Artificial Intelligence and Language Models

ACS Publishing

Article

Modeling and Benchmarking GraphRAG for Indonesian Legal Question Answering

Dea Nabila[1]*, Arbi Haza Nasution[1], Yohei Murakami[2], Stefan Koos[3], Ahmet Emre Ergun[4]

[1] Department of Informatics Engineering, Universitas Islam Riau, Indonesia
[2] Faculty of Information Science and Engineering, Ritsumeikan University, Japan
[3] Faculty of Economics and Organizational Sciences, Universität der Bundeswehr München, Germany
[4] Department of Computer Engineering, Faculty of Engineering and Architecture, Izmir Katip Celebi University, Türkiye
* Correspondence: deanabila389@gmail.com;

Abstract: This study explores the integration of Graph Retrieval-Augmented Generation (GraphRAG) with legal question answering in the context of Indonesian civil law (KUH Perdata). Unlike traditional RAG systems, GraphRAG leverages graph-structured knowledge representations in Neo4j to capture the hierarchical and relational nature of legal texts, enabling more precise and contextually faithful responses. Using 2,128 legal articles as the source corpus, the Indonesian Legal GraphRAG model supports structured retrieval across books, chapters, sections, and articles of the Civil Code. Several large language models (LLMs) of varying scales—very large, large, and mid-sized—were benchmarked using RAGAs metrics for faithfulness, answer relevancy, and context entity recall. Results show that Llama4-Maverick demonstrates higher performance than GPT-4o in specific metrics such as faithfulness and contextual grounding. These findings highlight the effectiveness of graph-based retrieval modeling for enhancing factual consistency and contextual relevance in legal QA and provide a new resource and benchmark for the Indonesian legal domain.

Keywords: Llama4-Maverick, GPT-4o, Large Language Models, Graph Retrieval-Augmented Generation, Question Answering, Indonesian Civil Law

Citation: Dea Nabila, Arbi Haza Nasution, Yohei Murakami, Stefan Koos, Ahmet Emre Ergun. Modeling and Benchmarking GraphRAG for Indonesian Legal Question Answering. *Artificial Intelligence and Language Models* 1–12. https://doi.org/

Received: 15 March 2026
Revised: 26 March 2026
Accepted: 27 March 2026
Published: 27 March 2026

1. Introduction

The advancements in Large Language Models (LLMs) have significantly enhanced Natural Language Processing (NLP) with remarkable success in many fields [1–5]. LLMs' ability to respond to queries when prompted [6] has led it to research with a wide variety of contexts. Despite that, there is not yet any research using the Indonesian Civil Law (KUH Perdata). Any person can perform legal acts [7], and in Indonesia, KUH Perdata has a big role in addressing this civil legal matter. AI chatbots in a legal context can enhance user experience [8]. With legal Question Answering (QA), people can be assisted when encountering civil justice problems [9] in their daily lives. This highlights the importance of an effective legal information retrieval system [10] to improve the QA. Information retrieval is the act of finding materials of an unstructured nature that contain the information needed from a large collection [11]. The quality of the retrieval system decides the precision and recall of the provided answers.

LLMs have notable limitations in legal applications. Such as the inability to access up-to-date or case-specific context in real-time and "hallucination", which is when the models generate statements that are inaccurate or misleading [12]. In the legal field, source texts and statutes are essential. Those hallucinations can lead to incorrect or harmful advice. For example, a lawyer used ChatGPT and was sanctioned because of a fabricated non-existent case citation (hallucinations).

Retrieval-Augmented Generation (RAG) emerged as an approach to help improve LLMs [13] [14]. RAG combines the strengths of information retrieval and generative models

[15], making more accurate responses by grounding LLM outputs in relevant documents. But even as RAG has shown success in contextualized generation [16], it fails to capture significant structured relational knowledge [17]. Native RAG relies on keyword matching and vector similarity, which is often inadequate for capturing multi-step legal reasoning processes for better accuracy and comprehensiveness [18]. For example, when asked about the connection between the books and articles from the KUH Perdata, the RAG system retrieves only directly related data, missing the relationship concepts such as books, chapters, and sections. This can lead to incomplete or inaccurate responses. Given that the KUH Perdata has a clear hierarchy of books, chapters, sections, and articles, the system needs to preserve the structure to maintain legal precision in its answer. It is necessary to include accurate law references [8].

Knowledge Graphs (KGs) offer a way to represent information with nodes and edges [19] [20]. The elements in the data are turned into nodes, and the edges are their relationships. KGs are used for domain-specific use cases, showing more concise and accurate sources [21]. Graph Retrieval-Augmented Generation (GraphRAG) leverages KGs to overcome the shortcomings of traditional RAG. GraphRAG serves as an alternative framework for enhancing LLM performance [22] [23] [24]. GraphRAG first uses the LLMs to construct a knowledge graph, then it partitions the graph into a hierarchy of communities, and after that, the GraphRAG answers queries by retrieving specific subgraph information with map-reduce processing [22]. By using GraphRAG, models will be more optimized and able to process complex knowledge [25]. The models will also have enhanced retrieval accuracy [26] and ground their answer in a trusted law source.

