KG-MuLQA: A Framework for KG-based Multi-Level QA Extraction and Long-Context LLM Evaluation

Nikita Tatarinov,✉ Vidhyakshaya Kannan,* Haricharana Srinivasa,* Arnav Raj, Harpreet Singh Anand, Varun Singh, Aditya Luthra, Ravij Lade, Agam Shah,✉ Sudheer Chava

Georgia Institute of Technology
✉ Corresponding Authors: {ntatarinov3, ashah482}@gatech.edu
* Indicates equal contribution

Abstract

We introduce KG-MuLQA (**K**nowledge-**G**raph-based **Mu**lti-**L**evel **Q**uestion-**A**nswer Extraction): a framework that (1) extracts QA pairs at multiple complexity levels (2) along three key dimensions – multi-hop retrieval, set operations, and answer plurality, (3) by leveraging knowledge-graph-based document representations. This approach enables fine-grained assessment of model performance across controlled difficulty levels. Using this framework, we construct a dataset of 20,139 QA pairs based on financial credit agreements and evaluate 16 proprietary and open-weight Large Language Models, observing that even the best-performing models struggle with set-based comparisons and multi-hop reasoning over long contexts. Our analysis reveals systematic failure modes tied to semantic misinterpretation and inability to handle implicit relations.

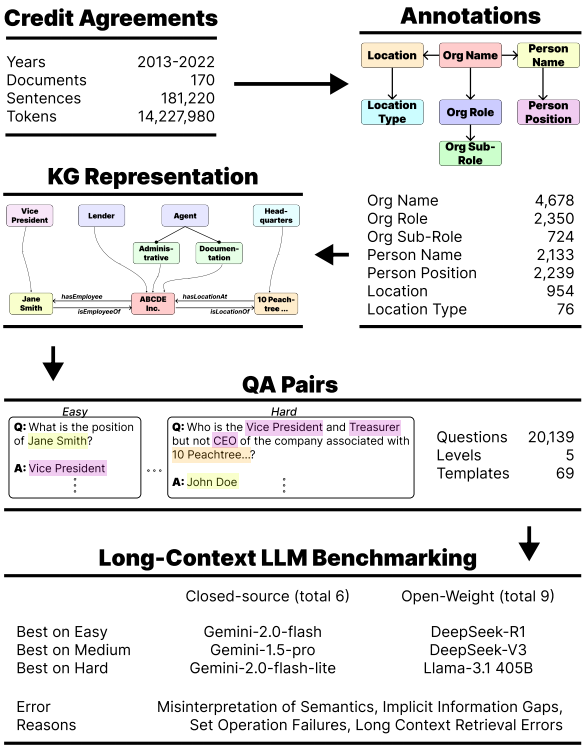


Figure 1: Overview of KG-MuLQA. Credit agreements are annotated to identify entities and their relationships, forming a knowledge graph representation. This graph is then used to *systematically extract multi-level QA pairs*, which serve as the basis for benchmarking long-context LLMs.

1 Introduction

The increasing context length of recent Large Language Models (LLMs) has led to growing interest in evaluating their capabilities. Researchers have studied challenges like the *Lost in the Middle* problem (Liu et al., 2024b), examined the limiting factors of scaling models to long context such as the constraints imposed by RoPE's base value (Men et al., 2024), and developed techniques to extend context length efficiently (Hooper et al., 2024; Ding et al., 2024). Others have focused on improving inference efficiency (Jiang et al., 2024b; Tang et al., 2024) or enhancing long-context utilization and mitigating information loss (Zhang et al., 2024d; Lin et al., 2024; Zhang et al., 2024c). Table 1 highlights numerous benchmarks developed to evaluate long-context LLMs on different tasks.

