Assessing RAG to Retrieve Information from Construction Contracts

Ramesh Balaji[1], Murali Jagannathan, Ph.D.,[2] and Mateja Durovic Ph.D.[3]

[1]Ph.D Scholar, Indian Institute of Technology Madras, Chennai, 600036; email: ramesh.balaji@dsai.iitm.ac.in

[2]Assistant Professor, Indian Institute of Technology Madras, Chennai, 600036; email: muralij@civil.iitm.ac.in

[3]Professor, King's College London, United Kingdom, WC2R 2LS; email: mateja.durovic@kcl.ac.uk

ABSTRACT:

Due to time constraints, construction contract parties often skip thorough contract risk assessments, leading to costly disputes during project execution. While Large Language Models (LLMs) can aid slow manual analysis, their reliance on domain-specific data limits effectiveness. This study explores Retrieval Augmented Generation (RAG) techniques with LLMs like GPT4 and Llama2-70B. Testing 38 models on two contracts, the study found 25% to 76% accuracy. Challenges included contextual ambiguities and domain-specific meanings, suggesting that high accuracy requires tailored RAG-LLM combinations to address construction contract complexities effectively.

AUTHOR KEYWORDS:

Construction; Contracts; AI; LLM; LM; RAG; Risk; RAG

INTRODUCTION:

Construction contracts govern employer-contractor relationships, but time constraints during bidding often prevent thorough analysis, leading to risks like disputes and cost overruns. The time available for the complete process until bid submission varies from one project to another, ranging from a few days to a few months. However, the bigger problem is that the bidder is involved in bidding on many other projects simultaneously and cannot devote sufficient time to thoroughly analyse the bid documents (Eken 2022). Lack of sufficient time, coupled with the time-consuming manual risk assessment process (Khalef and El-adaway 2021), results in suboptimal risk reviews, often leading to interpretation differences, conflicts, claims, and disputes during the project execution stage (Iyer 1996). Further, ambiguous clauses and poor risk allocation contribute to delays and blame-shifting. AI tools, like RAG, could improve risk assessment by analysing contracts efficiently, but challenges such as the need for large training datasets and skill gaps persist. This study explores RAG as a promising solution for improving contract management and mitigating risks without requiring extensive training data.

LITERATURE REVIEW AND GAPS:

Choosing LLMs for the study

This study classifies language models (Zhao et al. 2023) into categories of closed (e.g., GPTs (Saka et al. 2023b)) and open (e.g., Zephyr, Llama). Zephyr 7B (Tunstall et al. 2023a) and Llama2-70B (Touvron et al. 2023) were selected for an RAG-based tool based on strong performance and Reinforced Learning with Human Feedback (RLHF) alignment. GPT-4 was chosen as the best-performing closed model and aligned with RLHF (Tunstall et al. 2023a).

Models for Contract Risk Assessment

Automatic assessment of risks in contraction contracts has gained significant traction in the last decade, thanks to advancements in Natural Language Processing (NLP). Initially,

rule-based NLP algorithms were developed to identify problematic areas of construction contracts like the extraction of poisonous clauses (Lee et al. 2019), exculpatory provisions (Padhy et al. 2021), and the assessment of the extent of a party's obligations (Agrawal et al. 2021). Along these lines, (Choi and Lee 2022) used an ontological semantic model and Bi-long short-term memory techniques to assess risks in the bid documents of Engineering, Procurement, and Construction (EPC) projects. However, due to the inherent complexity of the contract document's structure, it was challenging to define a set of comprehensive rules to extract required information from a contract automatically. Additionally, rules were often defined by referring to some well-known standard forms of contract (like the International Federation of Consulting Engineers or FIDIC), limiting the generalizability of such models (Lee et al. 2020; Serag 2010). Realising the need to understand the subtle semantic and syntactic challenges associated with understanding the text used in construction domains (Okonkwo et al. 2023, in the context of information extraction from building regulation documents, evaluated word embeddings and transformers for their ability to identify and extract semantic regularities in domain-specific documents and noted that certain models, when trained and curated with domain-specific information, can indeed aid in successful extraction of data from a given document. Some articles specifically focus on certain aspects of extraction, like extracting party obligations (Al Qady and Kandil 2010) and scheduling requirements (Hassan and Le 2022). Taking a cue from studies in the legal domain (Al Qady and Kandil 2010) developed a rule-based model to extract party obligations, highlighting the need to generalise the model capabilities for various querying styles. Still the Rule based Algorithms are not robust enough to solve the risks which opened the way for domain-specific models (Jayakumar et al. 2023; Zheng et al. 2022), word embeddings, and transformers improved semantic extraction. LLMs and classification models (Chakrabarti et al. 2018; Hassan and Le 2021) aid in summarization (Xue et al. 2022) and contract management (Saka et al. 2023a),

with RAG showing promise for better search and retrieval in construction contracts. Recently, a study in the context of automated compliance checking by (He et al. 2025) proposed using an improved retrieval-augmented generation (RAG) framework to conduct question-answering (QA)-based construction quality checks. The framework contains a novel hybrid search engine that integrates term frequency-inverse document frequency (TF-IDF)-based keyword search with text-embedding search to facilitate domain semantic-aware regulation information extraction. Another recent article employed LLM and RAG to assess risks in a construction contract document with an accuracy of 76.7% (He et al. 2024). However, continuous development in RAGs and LLMs can be explored further to understand if improvements could be made.