This study provides a comprehensive analysis of GraphRAG effectiveness in improving QA in the Indonesian Civil Law context. By addressing the hierarchical structure of Indonesian legal documents, organized into books, chapters, sections, and articles, the research introduces a legal QA system that leverages GraphRAG to improve contextual understanding and retrieval accuracy. This approach contributes to the advancement of legal domain-specific AI applications.

The research implements Neo4j for its KG construction and GraphRAG workflows, using its advanced performance in handling complex, interconnected legal data. GraphRAG uses structured KGs to model entity connections, reducing ambiguity and improving answer precision. By integrating KG with Indonesian legal law, this work advances context-aware legal chatbots.

As GPT-4o has shown superior performance [27] [28] in its multimodal reasoning capabilities and legal comprehension, it is chosen as the language model for the QA along with the open models (Gemma3-27B, Llama3.3-70B, Phi3-14B, Phi4-14B, Qwen2.5-32B, Llama4-Scout, and Llama4-Maverick). The models are carefully selected based on their performance data in another research study by [29] and small experiments conducted to ensure greater reliability. As for Llama4-Scout and Llama4-Maverick, they have been selected for this research due to their promised superior performance as the newest part of the Meta Llama 4 suite. These models introduce significant advancements in natural language processing capabilities, particularly in their ability to handle complex tasks across various domains.

2. Methods

2.1. Dataset

The dataset contains a total of 4 books, 66 chapters, 160 sections, and 2,128 KUH Perdata articles. Each book, chapter, and section has its own title to describe the content. In addition, there is also a dataset that contains a list of queries. It has 66 questions that cover all chapters of KUH Perdata. They were made while bearing in mind what might be used in real-life cases of civil law.

Table 1. Relationship Types in the Indonesian Legal Knowledge Graph

Relationship	Source → Target	Description
`BELONGS_TO`	Book → Document Chapter → Book Section → Chapter Article → Section	Hierarchical ownership showing inclusion of smaller units.
`CONTAINS`	Document → Book Book → Chapter Chapter → Section Section → Article Article → Chunk	Compositional relationship showing content inclusion.
`PART_OF`	Article → Section Section → Chapter Chapter → Book	Defines structural hierarchy from smaller to larger units.
`HAS_ENTITY`	Chunk → Category Article → Category	Connects legal text segments to identified entities or topics.
`NEXT_CHUNK`	Chunk ↔ Chunk	Sequential connection between consecutive text segments.
`SIMILAR`	Chunk ↔ Chunk Article ↔ Article	Semantic similarity connection between related content.

2.2. Modeling Indonesian Legal GraphRAG

The knowledge graph models the hierarchical structure of the Indonesian Civil Code, where nodes represent textual and conceptual elements and relationships capture both structural and semantic dependencies. The schema enables both vertical navigation (e.g., from an Article to its parent Chapter) and horizontal traversal (e.g., between semantically similar Articles or Chunks). This dual-layer representation supports context-aware retrieval and reasoning in the GraphRAG pipeline.

To construct the knowledge graph from unstructured legal text, we employed the Neo4j LLM Knowledge Graph Builder [1], a no-code platform that uses large language models to extract entities and relationships. The system ingests CSV documents (KUH Perdata) and transforms them into a structured knowledge graph using underlying LLMs from OpenAI and open models accessed via Ollama API.

The nodes in the graph are:

- **Document**: The legal document itself.
- **Chunk**: Text segments extracted from the document.
- **Book**: The KUH Perdata Books related in the document.
- **Chapter**: Chapters inside of the book.
- **Section**: Sections within chapters.
- **Article**: Articles within the document.
- **Category**: Classification category.

The knowledge graph enables contextual navigation when a legal query or question is given. The system traverses the graph to find relevant articles using a combination of hierarchical navigation and semantic similarity matching, enabling multi-hop retrieval across related legal provisions as listed in Table 1. It also enables connections between relevant data from different sections or chapters within the documents. This makes the system more precise and contextually appropriate compared to the traditional keyword-based approaches.

Figure 1 illustrates the hierarchical and semantic structure of the Indonesian Legal Knowledge Graph used in this study. The nodes represent the primary components of the Civil Code, from the full `Document` down to its granular `Chunk` level. Hierarchical relationships such as `CONTAINS`, `BELONGS_TO`, and `PART_OF` capture the compositional structure of

[1] https://github.com/neo4j-labs/llm-graph-builder

```
(Document)
 |-- CONTAINS -> (Book)
 |    |-- CONTAINS -> (Chapter)
 |    |    |-- CONTAINS -> (Section)
 |    |    |    |-- CONTAINS -> (Article)
 |    |    |    |    |-- CONTAINS -> (Chunk)
 |    |    |    |    |-- HAS_ENTITY -> (Category)
 |    |    |    |    |-- SIMILAR <-> (Article)
 |    |    |    |    \-- PART_OF -> (Section)
 |    |    |    \-- PART_OF -> (Chapter)
 |    |    \-- PART_OF -> (Book)
 |    \-- BELONGS_TO -> (Document)
 \-- SIMILAR <-> (Document)
```

Figure 1. Knowledge Graph Schema Overview

books, chapters, sections, and articles, while semantic connections such as `HAS_ENTITY`, `NEXT_CHUNK`, and `SIMILAR` enable cross-references and context continuity between related legal provisions. This structured representation enables both hierarchical navigation—such as moving from an `Article` to its parent `Chapter`—and semantic retrieval through similarity and entity connections. By integrating multiple relational layers, the GraphRAG framework facilitates multi-hop reasoning and context-aware retrieval, allowing it to capture the logical dependencies within the Indonesian Civil Code more effectively than conventional vector-based approaches.