We introduce the **K**nowledge-**G**raph-based **Mu**lti-**L**evel **Q**uestion-**A**nswer Extraction (**KG-MuLQA**) framework, outlined in Figure 1. Our approach uses knowledge-graph document representation to *programmatically* extract question-answer pairs of varying complexity *from gold annotations*. We define *question templates* and determine their complexity through three dimensions: the number of hops, the use of set operations, and the plurality of the answer. This structured approach ensures scalability, flexibility, and controlled dataset construction, enabling a more precise evaluation of long-context LLMs.

We apply KG-MuLQA to 170 credit agreements, resulting in **KG-MuLQA-D**, a dataset of 20,139 QA pairs categorized by five complexity

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3323–3359

Table 1: Comparison of KG-MuLQA-D with existing long-context LLM benchmarks. ***Multi-hop reasoning*** marks whether the dataset requires reasoning across multiple evidence pieces; KG-MuLQA-D explicitly defines and scales multi-hop paths through its knowledge-graph structure, enabling controlled evaluation of relational inference depth. ***Systematic Q-Extraction*** denotes whether questions are generated deterministically rather than manually or by LLMs; KG-MuLQA-D automates QA generation from annotated graphs, preserving real document semantics without synthetic injection. ***Question complexity*** shows whether the benchmark categorizes questions by difficulty; KG-MuLQA-D introduces a principled, multi-dimensional notion of complexity (hops + set operations + plurality), allowing granular performance analysis across reasoning levels.

Dataset	Long context	Multi-hop reasoning	Systematic Q-extraction	Question complexity	Document Type	# Samples
L-Eval (An et al., 2024)	✔	✘	✘	✘	Earnings Calls + Misc	2,008
CLongEval (Qiu et al., 2024)	✔	partially	✘	✘	THUnews + Misc	7,267
Marathon (Zhang et al., 2024a)	✔	partially	✘	✘	Prior Datasets	1,530
MMLongBench-DOC (Ma et al., 2024)	✔	partially	✘	✘	arXiv + Misc	1,082
DocFinQA (Reddy et al., 2024)	✔	partially	✘	✘	SEC 10-K & 10-Q	7,437
NOCHA (Karpinska et al., 2024)	✔	✔	✘	✘	Fictional Books	2,002
LooGLE (Li et al., 2024)	✔	✔	✘	✘	arXiv + Misc	6,448
BAMBOO (Dong et al., 2024)	✔	✔	✘	✘	ACL Anthology + Misc	1,502
LongBench (Bai et al., 2024)	✔	✔	✘	✘	Previous Datasets	4,750
Loong (Wang et al., 2024)	✔	✔	✘	✘	SEC + Misc	1,600
FinDVer (Zhao et al., 2024b)	✔	✔	✘	✘	SEC 10-K & 10-Q	2,100
M4LE (Kwan et al., 2024)	✔	✘	partially	✘	Prior Datasets	unspecified
Michelangelo (Vodrahalli et al., 2024)	✔	partially	partially	✘	synthetic	unspecified
∞Bench (Zhang et al., 2024b)	✔	partially	needles	✘	Novels + Misc	3,946
mLongRR (Agrawal et al., 2024)	✔	✔	needles	✘	BBC News	unspecified
(Gupta et al., 2024)	✔	✘	✘	partially	Financial News	560
DocMath-Eval (Zhao et al., 2024c)	✔	partially	✘	partially	SEC 10-K & 10-Q	4,000
FanOutQA (Zhu et al., 2024)	✔	✔	✘	partially	Wikipedia	7,305
RULER (Hsieh et al., 2024)	✔	✔	needles	partially	SEC 10-K & 10-Q	30,000
BABILong (Kuratov et al., 2024)	✔	✔	partially	partially	Books + Misc	60,000
MuSiQue (Trivedi et al., 2022)	✘	✔	✔	✔	Wikipedia	24,814
KG-MuLQA-D (ours)	✔	✔	✔	✔	SEC Credit Agreements	20,139

levels. We use it to evaluate 16 long-context LLMs and find that even on simple factual queries, some models fail to extract basic information; performance further drops with multi-hop paths and questions involving set operations. Our analysis shows that set-based comparisons are especially challenging, with high rates of "Not Found" responses and semantic misinterpretation.