RAG and Advanced RAG

In Retrieval-Augmented Generation (RAG) (Lewis et al. 2021), contract documents are stored as embedding vectors in a Vector Database. When a query is made, it is converted into an embedding and searched using semantic similarity. Matching document sections are retrieved and provided as context to a Large Language Model (LLM) for a response. The pictorial representation of the RAG process is shown in Figure 1.

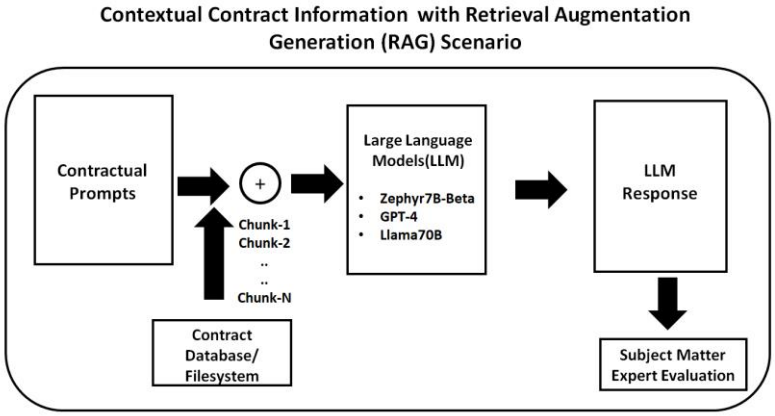

Figure 1: RAG process with manual validation

Recent RAG advancements, including Prompt Techniques (Tang et al. 2024), Constitutional AI (Bai et al. 2022), Contextual Compression (Witteveen n.d.), and Sentence Window (Yang 2023), aim to improve response accuracy. These methods, not yet tested on construction contracts, are considered in the study, making them more sophisticated than basic embedding-based retrieval.

Research Gaps

AI and Machine Learning have potential in construction contract management, but limited research and publicly available training data (Choi and Lee 2022; Zheng et al. 2022) hinder progress (Ghimire et al. 2023). Rule-based models (Choi and Lee 2022) face challenges due to contract drafting styles, while RAG models offer flexibility by retrieving relevant information without predefined rules, improving accuracy. Certain provisions in construction contracts may pose risks when linked to other clauses, but current AI struggles to identify these interconnected risks (Abdul-Malak and Jamileh 2019). Further, implicit contract risks arise from missing details (Lee et al. 2020, 2023; Serag 2010), leading to disputes. Existing studies focus on standard contracts, but bespoke contracts may not fit these templates. Moreover, a challenge in developing a user-question-based retrieval algorithm for contract risk assessment is the variability in how users phrase questions and how contractual provisions are structured, depending on their experience, expertise, and jurisdiction. RAG, by design, should accommodate this variability in prompts and contexts. Finally, ambiguity detection (Wu et al. 2022) in construction contracts is crucial but challenging, involving both word-level (Anish et al. 2019) and document-wide ambiguities. Current research emphasizes document comparison (Candaş and Tokdemir 2022; Roshnavand et al. 2019; Zhang and Ma 2023), not single-document detection. While fully automated tools are still developing, an information retrieval tool could simplify ambiguity detection, but further studies are needed to assess its effectiveness for contract analysis. RAG can potentially assist users in overcoming the above

gaps. However, RAG and its advanced variations, combined with LLMs, have not received much attention in assessing risks in construction contracts.

METHODOLOGY:

A six-step methodology is followed in this study to generate answers to a set of user questions on a given contract document.

Step 1: Create a set of user questions with human-generated answers

The General Conditions of Contract (GCC) document of a public sector firm involved in Indian railway infrastructure projects is analysed. A set of 20 questions covering key contract provisions is created. Two of these questions are repeated with slight modifications to test the model's sensitivity to phrasing. The correct answers are manually recorded to ensure consistency when comparing responses.

Step 2: Testing LLM robustness without a context

The 20 questions are fed as zero-shot prompts (zero-shot prompts are questions fed into the LLM without providing any context to understand if the LLM is pre-trained with adequate data to respond without the need for any context (contract document), just by mentioning the name of the contract document in the prompt, followed by user questions.

Step 3: RAG

In this RAG stage, the reference contract document will be queried based on the questions developed in the previous stage. The generated responses are tabulated against the respective questions.

Step 4: Choosing the best LLM

As discussed earlier in the literature review section, three LLMs (Zephyr7B Beta, GPT4, and Llama2-70B) are tested in this study. **Step 5: Automatic evaluation**

In this step, the validation process is automated using a validation LLM, which compares the model-generated responses with human-generated ones. A score of '1' is given for matches based on correctness, accuracy, and factualness, and '0' for mismatches. The ideal score is 20, reflecting the 20 questions. GPT-4 is used for evaluation, but manual validation is retained to compare results and discuss differences between the two methods.