Although GraphRAG has been introduced in prior works, our contribution lies in adapting it for the Indonesian Civil Law corpus (KUH Perdata), which poses unique linguistic and structural challenges. Constructing the knowledge graph from these legal documents required addressing several key difficulties:

- **Hierarchical document organization:** The KUH Perdata is deeply nested into *Books*, *Chapters*, *Sections*, and *Articles*. This hierarchy must be preserved in the knowledge graph to ensure contextual integrity. We represent each level as a node type (`Book`, `Chapter`, `Section`, `Article`), linked by edges such as `CONTAINS`, `BELONGS_TO`, and `PART_OF`.
- **Cross-referencing complexity:** Many articles cite or implicitly reference other sections (e.g., "menurut ketentuan Pasal 1320"). As there is no standardized citation format, these references were semantically extracted by the Neo4j LLM Graph Builder and represented through `HAS_ENTITY` and `SIMILAR` relationships rather than explicit cross-reference edges.
- **Legal language variation:** The same legal concept may appear in multiple linguistic forms (e.g., "perjanjian" vs. "kontrak"). Entity extraction and normalization were required so that related concepts could be connected through `HAS_ENTITY` links between `Chunk` and `Category` nodes. To ensure consistency in entity representation, extracted entities were normalized using lexical matching and semantic grouping. Specifically, synonymous legal terms (e.g., "perjanjian" and "kontrak") were mapped to the same Category node using embedding-based similarity thresholds. Additionally, similarity relationships (SIMILAR) between Articles and Chunks were generated using cosine similarity over sentence embeddings, with a predefined threshold to connect semantically related legal provisions.

2.3. Benchmarking LLMs in the Indonesian Legal GraphRAG

To evaluate the effectiveness of the proposed GraphRAG framework in the Indonesian legal context, we benchmarked a diverse set of Large Language Models (LLMs) representing three capability tiers: very large, large, and mid-sized. This categorization allows for a comparative understanding of how model size and training alignment influence reasoning faithfulness and retrieval-grounded performance.

The very large models include GPT-4o, Llama4-Maverick, and Llama4-Scout, each exceeding 100 billion parameters and designed for complex, multi-hop reasoning tasks. These models are expected to perform best in maintaining factual accuracy and semantic grounding when integrated with GraphRAG, given their superior comprehension and alignment tuning. In particular, Llama4-Maverick introduces improvements in retrieval alignment and response conservativeness, making it suitable for domains requiring high factual integrity such as legal reasoning.

The large model, Llama3.3-70B, represents the prior generation of open models that, despite its substantial scale, was trained without advanced graph or retrieval alignment techniques. It serves as a baseline to assess how well a large, general-purpose model can utilize graph-based context without additional fine-tuning.

The mid-sized models—Gemma3-27B, Qwen2.5-32B, Phi3-14B, and Phi4-14B—are more computationally efficient but often exhibit weaker grounding to external context, making them useful for assessing the lower bound of GraphRAG's effectiveness in constrained-resource settings. Their inclusion provides insight into how smaller models leverage structured graph retrieval compared to relying solely on internal parametric knowledge.

All models were integrated within the same GraphRAG pipeline, connected to the Neo4j-based Indonesian Legal Knowledge Graph. Each model received identical graph-derived context during generation to ensure a fair comparison. This setup enables the evaluation of how different LLM architectures—proprietary and open-source—interact with structured legal knowledge to generate factually grounded and contextually relevant answers.

The benchmarking not only assesses model accuracy but also examines how effectively each LLM leverages graph-based retrieval for multi-hop legal reasoning. This provides a foundation for understanding the interplay between model capability, structured knowledge grounding, and reliability in Indonesian legal question answering.

2.4. Evaluation Guidelines

The research focuses on generating contextually accurate responses about Indonesian civil law. Each model will need to answer all the questions given and will be evaluated based on the generated answers using the Retrieval Augmented Generation Assessment (RAGAs) score of faithfulness, answer relevancy, and context entity recall [30]:

- **Faithfulness (F):** Measures factual consistency between the generated answer (A) and the retrieved legal context (C):
$$F = \frac{|S_{A \cap C}|}{|S_A|}$$
where S_A is the set of factual statements in A, and $S_{A \cap C}$ are those verifiable in C. Higher faithfulness means fewer hallucinated or unsupported statements.
- **Answer Relevancy (R):** Evaluates semantic similarity between the model's answer and the gold reference answer (A_{ref}) using cosine similarity on sentence embeddings:
$$R = \cos(\text{emb}(A), \text{emb}(A_{ref}))$$
capturing how closely the model addresses the user's query.
- **Context Entity Recall (CER):** Measures how many legal entities or article references from the context appear in the generated answer:
$$CER = \frac{|E_A \cap E_C|}{|E_C|}$$
where E_A and E_C denote entities in the answer and context. High CER indicates strong grounding in the retrieved legal passages.