Our contributions are as follows. (1) We present a framework based on knowledge graphs, enabling systematic extraction of QA pairs at varying complexity levels. (2) We release a portion of the constructed dataset[1] to encourage future work (see Appendix G for details). (3) We release our inference pipeline and benchmarking codes[2] to facilitate reproducible evaluation of long-context models.

[1] https://huggingface.co/datasets/gtfintechlab/KG-MuLQA-D

[2] https://github.com/gtfintechlab/KG-MuLQA

2 The KG-MuLQA Benchmark

KG-MuLQA systematically extracts structured QA pairs across varying complexity levels for long-context evaluation. Using credit agreements (Section 2.1) annotated with a structured label schema (Section 2.2, Appendix A), we construct document-level knowledge graphs (Section 2.3) to enable systematic query construction. Questions are designed along three dimensions – multi-hop reasoning, set operations, and answer plurality (Section 2.4) – ensuring scalable, diverse, and automated question extraction (Section 2.5). The dataset balances complexity and document coverage (Appendix A.6), making it a robust long-context LLM benchmark.

2.1 Credit Agreements

We sampled 17 credit agreements per year from 2013 to 2022 from SEC EDGAR, totaling 170 documents. Table 2 includes their length by token and sentence count. Their suitability for long-context LLM benchmarking is detailed in Appendix A.1.

2.2 Annotations

Our annotation schema captures key entities and relationships in credit agreements, defining seven label types covering persons, organizations, roles, and locations to represent their structure. We outline label definitions in Appendix A.2, placement rules and relationship guidelines in Appendix A.3, while Table 2 includes the frequency statistics for each entity and relation type across documents. Each document was independently annotated by 3 annotators. Between the second and third (final) rounds of annotating, the inter-annotator agreement (Cohen's Kappa) reached 80.76%, indicating substantial consistency (see Appendix A.5 for annotation process details).

2.3 Knowledge Graph Representation

To enable systematic and scalable QA extraction, we transform annotated credit agreements into knowledge graph representations. These graph encode entities of people, organizations, roles, and locations, along with their relationships, in a structured and queryable form. Figure 2 illustrates a graph which captures that `Jane Smith` holds the position of `Vice President` at `ABCDE Inc.` – an organization that serves as a `Lender` as well as an `Administrative` and `Documentation` `Agent`, and has its `Headquarters` at `10 Peachtree...`. By abstracting away document-specific phrasing while preserving these logical links, the graphs support consistent interpretation across agreements and form the basis for creating questions of varying complexity through structured queries.

We construct knowledge graphs using the Resource Description Framework[3] (RDF), which represents data as subject-predicate-object triples, forming a directed labeled graph for knowledge representation. Stored in Turtle (*.ttl) format for readability, this structure ensures scalability and efficient querying across documents using SPARQL[4] (see Appendix B for KG construction details).

2.4 Question Templates and Complexity Dimensions

We begin by identifying all possible 1-hop QA templates from the knowledge graph, illustrated in Figure 2: questions formed from direct relationships between two entities, such as "*What is the position of* [`Person Name`]?" or "*Who is the representative of* [`Org Name`]*?*". These base templates are constrained to yield singular answers, forming the simplest form of retrieval. It should be noted that not all templates are universally applicable – for instance, the latter question is only valid when the document specifies exactly one representative for the given organization. We then systematically expand these templates along three dimensions.

[3] https://www.w3.org/TR/rdf11-concepts/
[4] https://www.w3.org/TR/sparql11-query/

Table 2: Statistics of the source documents and their structured annotations used to construct KG-MULQA-D. The table summarizes token and sentence counts, as well as the distribution of annotated entities and relations per document. The low minimum values are due to an edge case discussed in Appendix A.4.