Step 6: Improving the RAG output

The study employs four techniques—Prompt Techniques, Prompt Techniques with Constitutional AI, Prompt Techniques with Contextual Compression, and Sentence Window—to assess their impact on RAG output accuracy. Consistent parameters include a temperature of 0.1 and a maximum new token of 2048, with top_p and num_beams set to default.

The passage discusses techniques to improve LLM performance for context-specific queries:

- **Prompt Techniques (PT)** ensure context-based responses (Tang et al. 2024).
- **PT with Constitutional AI (PT-CAI)** (Bai et al. 2022) uses predefined guidelines for improved response quality.
- **PT with Contextual Compression (PT-CC)** (Witteveen n.d.) focuses on relevant information by compressing dense text.
- **Sentence Window (SW)** (Yang 2023) enhances retrieval by adding contextual sentences.

The study investigates two questions: 1) Whether 20 questions represent contract risks, and 2) How document changes affect answer accuracy. It tests various LLMs and RAG techniques

with 20 questions, validating results against a larger set of 87 questions. Further validation uses a second contract (EPC), applying the same questions and techniques as GPT-4 and RAG + PT.

RESULTS:

The step-wise results are presented in this section.

Step 1: Create a set of user questions with human-generated answers.

Considering the article word and page limits, a sample set of two questions (out of 20 used in the study), along with human-generated answers, is shown in Table 1.

Table 1. Selected questions and expected human-generated answers

Serial No	Questions	Expected Responses
1	When should the contractor submit the Bank Guarantee for performance?	Within 21 days from the date of the Letter of Acceptance. (Clause Reference: 16.4 (a))
13	Is the time of essence even if the Employer accords sanction according to clauses 17 and 17A?	The contract is silent about the time being the essence of the contract when the Employer accords sanction according to clauses 17 and 17A. (Clause Reference: 17B)

Step 2: Testing LLM robustness without a context

The summary of non-RAG extraction against the 2 questions in Table 1 across 3 LLMs is tabulated in Table 2.

Table 2. Results of the Non-RAG Extraction Attempt

Technique	Question	Zephyr7B-Beta	GPT-4	Llama2
Non-RAG	When should the contractor submit the Bank Guarantee for performance?	The contractor shall submit the Bank Guarantee for performance to the Engineer at least thirty (30) days before the commencement of the work.	The contractor should submit the Bank Guarantee for performance typically before the start of the project or as stipulated in	The contractor should submit the Bank Guarantee for performance when the contract is awarded and before the commencement of work. This ensures that the contractor has the

			the contract agreement.	necessary financial resources to fulfil the obligations of the contract and provides the project owner with a level of assurance that the contractor will complete the work as agreed.
	Is the time of essence even if the Employer accords sanction according to clauses 17 and 17A?	17 and 17A are the clauses which are applicable in case of transfer of an officer from one zone to another zone. In such cases, the officer is entitled to the next increment in the new zone. But, the time of essence is not extended in such cases. The officer has to serve the full period of 3 years in the new zone to become eligible for the next promotion.	The answer would depend on the specific terms and conditions outlined in clauses 17 and 17A. Generally, the concept of "time is of the essence" in contracts means that the specified times for performance are considered vital and failure to perform within the specified time may be considered a breach of contract.	Yes, the time is of the essence even if the Railways accord sanction according to clauses 17 and 17A. The time limits specified in the agreement are essential and must be strictly adhered to. Failure to comply with the time limits may result in the cancellation of the agreement and the forfeiture of the performance security.

Step 3: RAG

The results obtained in this step are tabulated in Table 3, showing the RAG-based responses to the two questions presented in Table 1.

Table 3. Selected questions and responses – RAG

Questions	**Expected Responses**	**Zephyr7BBeta-Responses**	**GPT4-Responses**	**Llama2-70b-Responses*)***

When should the contractor submit the Bank Guarantee for performance?	Within 21 days from the date of the Letter of Acceptance. (Clause Reference: 16.4 (a))	The contractor should submit the Bank Guarantee for performance within 21 days from the date of issue of Letter of Acceptance (LOA), as outlined in Sub-Clause 16. (4)(b) of the clause…..	The successful bidder or contractor should submit the Performance Guarantee within 21 days from the date of issue of the Letter of Acceptance (LOA)…..	The contractor should submit the Bank Guarantee for performance within 21 days from the date of issue of Letter of Acceptance (LOA)….
Is the time of essence even if the Employer accords sanction according to clauses 17 and 17A?	The contract is silent about the time being the essence of the contract when the Employer accords sanction according to clauses 17 and 17A. (Clause Reference: 17B)	No, the time of essence is not necessarily applicable even if the Employer accords sanction according to clauses 17 and 17A. …	The provided text does not provide specific information on whether the time of essence applies even if the Employer accords sanction according to clauses 17 and 17A.	Please select one of the following options: Yes No I don't know.

Step 4: Choosing the best LLM

Based on a manual evaluation of LLMs, the score for each LLM is shown in Table 4 under the respective columns titled “Manual Score.”