Table 2. Evaluation Metrics Across Different Language Models

Model	Faithfulness	Answer Relevancy	Context Entity Recall
Very Large Models			
GPT-4o	0.762	**0.920**	0.366
Llama4-Maverick	**0.818**	0.498	**0.380**
Llama4-Scout	0.694	0.482	0.270
Large Models			
Llama3.3-70B	0.195	0.629	0.035
Mid-sized Models			
Gemma3-27B	0.544	0.298	0.190
Qwen2.5-32B	0.280	0.475	0.085
Phi3-14B	0.090	0.478	0.030
Phi4-14B	0.240	0.479	0.070

3. Result

Each of the models answered the same set of legal questions using the GraphRAG pipeline with identical knowledge graph and retrieval settings. The models include Llama4-Maverick, Llama4-Scout, Phi3-14B, Phi4-14B, Gemma3-27B, Qwen2.5-32b, and Llama3.3-70B. These models were chosen to represent a range of sizes and presumed capabilities, with Llama4-Scout and Maverick expected to have strong performance due to their latest-generation architecture, and others included based on prior performance indications. We report their evaluation scores in Table 2 and highlight key findings below. Overall, GraphRAG improved the accuracy of legal QA across all models, but the degree of improvement and the balance between faithfulness and relevancy varied considerably between models. Table 2 illustrates a comparison of all average scores under the GraphRAG evaluation. Rather than reiterating numerical values, we focus on interpreting key performance trends across model categories. Several trends and notable results emerged on very large models, large models, and mid-sized models.

3.1. Very Large Models

Advanced models delivered the best performance on GraphRAG. GPT-4o achieves the highest answer relevancy (0.920), indicating strong alignment with user queries and effective utilization of retrieved context. It also maintained a faithfulness score of (0.762), indicating that even when using the graph, GPT-4o rarely introduced content that wasn't supported by the retrieved legal context. But even so, GPT-4o scored below Llama4-Maverick for other metrics.

The Llama4-Maverick model surprisingly achieved the highest faithfulness with 0.818 and context entity recall with 0.380, surpassing GPT-4o. This suggests that Maverick is extremely conservative about sticking to the provided knowledge, a desirable trait in the legal domain, as it minimizes hallucinations. Indeed, qualitative examination of Maverick answers shows it would often explicitly say "based on the given context, there is no information [to answer]..." when the graph documents did not contain a clear answer, rather than attempting to invent one. This behavior yields a high faithfulness score since it never states unsupported facts, outperforming even GPT-4o in factual grounding, although it sometimes came at the cost of answer relevancy. The answer relevancy score was around 0.498 for Maverick, meaning about half of its answers were fully relevant. In many cases where the answer was incomplete or not directly addressing the question, Maverick had chosen not to go beyond the evidence. Llama4-Maverick excels in faithfulness and context entity recall, outperforming GPT-4o by 6.17% and 5.26% respectively.

Llama4-Scout showed a similar pattern. It achieved a solid faithfulness of 0.694 and a relevancy of around 0.482. These results for the Llama4 models indicate that the Meta fine-tuned Llama 4 series can be very reliable and faithful, but slightly less forthcoming in

answering if the graph retrieval does not provide a clear path. In practice, it leads these models to sometimes respond with "I do not have enough information from the provided sources" or give partial answers, which, while not fully satisfying to a user, is certainly safer legally than fabricating an answer. Notably, all these larger models (GPT-4o, Llama4-Maverick, Llama4-Scout) also demonstrated relatively high context entity recall, which is around 0.27–0.38 on average, with GPT-4o and Maverick at 0.36–0.38, Scout at 0.27. This shows they not only stayed truthful but also pulled in the relevant legal terminology and references from the graph context frequently. For example, GPT-4o and Maverick often cited specific article numbers or legal terms or entities from the Civil Code in their answers, which boosts the confidence that the answer is grounded in actual law text.

3.2. Large Model

The large model showed unique strengths and weaknesses. Llama3.3-70B, a large 70B model from a previous generation, exhibited almost the opposite pattern. It scored the second-highest in answer relevancy at 0.629, only behind GPT-4o, but its faithfulness was very low (0.195), compared to the much smaller models. This indicates that Llama3.3-70B frequently produced answers that sounded relevant and comprehensive, likely due to its large parametric knowledge and language ability, but those answers were not reliably grounded in the retrieved graph data. In essence, Llama3.3-70B might "know" a lot about Indonesian civil law from pre-training and thus can answer many questions reasonably well without needing the retrieval. However, if it relies on that internal knowledge and the answer is not strictly cross-checked with the graph context, it gets penalized in faithfulness.