Stats per doc (total 170 docs)	Min	Avg	Max
Documents			
Tokens	32,562	83,694	233,207
Sentences	470	1,066	3,541
Entities			
`Org Name`	2	27.52	390
`Org Role`	1	13.82	52
`Org Sub-Role`	0	4.26	35
`Person Name`	0	12.55	88
`Person Position`	0	13.17	88
`Location`	0	5.61	168
`Location Type`	0	0.45	7
Relations			
`Org Name` → `Org Role`	1	21.17	79
`Org Role` → `Org Sub-Role`	0	3.74	32
`Org Name` → `Person Name`	2	12.55	88
`Org Name` → `Person Position`	0	13.17	88
`Person Name` → `Person Position`	0	12.93	86
`Org Name` → `Location`	0	6.32	178
`Location` → `Location Type`	0	0.45	18

Number of Hops Increasing retrieval depth via intermediate nodes, e.g., "*Who is the* [`Person Position`] *of the organization located at* [`Location`]*?*".

Set Operations Introducing comparisons across entities using set logic, e.g., "*What are the positions held by* [`Person Name A`] *but not* [`Person Name B`] *or* [`Person Name C`]*?*".

Plurality Moving from expecting a singular answer to allowing multiple valid answers, e.g., "*What companies are the* [(`Org Sub-Role` +) `Org Role`] *in the agreement?*".

2.5 Extraction of QA Pairs from Templates

Table 3 illustrates how questions are systematically constructed by incrementally increasing complexity across three dimensions introduced in Section

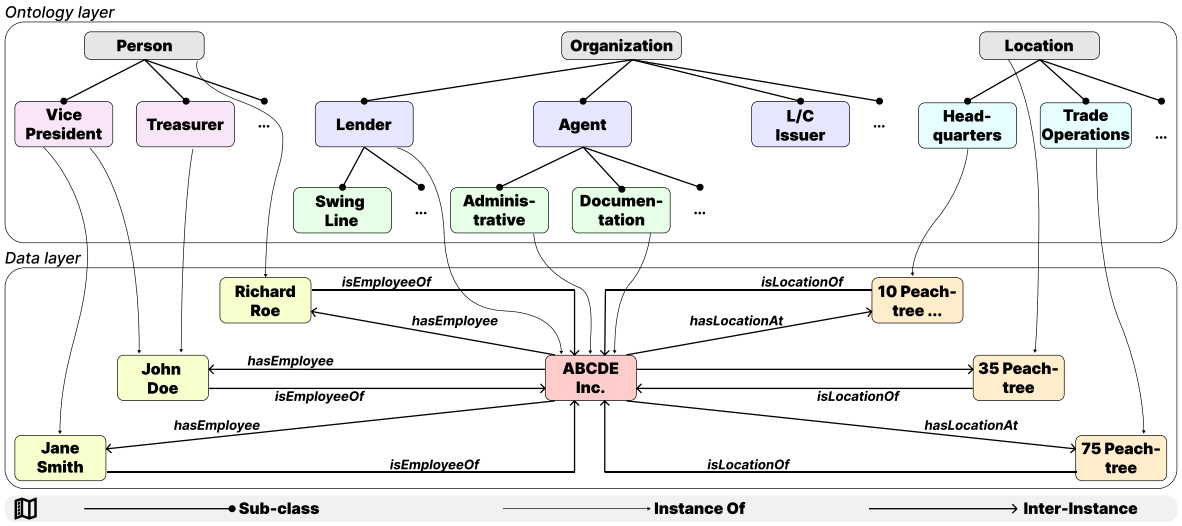


Figure 2: Knowledge graph representation of annotated credit agreements. The ontology layer defines high-level entity types and their sub-classes (e.g., Agent, Lender), while the data layer contains document-specific instances connected via labeled relations (e.g., hasEmployee, isLocationOf). This abstraction supports systematic QA extraction across documents by removing company-specific constraints. All entity names in this figure, including individuals, organizations, and addresses, are fictional and used for illustrative purposes only.