Table 4. Question-wise response for the initial run with RAG

Questions	**Zephyr7B-Beta**			**GPT-4**			**Llama2-70B**		
	GPT 4 score	**Manual Score**	**Difference (Mismatch)**	**GPT4 Score**	**Manual Score**	**Difference (Mismatch)**	**GPT4 Score**	**Manual Score**	**Difference (Mismatch)**
1	1	1	0	1	1	0	1	1	0
2	0	0	0	0	0	0	0	0	0
3	1	1	0	1	1	0	0	0	0
4	1	1	0	1	1	0	0	1	***-1***
5	0	1	***-1***	1	1	0	0	1	***-1***
6	1	1	0	1	1	0	1	1	0
7	1	1	0	1	1	0	1	1	0
8	0	0	0	1	1	0	0	1	***-1***

9	0	1	***-1***	0	1	***-1***	0	1	***-1***
10	0	0	0	0	0	0	0	0	0
11	1	1	0	1	1	0	1	1	0
12	1	1	0	1	1	0	1	1	0
13	1	1	0	1	1	0	0	0	0
14	0	0	0	0	0	0	0	0	0
15	0	0	0	1	0	***1***	1	0	***1***
16	1	0	***1***	1	1	0	0	1	***-1***
17	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0
TOTAL	**9**	**11**	**Count: 3**	**12**	**12**	**Count: 2**	**6**	**10**	**Count: 6**

Step 5: Automatic evaluation

The results of the automatic evaluation of the extraction results, considering GPT-4 as the validation LLM, are shown in the column (titled GPT4 Score) of Table 4. Although the total algebraic difference is only "two," mismatches were observed in three cases. In some instances, the automatic evaluation scored ‘1’ while the manual evaluation gave ‘0’, and vice versa. The positive and negative differences cancel each other out, so tracking the exact number of mismatches is crucial.

Step 6: Improving RAG output

As discussed earlier, four methods (PT, PTCAI, PT-CC, SW) are tried to improve the solutions generated from various LLMs. The llama2-70 B model was discontinued for further analysis for two reasons. Firstly, for the given task, the accuracy of the Llama2-70B model was lower compared with the other two LLMs. Secondly, the large size of the Llama2-70B model (70 billion parameters) demanded significant computing resources compared to the other models. Instead of Llama2-70B, an LLM rated next best to Zephyr7B per the MT-Bench score, the MistralAI Mixture of Experts (MoE) model is chosen (Tunstall et al. 2023b). The MoE concept is an ensemble learning technique initially developed within artificial neural networks.

It introduces the idea of training experts on specific subtasks of complex predictive modelling problems (Sanseviero et al. 2023). Results are tabulated in Table 5.

Table 5. Automatic evaluation of RAG with improvement techniques (numbers in the brackets indicate manual evaluation, mismatch instances)

Technique	Zephyr7B-Beta	GPT-4	Mistral7B_MOE
RAG	9 (11,3,55%)	12 (12,2,60%)	12 (6,8,30%)
RAG + PT	10 (10,2,50%)	14 (11,3,55%)	14 (12,2,60%)
RAG + PT-CAI	12 (10,2,50%)	13 (11,4,55%)	12 (11,1,55%)
RAG + PT-CC	8 (5,3,40%)	6 (6,2,30%)	8 (5,3,25%)
RAG + SW	14 (13,1,65%)	14 (13,1,65%)	11 (11,0,55%)

To check the sensitivity of the responses to the window size in the case of SW, the initial window size was set at three and steadily improved to 12. All along, it has been seen that the results have improved. However, beyond 12, there is a reduction in output quality; thereby, the window size is restricted to 12. The results of the variation in window size (for sizes 12 and above) are shown in Table 6. Results from the improved techniques from Table 5 show that the SW technique performed well in the case of Zephyr and the GPT models, whereas the PT worked well in the case of GPT and the Mistral models, with the maximum correct score reaching about 13 (out of 20).

Table 6. Automatic evaluation of RAG with varying window sizes in SW (numbers in the brackets indicate manual evaluation, mismatch instances, and assessment accuracy considering manual evaluation)

Technique	Zephyr7B-Beta	GPT-4	Mistral7B_MOE
SW – (Size 12)	14 (13,1,65%)	14 (13,1,65%)	11 (11,0,55%)
SW – (Size 15)	10 (8,4,40%)	12 (12,1,60%)	12 (8,6,40%)
SW – (Size 20)	9 (8,1,40%)	14 (13,1,65%)	11 (9,6,45%)

The results of operations on the larger dataset (87 questions) are presented in Table 7 shown below:

Table 7. Results of operations on the larger dataset (87 questions), numbers in the brackets indicate manual evaluation, mismatch instances, and assessment accuracy considering manual evaluation.

Technique	Zephyr7B-Beta	GPT-4	Mistral7B_MOE

RAG + PT	58 (51,15,59%)	61 (50,11,57%)	57 (60,11,69%)
RAG + PT-CAI	67 (57,14,66%)	-	-
RAG + PT-CC	58 (49,9,56%)	-	56 (39,25,45%)
SW – (Size 12)	73 (66,7,76%)	-	-
SW – (Size 15)	65 (57,12,66%)	-	-
SW – (Size 20)	68 (56,12,64%)	-	-

Finally, the analysis of the second document, as shown in Table 8, indicates lower accuracy levels than the first, with correct responses ranging from 6 to 10 out of 20, while the best case in the first document was 13 (based on manual evaluation results). Further investigation revealed that five questions typically answered correctly in the first document had wrong answers in the second. These questions related to bank guarantees, deviation limits, price variation clauses (2 questions), and non-excusable delays.

