact assessment must be conducted in collaboration with information governance teams and the Caldicott Guardian to address data security and patient confidentiality concerns. Evaluation criteria should be established with clinical teams to assess the LLM's performance before full deployment. A shadow mode deployment should be considered—where the system operates in the background, collecting data for evaluation without influencing real world decision making. A robust post-deployment surveillance plan must also be implemented to continuously monitor system performance, ensuring it remains aligned with clinical standards.

Adhering to these rigorous guidelines ensures that LLMs conform to the same safety and quality standards as medical devices. Should regulatory bodies formally classify LLMs as medical devices in the future, manufacturers that have followed these protocols will already have the necessary documentation in place, preventing service disruptions and ensuring seamless compliance.

This pilot is a step towards addressing these challenges by exploring robust evaluation methodologies for LLMs in clinical contexts. Combining quantitative metrics with qualitative expert evaluation offers a holistic understanding of LLM performance. While hurdles remain, these findings provide insights for advancing regulatory compliance, clinical integration, and performance validation in healthcare.

5.2. Limitations and Future Work

This pilot study has limitations warranting further exploration. The small sample size (n=24) for expert evaluation affects the generalisability of the findings, particularly given the limited number of level three questions. Expanding the dataset to cover a broader range of clinical queries would enhance the robustness of the results. Additionally, conducting expert evaluations across multiple rounds would improve inter-rater reliability, addressing potential subjectivity in qualitative assessments.

Further investigation is required to refine Graph RAG's retrieval accuracy, particularly in mitigating errors from ambiguous queries. Enhancing retrieval mechanisms and evaluating alternative post-retrieval processing techniques may improve precision in contextual retrieval.

Lastly, this study does not assess long term performance stability, an essential factor for clinical deployment. Future research should explore continuous validation methodologies such as shadow deployment evaluations to prospectively track performance drift over time. These approaches would ensure the safe and effective integration of LLMs into healthcare workflows.

ACKNOWLEDGEMENTS

Sincere thanks are extended to the Oxford University Hospitals Medicines Management and Therapeutics Committee for providing access to the medicines information leaflets, and to the Pharmacy Department for granting access to the medicines information database, which enabled the collection of human reference answers.

REFERENCES

- [1] British Medical Association. 2023. An NHS under pressure. [Online]. [Accessed 18 August 2024]. Available from: https://www.bma.org.uk/advice-and-support/nhs-delivery-andworkforce/pressures/an-nhs-under-pressure
- [2] Innes, A. J., Bramley, D. M. and Wills S. 2014. The impact of UK Medicines In-formation services on patient care, clinical outcomes and medicines safety: an evaluation of healthcare professionals' opinions. Eur J Hosp Pharm. 21(4), pp.222-228.
- [3] NHS England. [no date]. About the NHS AI Lab. [Online]. [Accessed 16 August 2024]. Available from: https://transform.england.nhs.uk/ai-lab/about-the-nhs-ai-lab/
- [4] Koman, J., Fauvelle, K., Schuck, S., Texier, N. and Mebarki, A. 2020. Physicians' Perceptions of the Use of a Chatbot for Information Seeking: Qualitative Study. J Med Internet Res. 22(11), article no: e15185 [no pagination]
- [5] Morath, B., Chiriac, U., Jaszkowski, E., Deiß, C., Nürnberg, H., Hörth, K., Hoppe-Tichy, T. and Green, K. 2023. Performance and risks of ChatGPT used in drug information: an exploratory realworld analysis. Eur J Hosp Pharm. 31(6), pp. 491-497.
- [6] Parry, C. 2023. AI chatbots in pharmacy: a brave new world or looming threat?. Pharm. J. 311, article no: 7976 [no pagination]
- [7] Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S. and Li, Q. 2024. [Pre-print]. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. arXiv:2405.06211v3 [cs.CL].
- [8] Ke, Y. H., 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. 2024. Development and Testing of Retrieval Augmented Generation in Large Language Models - A Case Study Report. arXiv:2402.01733v1 [cs.CL].
- [9] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M. and Wang, H. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997v5 [cs.CL].
- [10] Trajanoska, M., Stojanov, R. and Trajanov, D. 2023. Enhancing Knowledge Graph Construction Using Large Language Models. arXiv:2305.04676v1 [cs.CL].