In a legal setting, the Llama3.3 answers, while often relevant, could be unsupported by actual cited law, which is a serious shortcoming if we require evidence-backed responses. Its low context entity recall of 0.035 confirms that it seldom referenced the specific entities from the documents, preferring its own wording. Interestingly, this result highlights that bigger model size alone does not guarantee better use of external knowledge – model training and fine-tuning approach matter. Llama4-Maverick dramatically outperformed the older Llama3.3 of greater size in terms of grounding and factuality, illustrating advancements in model alignment with retrieval.

3.3. Mid-sized Models

Mid-sized general models such as the Phi and Qwen series had mixed results, generally maintaining moderate relevancy but low faithfulness. The models Phi3-14B, Phi4-14B, and Qwen2.5-32B all scored in the same ballpark for answer relevancy, approximately 0.47–0.48 on average, meaning roughly half of their answers were on topic. This suggests that even these smaller or less specialized models were often able to pick up on the intent of the question and attempt an answer when using GraphRAG. However, their faithfulness scores were dramatically lower, at 0.09, 0.24, and 0.28, respectively. Such low faithfulness indicates that these models frequently introduced information not found in the retrieved context, and often hallucinated or pulled from their prior knowledge incorrectly. In other words, while they might generate an answer relevant to the question, they were not reliably using the graph-provided evidence to ground that answer. For instance, for a question about a specific legal term, Phi4-14B might give a definition or rule that sounds plausible for that term, keeping the answer relevant, but if that detail was not present in the retrieved legal text, it would count as not faithful.

This pattern highlights a risk for smaller LMs, even with GraphRAG, they may default to their internal knowledge base, especially if they do not fully understand how to use the graph context. We also observed that their context entity recall was very low (phi3: 0.03, phi4: 0.07, Qwen: 0.085). Their answers were often phrased generically and failed to mention the concrete entities, such as law articles from the context. This reinforces that these models underutilize the graph, either due to limited capacity or insufficient fine-tuning for retrieval augmentation, reducing both the factual accuracy and the richness of their answers. In practical terms, while GraphRAG gave these models a chance to find the right

information, they did not always follow through, potentially giving confident-sounding but unverified answers. Such behavior is dangerous in legal QA, as it could mislead users with incorrect legal information. The stark contrast between their relevancy and faithfulness means that caution is needed when employing smaller models, as they may need additional training to better ground their answers in provided references.

The Gemma3-27B model achieved a faithfulness of 0.544, which is substantially higher than the phi and Qwen models, indicating it incorporated the graph evidence more often. However, its answer relevancy was only 0.298, the lowest relevancy among all models tested. On examining the Gemma3 outputs, we found it often summarized parts of the context without actually answering the precise question asked. For instance, if the question was about the primary obligation of a lessee in a rental agreement, Gemma3 might recite several obligations from the law (as found in the retrieved graph context) but not clearly identify the "primary" one, or it might provide legal context that does not directly resolve the query. This behavior yields decent faithfulness because it is quoting the law correctly. It is on-topic factually, but poor in relevancy because it does not directly address the question asked. Such a model might be leveraging GraphRAG to fetch the right documents, but it lacks the decisiveness or clarity in its generation to form a reliable answer.

Among open models, Llama4-Maverick emerged as a strong candidate for GraphRAG in legal QA, achieving the highest faithfulness in most answers and a decent relevancy, which is a promising result for a model outside the proprietary GPT-4 family. Llama4-Scout was not far behind, also showing a good balance. These suggest that the newest generation open models, when paired with GraphRAG, can approach the reliability of GPT-4o, at least in terms of factual correctness, though they may still trail in completeness of the answer. Other models, like Gemma3-27B, while utilizing the graph to stay factual, need improvement in how directly they answer the question. And models like Phi4-14B, Qwen-32B, or Llama3.3-70B would likely require further fine-tuning or prompt engineering to reduce their hallucination tendency when using GraphRAG. Notably, there is a strong correlation between faithfulness and context recall across the models. Those models that are better at grounding their answers also tend to include more specific context details with high entity recall, as seen with GPT-4o and Llama4 variants. This correlation reinforces that the benefit of GraphRAG is maximized only when the model properly attends to the graph-provided evidence.

4. Discussion

The above results highlight several important insights into using Graph Retrieval-Augmented Generation for legal QA, with Llama4-Maverick having the highest score for faithfulness and context entity recall, and GPT-4o having the highest score for answer relevancy.

4.1. Model-Specific Performance

The second part of our results compared GraphRAG performance across various LLMs. This analysis reveals that while GraphRAG provides a strong foundation for retrieval, the ability to utilize that foundation is critical. Each model had distinct behavior:

4.1.1. GPT-4o

With GraphRAG, GPT-4o nearly always grasped the legal context correctly and provided a thorough answer. Errors were rare; when they occurred, they were usually cases of slight over-generalization. For instance, if a law article had two exceptions to a rule but the model only mentioned one, or if it paraphrased a legal definition a bit loosely. Overall, GPT-4o produced highly coherent, precise answers that included the necessary legal references.