2.4: number of hops (H), plurality (P), and number of set operations (#SO). Each row in the table corresponds to a specific question template, with associated values for P, H, and #SO, the query path in a KG, and the set operation logic used.

Question instances are constructed by dynamically substituting entities into these templates using SPARQL queries over the knowledge graph described in Section 2.3. This ensures precise and scalable QA pair creation, ***with answer labels grounded in annotated relationships***. While our framework supports arbitrarily complex logical combinations, we observed that highly compositional templates often become inapplicable to real documents. For instance, there are no individuals in an agreement who hold three distinct positions or no organization with the combination of roles required to fulfill a complex query. Thus, we have to limit template depth since such complex conditions are typically not present in the source documents. The list of all the created templates can be found in Appendix C.

To support evaluation, we define a composite question complexity level $L = H + P + \#SO$, and group QA pairs into three categories. **Easy** ($L = 1$): simple, single-hop, singular-answer queries. **Medium** ($2 \leqslant L \leqslant 4$): questions with moderate complexity via hops, set operations, or plurality. **Hard** ($L = 5$): questions with the most complex combinations of all three dimensions.

The resulting dataset, KG-MuLQA-D, comprises 20,139 QA pairs from 170 credit agreements, surpassing most long-context QA benchmarks in size (Table 1). This scale is achieved through efficient question extraction from structured annotations using reusable templates.

3 Evaluation

In this section, we describe the setup for our experiments. We evaluate multiple long-context LLMs listed in Section 3.1 across different question complexity levels, as defined in Section 2.4. Performance is measured using four metrics detailed in Section D.2. In addition, we conduct human evaluation of LLM responses on a small sample of QA pairs (Section D.3) to discover that F1-score and the LLM-as-a-Judge metrics correlate with human's expectations most. Our prompting strategy is outlined in Section D.1. The results of our level-based evaluation are presented in Table 4, with the analysis provided in Section 4.

3.1 Baselines

We evaluate 16 proprietary and open-weight LLMs on KG-MuLQA-D benchmark.

Proprietary LLMs: Gemini-2.0-Flash & Gemini-2.0-Flash-Lite (Team et al., 2024a), Gemini-1.5-Pro & Gemini-1.5-Flash (Team et al., 2023, 2024a), and GPT-4o & GPT-4o-Mini (Achiam et al., 2023; Hurst et al., 2024).

Table 3: This table illustrates the question templates used to construct KG-MULQA-D, structured along three dimensions: plurality (P), number of hops (H), and set operations (#SO). It includes example templates, corresponding knowledge graph query paths, and logical operations involved. These dimensions are used to compute the overall complexity level for each QA pair. The full list of templates can be found in Appendix C.