- [11] Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., Zhang, Y. and Tang, S. 2024. Graph Retrieval-Augmented Generation: A Survey. arXiv:2408.08921v1 [cs.AI].
- [12] MHRA. 2024. Software and AI as a Medical Device. [Online]. [Accessed 12 December 2024]. Available from: www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device
- [13] Feldman, R., Aldana, E. and Stein, K. 2019. Artificial Intelligence in the Health Care Space: How We Can Trust What We Cannot Know. Stan. L. & Pol'y Rev. 30(2), pp.399-419.
- [14] NICE. 2022. Evidence standards framework for digital health technologies. [Online]. [Accessed 12 August 2024]. Available from: www.nice.org.uk/corporate/ecd7/chapter/how-to-meet-the-standards
- [15] Gao, M., Hu, X., Ruan, J., Pu, X. and Wan, X. 2024. [Pre-print]. LLM-based NLG Evaluation: Current Status and Challenges. arXiv:2402.01383v2 [cs.CL].
- [16] Gehrmann, S., Clarke, E. and Sellam, T. 2022. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. arXiv:2202.06935v1 [cs.CL].
- [17] Benary, M., Wang, X. D., Schmidt, M., Soll, D., Hilfenhaus, G., Nassir, M., Sigler, C., Knödler, M., Keller, U., Beule, D., Keilholz, U., Leser, U. and Rieke, D. T., 2023. Leveraging Large Language Models for Decision Support in Personalized Oncology. JAMA Netw Open. 6(11), article no: e2343689 [no pagination]
- [18] Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G. and Rueckert, D. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 30, pp.2613-2622.
- [19] Sblendorio, E., Dentamaro, V., Cascio, A., Germini, F., Piredda, M. and Cicolini, G. 2024. Integrating human expertise & automated methods for a dynamic and multi-parametric evaluation of large language models' feasibility in clinical decision-making. Int. J. Med. Inform. 188, article no: 105501 [no pagination]
- [20] Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCar-thy, K. R., Osterhoudt, H., Wu, X., Visweswaran, S., Fu, S., Mathur, P., Cacciamani, G. E., Sun, C., Peng, Y. and Wang, Y. 2024. A framework for human evaluation of large language models in healthcare derived from literature review. npj Digit Med. 7, article no: 258 [no pagination]
- [21] Lee, C., Gatt, A., Miltenburg, E. and Krahmer, E. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. Computer Speech & Language. 67, article no: 101151 [no pagination]
- [22] NHS UK Medicines Information. 2024. Guidance on assessing enquiries completed by Medicines Advice Services. [Online]. [Accessed 12 September 2024]. Avail-able from: https://www.ukmi.nhs.uk/Resources
- [23] Xu, S., Wu, Z., Zhao, H., Shu, P., Liu, Z., Liao, W., Li, S., Sikora, A., Liu, T and Li, X. 2024. Reasoning before Comparison: LLM-Enhanced Semantic Similarity Metrics for Domain Specialized Text Analysis. arXiv:2402.11398v2 [cs.CL].
- [24] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675v3 [cs.CL].
- [25] Nelson, F. 2020. Why health is the 'worst case' scenario for text mining. [Online]. [Accessed 12 December 2024]. Available from: www.medicalrepublic.com.au/why-health-is-the-worst-case-scenario-for-text-mining/3171
- [26] Jeong, M., Sung, M., Kim, G., Kim, D., Yoon, W., Yoo, J., and Kang, J. 2021. Transferability of Natural Language Inference to Biomedical Question Answering. arXiv:2007.00217v4 [cs.CL].
- [27] Blease, C.R., Locher, C., Gaab, J., Hägglund, M. and Mandl, K.D. 2024. Generative artificial intelligence in primary care: an online survey of UK general practitioners. BMJ Health & Care Informatics. 31(1), article no: e101102 [no pagination]
- [28] Neo4j. 2024. Graph Database [Computer software], Version 5.20.0
- [29] LangChain. 2023. LLMGraphTransformer [Computer software], Version 0.3.4.
- [30] Wood, T. 2024. Drug Named Entity Recognition [Computer software], Version 2.0.4.
- [31] Nicoomanesh, A. 2024. Token Efficiency and Compression Techniques in Large Language Models: Navigating Context-Length Limits. [Online]. [Accessed 27 Octo-ber 2024]. Available from: https://medium.com/@anicomanesh/token-efficiency-and-compression-techniques-in-large-languagemodels-navigating-context-length-05a61283412b

- [32] Soman, K., Rose, P. W., Morris, J. H., Akbas, R. E., Smith, B., Peetoom, B., Villouta-Reyes, C., Cerono, G., Shi, Y., Rizk-Jackson, A., Israni, S., Nelson, C. A., Huang, S., and Baranzini, S. E. 2024. Biomedical knowledge graph-optimized prompt generation for large language models. Bioinformatics.40(9), article no: btae560 [no pagination]
- [33] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. arXiv:2007.15779v6 [cs.CL].
- [34] Dinno, A. 2015. Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. The Stata Journal. 15(1), pp. 292–300.
- [35] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M. B. A. 2019. Publicly Available Clinical BERT Embeddings. arXiv:1904.03323v3 [cs.CL].

APPENDIX A

Table 7. Summary of statistical results for embedding-based metrics (to five decimal places)

	Cosine Similarity			CBERT F1 Scores		
	ChatGPT 40	Vanilla RAG	Graph RAG	ChatGPT 40	Vanilla RAG	Graph RAG
Mean	0.91528	0.90862	0.91146	0.76313	0.75409	0.75214
Std	0.03439	0.03980	0.03451	0.03428	0.04004	0.02980
Min	0.82063	0.81234	0.80520	0.67284	0.68803	0.68522
Lower Q	0.89875	0.88063	0.89416	0.74178	0.72178	0.72743
Median	0.92404	0.91030	0.91955	0.76653	0.75590	0.75327
Upper Q	0.91887	0.90626	0.91911	0.76299	0.75411	0.75141
Max	0.96618	0.96345	0.95790	0.81924	0.81795	0.80894

© 2025 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.