The combination of GraphRAG with GPT-4o resulted in answers that often read like a well-prepared legal explanation, complete with citations. This indicates that GPT-4o's few mistakes can often be traced to limitations in retrieved info or brevity in answer, rather

than misunderstanding; it shows excellent integration with the graph. The implication for deployment is clear: if resources permit, GPT-4 with GraphRAG offers assurance of quality, making it suitable for applications where accuracy is paramount (e.g., a legal research assistant or a tool for verifying legal arguments).

4.1.2. Llama4-Maverick and Llama4-Scout

As noted, Llama4-Maverick exceeded GPT-4o in faithfulness and context entity recall, which is an impressive feat. The error profile of this model shows almost no fabrication of facts, a testament to their fine-tuning, likely alignment tuning to refuse answering when unsure. This is a very desirable trait in the legal domain, as it minimizes unsupported facts. In practice, the Llama4-Maverick answer would sometimes be incomplete or contain an explicit disclaimer of missing information rather than risk an unsupported claim.

Llama4-Scout is similar, though slightly less strict; it had a couple more instances of trying to answer and getting it partially wrong, which explains its somewhat lower faithfulness than Maverick. Both models leveraged the graph well when the info was present, they delivered it correctly. Their relatively high context recall means they frequently used the exact language of the law, which is ideal for legal explanations.

For example, on a question asking for something not explicitly stated in the context (e.g., "who can oppose a demand for separation of marital assets?" when the context didn't specify "who"), Maverick responded that the information wasn't found in the provided context, whereas GPT-4o tried to infer an answer from general knowledge. This conservative approach means Llama4-Maverick answers are extremely trustworthy (everything stated is backed by the graph), but the trade-off is the relevancy. The user might not get a full answer if the system does not retrieve the needed fact. Some may view this as a drawback, but in a legal assistant context, it is preferable to silence over speculation.

These evaluations validate the expectation that larger LLMs can be aligned to prioritize faithfulness, making them good candidates for domains requiring high precision. The key implication is that, with GraphRAG providing relevant facts, models like Llama4-Maverick and Llama4-Scout will handle them very carefully. For system designers, this suggests that such models might be paired with strategies to ensure retrieval returns enough information. If the graph retrieval is more comprehensive, these models will likely produce a correct and relevant answer. Conversely, if something is missing, they will signal that rather than guessing, which could be used as a trigger to retrieve more information or escalate to a human expert.

4.2. Implications and Trends

From the above, a few key trends emerge:

- **There is a positive correlation between the capability of the model to provide faithfulness and its ability to recall context entities in the answer**. Models like GPT-4o and Llama4-Maverick that scored high on faithfulness also tended to quote laws and use legal terms from the documents frequently. This makes the answers very concrete and useful for legal purposes. Models that scored low on faithfulness (phi series, Llama3.3) rarely included such specifics, presumably because they were not drawing from the documents in detail. This confirms that to get detailed and trusted answers, the model must be properly leveraging the graph input. Simply supplying a knowledge graph is not enough if the model is not able to use it fully.
- **The size and quality of the language model significantly affect outcomes**. Larger, more advanced models make far better use of GraphRAG. Smaller or less refined models either underutilize the graph or misuse it. This suggests that GraphRAG is not a silver bullet by itself. It works in tandem with the capabilities of the LLMs. For developers, this means that to achieve optimal results, one should use the best model available. If using a smaller model, additional training is necessary to ensure it behaves in a factually-grounded way. But even so, the fact that Llama4-Scout and Maverick performed so well is encouraging. It means that non-proprietary models are

catching up in their ability to handle complex, structured knowledge. This suggests the potential to deploy a reliable legal QA system without exclusive access to LLMs like GPT-4. The trend appears to be that each new generation of LLM (e.g., Llama 4 vs Llama 3) brings improvements in how well they follow evidence. This aligns with reports that newer models have better hallucination suppression and reasoning, especially when fine-tuned on instruction datasets.

- **Balancing completeness and correctness is a challenge**. We saw this with the contrast between GPT-4o and Llama4-Maverick. GPT-4o sometimes took slight liberties with generally correct info to give a more complete answer, whereas Maverick was strictly correct but occasionally incomplete. The ideal system in the legal domain likely needs a combination of both traits. We need the answer to be as complete as possible and completely faithful. One possible way to achieve this is a hybrid approach. First, use a conservative model to gather all facts, ensuring nothing false is introduced, then use a more expressive model to rephrase or expand as needed. As another approach, further fine-tuning could potentially push a model like Maverick to increase relevancy without losing faithfulness. The current data shows a slight trade-off between relevancy and faithfulness for some models, which is an interesting point for future research.

4.3. Qualitative Evaluation with Case Examples

To qualitatively analyze GraphRAG performance, we compared its outputs against a traditional vector-based RAG on representative legal questions.