P	H	#SO	Example Template	Hop Path	Set Operation
0	1	0	What is the position of [Person Name]?	Person→Position	None
1	1	0	What are the positions of [Person Name]?	Person→Position	None
0	1	1	What position is held by both [Person A] and [Person B]?	Person→Position	$A \cap B$
0	1	2	What position does [Person A] hold that [Person B] doesn't?	Person→Position	$A \cap \neg B$
0	1	3	What position is held by [Person A] but not [Person B] or [Person C]?	Person→Position	$A \cap \neg B \cap \neg C$
1	1	1	What are the positions held by both [Person A] and [Person B]?	Person→Position	$A \cap B$
1	1	2	What are the positions held by [Person A] but not [Person B]?	Person→Position	$A \cap \neg B$
1	1	3	What are the positions held by [Person A] but not [Person B] or [Person C]?	Person→Position	$A \cap \neg B \cap \neg C$
0	2	0	Who is the [Position] of [Org]?	Org→Pos→Per	None
1	2	0	What are the roles of [Org] in the agreement where [Person] is employed?	Per→Org→Role	None
0	2	1	Who is the [Position A] and [Position B] of [Org]?	Org→Pos→Per	$A \cap B$
1	2	1	Who are both [Position A]s and [Position B]s of [Org]?	Org→Pos→Per	$A \cap B$
0	3	0	Who is the [Position] of the company associated with [Location]?	Loc→Or→Ps→Pe	None
1	3	0	Who are the [Position]s of the company associated with [Location]?	Loc→Or→Ps→Pe	None
0	3	1	Who is both [Position A] and [Position B] of the company associated with [Location]?	Loc→Or→Ps→Pe	$A \cap B$
0	3	2	Who is [Position A] but not [Position B] of the company associated with [Location]?	Loc→Or→Ps→Pe	$A \cap \neg B$
0	3	3	Who is [Position A] and [Position B] but not [Position C] of the company associated with [Location]?	Loc→Or→Ps→Pe	$A \cap B \cap \neg C$
1	3	1	Who are both [Position A] and [Position B] of the company associated with [Location]?	Loc→Or→Ps→Pe	$A \cap B$
1	3	2	Who are [Position A]s but not [Position B]s of the company associated with [Location]?	Loc→Or→Ps→Pe	$A \cap \neg B$

Open-Weight LLMs: DeepSeek-R1 (DeepSeek-AI et al., 2025), DeepSeek-V3 (Liu et al., 2024a), GPT OSS 120B (OpenAI et al., 2025), Llama-4-Maverick-17B-128E-Instruct-FP8 (AI, 2024a), Llama-4-Scout-17B-16E-Instruct (AI, 2024b), Llama-3.1-405B-Instruct-Turbo & Llama-3.1-70B-Instruct & Llama-3.1-8B-Instruct (Dubey et al., 2024), Qwen3-235B-Instruct (Yang et al., 2025), Qwen2-72B-Instruct (Yang et al., 2024).

Given that our documents average 80k tokens, we restrict our evaluation to models supporting a context length of at least 128k tokens. Nonetheless, even long-context models require a carefully designed prompting strategy to extract relevant answers effectively (see Section D.1). We made several attempts to evaluate Gemma-2-8k (Team et al., 2024b), Mixtral-Small-32k, and Mixtral-8×22-65k (Jiang et al., 2024a), but these models returned "`Not found`" in the vast majority of the cases, making systematic evaluation infeasible.

For proprietary models and those accessed via Together.AI[5], we utilized the `LangChain`[6] framework, while open-weight models were employed using `vLLM` (Kwon et al., 2023) for efficient inference. We set the temperature to 0.0 to support reproducibility of our results.

[5] https://www.together.ai/

[6] https://www.langchain.com/

3.2 Experiment Setup

Our evaluation setup is designed to handle long financial documents and large batches of questions in a way that fits the context length constraints of modern language models. As seen in Figure 3, we split a document into chunks if it exceeds 128K tokens: most of the evaluated models have context length of 128K tokens, so **it is impossible to evaluate these models without chunking**. For the larger models, we **have to** keep the same chunking strategy to ensure a fair comparison between models (see Section 6 for the effect of chunking on evaluation). Then, we prompt the model with the document (or chunk) and a batch of 50 questions associated with it: long-context evaluation is very costly, so batching allows at most 3 inference runs per document (given almost 150 questions for some documents). This design ensures consistent and scalable evaluation across documents of varying lengths (see Appendix D.1 for further details).

We use four evaluation metrics: F1 score, Edit Distance, Cosine Similarity, and LLM-as-a-Judge. Among these, F1 and LLM-as-a-Judge are highlighted in Table 4 for their complementary strengths: F1 quantifies exact token-level overlap, while LLM-J captures semantic correctness with human-aligned scoring (see Appendix D.2 for details). Both metrics show the highest correlation with human evaluation metrics (see Appendix D.3).