Table 8. Results of operations on the second contract document (20 questions, numbers in the brackets indicate manual evaluation, mismatch instances, and assessment accuracy considering manual evaluation)

Technique	Zephyr7B-Beta	GPT-4	Mistral7B_MOE
RAG + PT	9 (10, 5, 50%)	10 (7, 3, 35%)	9 (6, 3, 30%)
RAG + PT-CAI	-	10 (7, 5, 35%)	-
RAG + PT-CC	-	5 (5, 2, 25%)	-
SW – (Size 12)	-	7 (6, 1, 30%)	-
SW – (Size 15)	-	6 (6, 5, 30%)	-
SW – (Size 20)	-	8 (6, 4, 30%)	-

DISCUSSION:

The study first tested a non-RAG approach, where LLMs pre-trained on general data were queried without domain-specific input. The results were inaccurate, with several models hallucinating answers. Hallucination refers to LLMs generating confident but false or fabricated information (Ghimire et al. 2023), such as the Zephyr-7B-Beta model creating a non-existent procedure for a bank guarantee submission. Other instances involved overly general answers, such as implications of employer delays, rather than contract-specific details. This confirms that without context, LLMs may be unreliable for contract analysis. Further testing

on a larger dataset showed that none of the model responses matched human-generated answers.

Strengths of RAG

The accuracy of responses improves significantly when using RAG (Retrieval-Augmented Generation) techniques compared to traditional methods. RAG enhances the model's ability to generate accurate answers by utilizing context and semantic similarity, which prevents hallucinations seen in non-RAG models. RAG outperforms keyword-based and ontology-based models in three key ways.

First, it expands beyond specific keywords by leveraging vector-based cosine similarity, enabling it to capture related and similar words. For example, it correctly identifies the deadline for a Performance Bank Guarantee, even without the word "deadline," and links "quantity variation" to "quantity deviation." Second, RAG extracts entire paragraphs, combines them with related sections, and summarizes the output effectively. For instance, when asked about the supply of construction power and water, RAG successfully pulled information from adjacent provisions and summarized it. Third, in cases of missing information, RAG-based models can infer details from related content, such as correctly identifying the role of the "Employer" based on contract context. Additionally, RAG facilitates holistic interpretation by extracting related elements spread across different sections. In one case, it identified a comprehensive definition of the term "drawing" by analysing two related definitions from different parts of the contract. This holistic approach allows RAG to provide a more accurate and comprehensive interpretation than simple search tools. Overall, RAG-based models offer substantial advantages in extracting, interpreting, and summarizing contract details, though consistent answers were observed for repeat questions across models, regardless of correctness.

However, there are observations that need further attention while developing RAG-based models for construction contract risk assessment.

RAG Limitations: Observations

Observation 1: Variation in Model Outputs

The accuracy of answers depends on the combination of RAG techniques and LLMs. Zephyr7B gave correct provisions but included irrelevant details, while Mistral7B-MOE with Sentence Window (SW) provided clearer answers. GPT-4 and Llama2-70B performed more reliably. These results highlight the sensitivity of model performance, suggesting further research to optimize LLM and RAG combinations for consistency.

Observation 2: RAG's Dependence on Availability of 'Similar' Words

The study highlights challenges with using RAG for analysing construction contracts. RAG's reliance on word similarity limits its effectiveness when terms are used in different contexts, such as with clauses about timely decision-making and deviation limits. It also struggled with varying terminology, like "performance guarantee" vs. "performance security." The tool failed to provide correct answers when contract terminology differed, showing the limitations of RAG in handling diverse language and context-specific issues. The study emphasizes that RAG needs refinement to handle the variability in construction contract terminology and improve accuracy in context-based extractions.

Observation 3: Overlooking Certain Legally-Sensitive Words While Summarisation

The new approach must not overlook crucial details, as shown by an instance where the tool omitted the word "conclusive" from "final and conclusive" in a contract, altering its legal meaning. The word "final" alone doesn't emphasize the finality of decisions as strongly, potentially affecting court challenges (Iyer and Satyanarayana 2001). Another issue arose with

models failing to properly identify "conditions precedent" for claims, often confusing them with impact or triggers for claims. Similarly, in loss mitigation cases, all models, except Mistral7B-MOE, incorrectly extracted indemnification provisions instead of relevant contractual obligations. These limitations highlight the importance of accurate legal interpretation during summarization.

Observation 4: Hallucinations are still an issue

Recalling the research gaps and the suitability of RAG to address the gaps, the initial hypothesis that with RAG, there is a lesser possibility of hallucination still holds good when compared with the non-RAG models. Notwithstanding the improvement, few cases of hallucinations are reported. In one of the responses extracted using the Zephyr7B-PT-CC (the question was on conditions precedent for the contractor's claim submission), the model includes provisions on conditions precedent, which cannot be found in the contract considered in this study. In the same model, hallucination is evident in the response to identifying 'collaborative' clauses in contracts (question 17). Hallucination is also seen in one of the responses (question 20) from Zephyr7B-PT. In construction contracts, such instances severely undermine the model outputs' reliability.