Case 1: Hierarchical legal reasoning. *Question:* "Apa saja syarat sahnya suatu perjanjian menurut KUH Perdata?" **Traditional RAG:** Retrieved only `Article 1320`, providing a partial answer. **GraphRAG:** Traversed `BELONGS_TO` and `HAS_ENTITY` edges to aggregate related nodes (*kesepakatan*, *kecakapan*, *hal tertentu*, *sebab yang halal*), delivering a complete and well-structured answer referencing each article.

Case 2: Controlled uncertainty. *Question:* "Siapa yang dapat menuntut pemisahan harta perkawinan?" **Traditional RAG:** Generated a speculative answer using general pretraining knowledge. **GraphRAG:** Accurately detected the absence of evidence in the context and responded conservatively: "Informasi tidak ditemukan dalam konteks hukum yang diberikan." This demonstrates higher faithfulness through explicit uncertainty handling.

Case 3: Cross-section contextualization. *Question:* "Bagaimana hubungan antara Buku II dan Buku III KUH Perdata?" **GraphRAG:** Followed `PART_OF` and `SIMILAR` relationships to describe the thematic continuity between property law and obligation law, a reasoning step impossible with flat retrieval.

These cases show that GraphRAG excels at hierarchical reasoning, explicit reference tracking, and conservative factual grounding — qualities essential for legal QA where precision and accountability are critical.

5. Limitations

This study has several limitations that should be acknowledged. First, the dataset is limited to 2,128 articles from the Indonesian Civil Code, which may not fully capture the diversity of real-world legal queries. Second, the evaluation is based on a fixed set of 66 questions, which may not represent the full complexity of legal reasoning tasks. Third, the benchmarking focuses on zero-shot usage without additional fine-tuning, which may disadvantage smaller models. Finally, while RAGAs metrics provide useful automatic evaluation, they may not fully capture nuanced legal correctness, which would require expert human validation. Future work should address these limitations by expanding dataset coverage, incorporating more diverse queries, and including human evaluation protocols.

6. Conclusion

In conclusion, this study demonstrates that GraphRAG serves as an effective framework for enhancing legal question answering in the Indonesian context. By leveraging a

graph-structured knowledge base in Neo4j, the system improves answer faithfulness and contextual grounding—key factors in applying LLMs to high-stakes domains such as law. Among the evaluated models, Llama4-Maverick achieved the best overall performance, showing higher scores than GPT-4o in faithfulness and context entity recall within this experimental setting. These results highlight the growing potential of open-source models to rival proprietary systems when coupled with structured retrieval mechanisms. Beyond the performance results, this research introduces the Indonesian Legal GraphRAG as a valuable resource and benchmarking framework for future studies. The demonstrated feasibility of combining large language models with graph-based retrieval opens opportunities for developing more transparent and trustworthy AI assistants across specialized knowledge domains.

References

1. Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; et al. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2411.15594* **2024**.
2. Nasution, A.H.; Onan, A. ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks. *IEEE Access* **2024**, *12*, 71876–71900. https://doi.org/10.1109/ACCESS.2024.3402809.
3. Hidayat, F.; Nasution, A.H.; Ambia, F.; Putra, D.F.; Mulyandri. Leveraging Large Language Models for Discrepancy Value Prediction in Custody Transfer Systems: A Comparative Analysis of Probabilistic and Point Forecasting Approaches. *IEEE Access* **2025**, *13*, 65643–65658. https://doi.org/10.1109/ACCESS.2025.3560254.
4. Nasution, A.H.; Monika, W.; Onan, A.; Murakami, Y. Benchmarking 21 Open-Source Large Language Models for Phishing Link Detection with Prompt Engineering. *Information* **2025**, *16*. https://doi.org/10.3390/info16050366.
5. Nasution, A.H.; Onan, A.; Murakami, Y.; Monika, W.; Hanafiah, A. Benchmarking Open-Source Large Language Models for Sentiment and Emotion Classification in Indonesian Tweets. *IEEE Access* **2025**, *13*, 94009–94025. https://doi.org/10.1109/ACCESS.2025.3574629.
6. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435* **2023**.
7. Ratnaningsih, I.D.A.S.; Dewi, C.I.D.L. Sahnya Suatu Perjanjian Berdasarkan Kitab Undang-Undang Hukum Perdata. *Jurnal Risalah Kenotariatan, [S. l.], v. 5, n. 1, p. 11–18, 2024. DOI: 10.29303/risalahkenotariatan.v5i1.204.* **2024**.
8. Faisal, D.; Darari, F.; Ryanda, R. Granularity-aware legal question answering: a case study of Indonesian government regulations. *International Journal of Advances in Intelligent Informatics* **2024**, *10*, 359–378. https://doi.org/10.26555/ijain.v10i3.1105.
9. Redelaar, F.; Van Drie, R.; Verberne, S.; De Boer, M. Attributed Question Answering for Preconditions in the Dutch Law. In Proceedings of the Proceedings of the Natural Legal Language Processing Workshop 2024; Aletras, N.; Chalkidis, I.; Barrett, L.; Goanță, C.; Preoțiuc-Pietro, D.; Spanakis, G., Eds., Miami, FL, USA, 2024; pp. 154–165. https://doi.org/10.18653/v1/2024.nllp-1.12.
10. Sansone, C.; Sperlí, G. Legal Information Retrieval systems: State-of-the-art and open issues. *Information Systems* **2022**, *106*, 101967. https://doi.org/https://doi.org/10.1016/j.is.2021.101967.
11. Wiggers, G. The relevance of impact: bibliometric-enhanced legal information retrieval **2023**.
12. Amazou, Y.; Tayalati, F.; Mensouri, H.; Azmani, A.; Azmani, M. Accurate AI Assistance in Contract Law Using Retrieval-Augmented Generation to Advance Legal Technology. *International Journal of Advanced Computer Science & Applications* **2025**, *16*.
13. Ni, B.; Liu, Z.; Wang, L.; Lei, Y.; Zhao, Y.; Cheng, X.; Zeng, Q.; Dong, L.; Xia, Y.; Kenthapadi, K.; et al. Towards Trustworthy Retrieval Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2502.06872* **2025**.
14. 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.; Kuo, C.F.; et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digital Medicine* **2025**, *8*, 187.
15. Barron, R.C.; Eren, M.E.; Serafimova, O.M.; Matuszek, C.; Alexandrov, B.S. Bridging Legal Knowledge and AI: Retrieval-Augmented Generation with Vector Stores, Knowledge Graphs, and Hierarchical Non-negative Matrix Factorization. *arXiv preprint arXiv:2502.20364* **2025**.
16. Pipitone, N.; Alami, G.H. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343* **2024**.
17. Peng, B.; Zhu, Y.; Liu, Y.; Bo, X.; Shi, H.; Hong, C.; Zhang, Y.; Tang, S. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921* **2024**.
18. Bruckhaus, T. Rag does not work for enterprises. *arXiv preprint arXiv:2406.04369* **2024**.
19. Procko, T.T.; Ochoa, O. Graph retrieval-augmented generation for large language models: A survey. In Proceedings of the 2024 Conference on AI, Science, Engineering, and Technology (AIxSET). IEEE, 2024, pp. 166–169.

20. Bahr, L.; Wehner, C.; Wewerka, J.; Bittencourt, J.; Schmid, U.; Daub, R. Knowledge graph enhanced retrieval-augmented generation for failure mode and effects analysis. *Journal of Industrial Information Integration* **2025**, *45*, 100807. https://doi.org/https://doi.org/10.1016/j.jii.2025.100807.
21. Abu-Salih, B. Domain-specific knowledge graphs: A survey. *Journal of Network and Computer Applications* **2021**, *185*, 103076. https://doi.org/https://doi.org/10.1016/j.jnca.2021.103076.
22. Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Metropolitansky, D.; Ness, R.O.; Larson, J. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130* **2024**.
23. Hu, Y.; Lei, Z.; Zhang, Z.; Pan, B.; Ling, C.; Zhao, L. Grag: Graph retrieval-augmented generation. *arXiv preprint arXiv:2405.16506* **2024**.
24. Zhang, Q.; Chen, S.; Bei, Y.; Yuan, Z.; Zhou, H.; Hong, Z.; Dong, J.; Chen, H.; Chang, Y.; Huang, X. A Survey of Graph Retrieval-Augmented Generation for Customized Large Language Models. *arXiv preprint arXiv:2501.13958* **2025**.
25. Dong, Y.; Wang, S.; Zheng, H.; Chen, J.; Zhang, Z.; Wang, C. Advanced RAG Models with Graph Structures: Optimizing Complex Knowledge Reasoning and Text Generation. In Proceedings of the 2024 5th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC). IEEE, 2024, pp. 626–630.
26. Masoudifard, A.; Sorond, M.M.; Madadi, M.; Sabokrou, M.; Habibi, E. Leveraging Graph-RAG and Prompt Engineering to Enhance LLM-Based Automated Requirement Traceability and Compliance Checks. *arXiv preprint arXiv:2412.08593* **2024**.
27. Shahriar, S.; Lund, B.D.; Mannuru, N.R.; Arshad, M.A.; Hayawi, K.; Bevara, R.V.K.; Mannuru, A.; Batool, L. Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency. *Applied Sciences* **2024**, *14*. https://doi.org/10.3390/app14177782.
28. Islam, R.; Moushi, O.M. Gpt-4o: The cutting-edge advancement in multimodal llm. *Authorea Preprints* **2024**.
29. Khalila, Z.; Nasution, A.H.; Monika, W.; Onan, A.; Murakami, Y.; Radi, Y.B.I.; Osmani, N.M. Investigating Retrieval-Augmented Generation in Quranic Studies: A Study of 13 Open-Source Large Language Models. *International Journal of Advanced Computer Science and Applications* **2025**, *16*. https://doi.org/10.14569/IJACSA.2025.01602134.
30. Es, S.; James, J.; Anke, L.E.; Schockaert, S. Ragas: Automated evaluation of retrieval augmented generation. In Proceedings of the Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2024, pp. 150–158.