The study highlights improvements in response quality with RAG techniques, showing significant gains with SW and PT models over plain RAG. However, CC reduces quality due to compressed context. SW's larger context better aligns with construction contracts' "whole document" rule. Increasing the window size beyond 12 does not improve accuracy due to the spread of relevant information. The study also notes challenges with incomplete or unreliable extractions when clauses vary across contexts. New techniques like HippoRAG may offer more efficient contract risk analysis, especially for holistic interpretations.

Observation 6: Limitations in Ambiguity Detection

The models perform well at detecting word-level ambiguity but fail to identify sentence-level ambiguity. In one instance, a clause specifying the precedence of technical specifications over drawings conflicted with another clause granting the employer final authority on interpretations, yet the tool missed this ambiguity. Similarly, in a second contract document, differing notice requirements for force majeure claims (15 days in one clause, 10 days in another) went undetected by the models, highlighting their inability to recognize such discrepancies and interpret conflicting provisions accurately. This suggests limitations in identifying contextual or logical ambiguities.

Observation7: LLM's Evaluation Inconsistencies

The study found inconsistent evaluations across models, with GPT4 and Mistral7B-MOE-PT offering varying answers. Despite occasional wrong extracts, some summaries matched expected answers, emphasizing the importance of context retrieval. While RAG is effective for quick information extraction, the highest accuracy was 13 out of 20, highlighting the need for further research, "Human-in-loop" involvement, and domain-specific AI tools to enhance LLMs for contract risk analysis.

RAG Limitations: bringing new risks

The key point of discussion here is the propensity of RAG-based risk assessment tools to introduce any new risks. If the contract manager is fully dependent on the RAG output, then, owing to the inherent limitations observed in the study, it is possible that certain risks are ignored, and that can cause disputes on a future date. It can also happen that hallucinations can result in flagging certain risks that may actually not be true, and this can lead to bidders considering additional contingencies, which in the case of the public sector tendering process can reduce the chances of bid success (owing to the concept of selection of the technically qualified least quoted bidder). Therefore, more research is required to improve the output

accuracy, and this can be achieved by bringing in some new techniques, like Retrieval Augmented Fine-Tuning (RAFT), Graph RAG, Agentic RAG, Finetuning and using the combination of a custom Contract Risk Ontology with RAG.

The overall inferences on the assessment of RAG models are summarised in Table 9.

Table 9. Results of RAG assessment

Gap	**RAG's strengths**	**Areas requiring improvement**
Need for the difficult-to-obtain elaborate training data to train the Language Models to extract answers to the user-defined questions accurately	RAG-based models answer user-defined questions with an accuracy of 65 to 75%. While this number could vary, depending on the document, it is still a good output given the ease of developing RAG-based Q&A tools	Contract risk analysis requires a nearly 100% output accuracy to avoid disputes arising from seemingly trivial issues. This means that in their current form, RAG models cannot fully replace the manual reading of contract documents.
Supervised models to retrieve information from the contract document are typically trained on individual paragraphs, not necessarily connected ones, preventing document analysis 'as a whole.'	RAG models can extract relevant pieces of information from different parts and then summarise the output in a concise manner	In the case of interconnected provisions without a semantical connection, the models do not capture the holistic meaning. Further, the summary falls short of a holistic interpretation when the information is spread over many clauses (more than two or three, like clauses on indemnification, delay events, etc.).
Extant studies assessing the implicit risks in a given contractual provision often use standard forms like FIDIC to identify missing information. In the absence of such reference documents, the tool may not be useful in identifying implicit risks	If information is missing to answer the user's question, the models highlight the absence of relevant information in the uploaded document.	While the models have been able to extract information from the given text, cases of hallucination are evidenced in the event of missing information. Domain-specific training seems inevitable to rein in such behaviors.
Both user questions and Construction contract document contents can vary to a high degree, making extraction challenging	As long as the terminologies used in the question and the uploaded document are commonly followed irrespective of the	- Although RAG is designed to extract relevant chunks from the document, the relevancy is more from a 'language' perspective, which may limit the model's capability when

	contract document (like indemnification, force majeure etc.), the models extracted correct answers from both documents. If the user has prior knowledge of the possible terms used in the given contract document, then the user can use appropriate terms in the question, thereby improving the output quality.	there are no 'similar' or 'equivalent' words or similar words used in a different context. - Some questions that yielded a correct answer in certain RAG models, did not generate a correct answer in others. - The accuracy of the RAG output depends on the terminologies in the question posed and the document contents. In construction contracts, the variability in both the questions posed and document contents is quite high, increasing the probability of incorrect outputs.

CONTRIBUTION AND LIMITATIONS:

In this age of rapidly advancing AI technology, domain researchers have an important role in understanding the extent to which the developing technology is applicable in solving domain-specific problems, which is the very objective of this research. The study contributes to the body of knowledge by evaluating the robustness of the RAG technique in specific construction contract management applications. The specific shortcomings highlighted in the study can help drive focused studies to analyse construction contract documents.

In terms of the contribution to the body of practice, it informs practitioners of the importance of their involvement and collaboration with data scientists to develop tools that can be widely applied in the construction industry. Since most of the observations stem from the fact that every contract document is drafted and structured in a unique manner, to tap into the potential of training/finetuning-independent quick risk assessment tools (as discussed in this study), practitioners should try to shift to the use of internationally acceptable standard form

contracts, to the extent practically possible. Modifications to suit the specific site/project requirements could be brought in through special conditions of the contract. If this is not possible, the way out is to provide domain training to LLMs trained on a general corpus. However, to achieve this, the industry should come together to create a large data repository that can be used for training/finetuning purposes.

Notwithstanding the contributions, the study is not without its limitations. Firstly, more public sector construction contract documents can be analysed to understand RAG better. Moreover, while this study focuses on RAG with selected LLMs, future studies can compare RAG with several non-RAG-based retrieval methods and combine other LLMs. Also, the authors manually evaluated the answers provided by the models, and future studies may consider external validation. Next, this study does not attempt role prompting; future studies can account for this. Although the study highlights various issues associated with RAG-based solutions, since the scope of this study is limited to assessing the effectiveness of RAG in contract risk assessment, no specific solutions are proposed/developed. Lastly, while open-source models' role in assuaging privacy fears is briefly touched upon, further studies on how such models impact privacy requirements can be tested as a part of a separate study.

CONCLUSION AND FUTURE DIRECTIONS:

The study highlights the effectiveness of RAG-based tools in construction contract risk assessment, addressing challenges like context and ambiguity. It suggests collaboration between experts and data scientists to standardize contracts, create data repositories, and optimize RAG techniques and LLM combinations for improved accuracy and interpretation across jurisdiction

The study discusses insufficient finetuning data for LLMs in construction contract management. It compares RAG and improved-RAG techniques, finding they align automatic evaluations with manual assessments but can't replace manual review. Future research should address contract complexities, with collaboration between the industry and researchers.

ACKNOWLEDGMENT:

We acknowledge the support of the Indian Institute of Technology Madras-King's College London Partnership Collaboration Awards in conducting this study.

REFERENCES:

Abdul-Malak, M.-A. U., and M. H. Jamileh. 2019. "Proposed Framework for the Rendering of Construction Contract Document Interpretations by Engineering Professionals." *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, 11 (3): 1–13. https://doi.org/https://doi.org/10.1061/(ASCE)LA.1943-4170.0000305.

Agrawal, A. K., M. Jagannathan, and V. S. K. Delhi. 2021. "Control Focus in Standard Forms: An Assessment through Text Mining and NLP." Journal of Legal Affairs and Dispute Resolution in Engineering and Construction, 13 (1): 1–11. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000441.

Al Qady, M., and A. Kandil. 2010. "Concept Relation Extraction from Construction Documents Using Natural Language Processing." J Constr Eng Manag, 136 (3).

Anish, P. R., A. Sainani, N. Ramrakhiyani, S. Pawar, G. K. Palshikar, and S. Ghaisas. 2019. "Towards Disambiguating Contracts for their Successful Execution-A Case from Finance Domain." *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*. Macao, China.

Bai, Y., S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. El Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan. 2022. "Constitutional AI: Harmlessness from AI Feedback." *ArXiv*.

Chakrabarti, D., N. Patodia, U. Bhattacharya, I. Mitra, S. Roy, J. Mandi, N. Roy, and P. Nandy. 2018. "Use of Artificial Intelligence to Analyse Risk in Legal Documents for a

Better Decision Support." *Proceedings of TENCON 2018 - 2018 IEEE Region 10 Conference*, 0683–0688. Jeju, Korea: IEEE.

Choi, S. W., and E. B. Lee. 2022. "Contractor's Risk Analysis of Engineering Procurement and Construction (EPC) Contracts Using Ontological Semantic Model and Bi-Long Short-Term Memory (LSTM) Technology." *Sustainability (Switzerland)*, 14 (11). MDPI. https://doi.org/10.3390/su14116938.

Eken, G. 2022. "Using natural language processing for automated construction contract review during risk assessment at the bidding stage." Doctoral. Middle East Technical University.

Ghimire, P., K. Kim, and M. Acharya. 2023. "Generative AI in the Construction Industry: Opportunities & Challenges." *arXiv Preprint*. Macao, China, .

Hassan, F. ul, and T. Le. 2021. "Computer-assisted separation of design-build contract requirements to support subcontract drafting." *Autom Constr*, 122. Elsevier B.V. https://doi.org/10.1016/j.autcon.2020.103479.

Hassan, F. ul, and T. Le. 2022. "Extraction of Activities Information from Construction Contracts Using Natural Language Processing (NLP) Methods to Support Scheduling." Construction Research Congress 2022, 773–781. ASCE.

He, Y., Y. Tang, and T. Chen. 2024. "A Study on Large Language Model-Based Approach for Construction Contract Risk Detection." *ACM International Conference Proceeding Series*, 136–141. Association for Computing Machinery.

Iyer, K. C. 1996. "Identification and evaluation of dispute-prone clauses in Indian construction contracts."

Jayakumar, T., F. Farooqui, and L. Farooqui. 2023. "Large Language Models are legal but they are not: Making the case for a powerful LegalLLM." *ArXiv*.

Khalef, R., and I. H. El-adaway. 2021. "Automated Identification of Substantial Changes in Construction Projects of Airport Improvement Program: Machine Learning and Natural Language Processing Comparative Analysis." Journal of Management in Engineering, 37 (6). American Society of Civil Engineers (ASCE). https://doi.org/10.1061/(asce)me.1943-5479.0000959.

Lee, J., Y. Ham, J.-S. Yi, and J. Son. 2020. "Effective Risk Positioning through Automated Identification of Missing Contract Conditions from the Contractor's Perspective Based on FIDIC Contract Cases." *Journal of Management in Engineering*, 36 (3). American Society of Civil Engineers (ASCE). https://doi.org/10.1061/(asce)me.1943-5479.0000757.

Lee, J., J.-S. Yi, and J. Son. 2019. "Development of Automatic-Extraction Model of Poisonous Clauses in International Construction Contracts Using Rule-Based NLP." *Journal of Computing in Civil Engineering*, 33 (3): 1–13. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000807.

Lee, S. H., S. W. Choi, and E. B. Lee. 2023. "A Question-Answering Model Based on Knowledge Graphs for the General Provisions of Equipment Purchase Orders for Steel

Plants Maintenance." *Electronics (Switzerland)*, 12 (11). MDPI. https://doi.org/10.3390/electronics12112504.

Okonkwo, O., A. Dridi, and E. Vakaj. 2023. "Leveraging Word Embeddings and Transformers to Extract Semantics from Building Regulations Text." 11th Linked Data in Architecture and Construction - LDAC2023, 176–188. Matera: CEUR Workshop Proceedings.

Padhy, J., M. Jagannathan, and V. S. K. Delhi. 2021. "Application of Natural Language Processing to Automatically Identify Exculpatory Clauses in Construction Contracts." Journal of Legal Affairs and Dispute Resolution in Engineering and Construction, 13 (4): 1–9. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000505.

Saka, A. B., L. O. Oyedele, L. A. Akanbi, S. A. Ganiyu, D. W. M. Chan, and S. A. Bello. 2023a. "Conversational artificial intelligence in the AEC industry: A review of present status, challenges and opportunities." *Advanced Engineering Informatics*. Elsevier Ltd.

Saka, A., R. Taiwo, N. Saka, B. Salami, S. Ajayi, K. Akande, and H. Kazemi. 2023b. "GPT Models in Construction Industry: Opportunities, Limitations, and a Use Case Validation." 1–58. https://doi.org/https://doi.org/10.48550/arXiv.2305.18997.

Sanseviero, O., L. Tunstall, P. Schmid, S. Mangrulkar, Y. Belkada, and P. Cuenca. 2023. "Mixture of Experts Explained." *Hugging Face*. Accessed January 10, 2024. https://huggingface.co/blog/moe.

Serag, E. 2010. "Semantic detection of risks and conflicts in construction contracts." *Proceedings of the CIB W78 2010: 27th International Conference*, 1–6.

Tang, C., T. Mohati Tahmineh, M. Nayeb, S. Wang, and H. Hemmati. 2024. "Prompt Engineering or Fine Tuning: An Empirical Assessment of Large Language Models in Automated Software Engineering Tasks." *ArXiv*, (1).

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. 2023. "LLaMA: Open and Efficient Foundation Language Models." *ArXiv*.

Tunstall, L., E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, N. Sarrazin, O. Sanseviero, A. M. Rush, and T. Wolf. 2023a. "Zephyr: Direct Distillation of LM Alignment." *arXiv Preprint*.

Tunstall, L., E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, N. Sarrazin, O. Sanseviero, A. M. Rush, and T. Wolf. 2023b. "Zephyr: Direct Distillation of LM Alignment." *arXiv Preprint*.

Witteveen, S. n.d. "samwit/langchain-tutorials." Accessed January 15, 2024a. https://github.com/samwit/langchain-tutorials.

Witteveen, S. n.d. "samwit/langchain-tutorials."

Wu, L. T., J. R. Lin, S. Leng, J. L. Li, and Z. Z. Hu. 2022. "Rule-based information extraction for mechanical-electrical-plumbing-specific semantic web." *Autom Constr*. Elsevier B.V.

Xue, X., Y. Hou, and J. Zhang. 2022. “Automated Construction Contract Summarization Using Natural Language Processing and Deep Learning.” *39th International Symposium on Automation and Robotics in Construction*, 459–466.

Yang, S. 2023. “Advanced RAG 01: Small-to-Big Retrieval.” *Towards Data Science*.

Zhao, W. X., K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen. 2023. “A Survey of Large Language Models.” *arXiv Preprint*.

Zheng, Z., X. Z. Lu, K. Y. Chen, Y. C. Zhou, and J. R. Lin. 2022. “Pretrained domain-specific language model for natural language processing tasks in the AEC domain.” *Comput Ind*, 142. Elsevier B.V. https://doi.org/10.1016/j.compind.2022.103733.