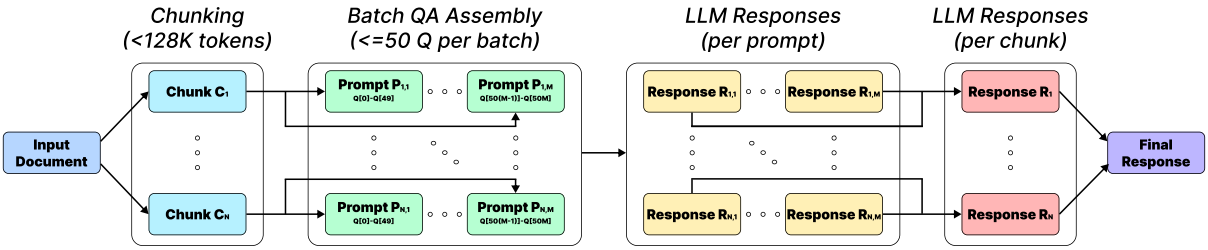


Figure 3: This figure illustrates the multi-stage process used to evaluate long-context LLMs: documents are split into chunks (if exceeding 128K tokens), paired with question batches ($\leqslant$50), and sequentially fed to the model. Model responses are collected per chunk and per question batch to ensure evaluation scalability across large documents.

3.3 Main Results

As seen in Table 4, the results for simple questions are straightforward to interpret: more advanced models mostly get better scores. For medium questions, DeepSeek-V3 outperforming DeepSeek-R1 and Gemini-1.5-Pro outperforming Gemini-2.0 mainly comes from fewer abstentions and "`Not Found`" cases, which raises recall and F1 but doesn't reflect stronger reasoning. Llama-3.1 405B behaves similarly: it answers more often, boosting both metrics through coverage rather than logic. In contrast, Qwen2's high judge score but low F1 reflects a different pattern: it tends to produce verbose, loosely correct text that seems semantically reasonable, so the judge rates it highly even when exact matches are wrong or incomplete. The GPT OSS 120B, however, gets good performance on medium difficulty from its ability to extract entities while minimizing errors, as it strikes a balance between answering and abstaining.

For hard questions on proprietary models, Gemini-2.0-Flash shows high judged quality but lower F1, while Flash-Lite reverses this trend; Gemini-1.5-Pro remains balanced. This difference comes from how often each model answers: more attempts raise F1 through partial overlaps with the gold, while fewer but cleaner attempts earn better judged quality (given responses are not verbose).

For open-weight models, Llama-4-Maverick achieves a high LLM-J score with moderate F1 for answering selectively with short, factual fragments that look semantically right but miss full token matches. Llama-3.1 405B, by contrast, produces longer and more complete answers that align better with references, yielding higher F1 while maintaining strong judged coherence, hence demonstrating the best performance with hard questions.

Overall, as task complexity increases, models split between answering broadly and answering cautiously. This divergence weakens the correlation between F1 and judge scores, since literal overlap and perceived semantic adequacy start reflecting different behaviors.

4 Analysis

4.1 Overall Error Trends

As question complexity increases, the LLM's ability to retrieve and generate correct responses degrades markedly. With increase in number of hops or set operations, the percentage distribution of each error type increases (see Figure 4 for details).

- **Not Found Responses.** Increases from 15.00% at Level 1 to 77.39%, reflecting increasing retrieval difficulty.
- **Low F1 Scores.** Rises from 32.55% to 84.25%, indicating less accurate extraction.
- **Low Cosine Similarity.** Jumps from 38.03% to 92.47%, showing semantic drift.
- **High Edit Distance.** Grows from 40.26% to 95.21%, reflecting surface mismatch.

4.2 Error Analysis

We categorize observed LLM failures into four major types, each of which presents recurring challenges as question complexity increases. For detailed error analysis on examples from each category, see Appendix E, Table 18.

Misinterpretation of Semantics LLMs frequently misinterpret task-specific terms such as "location type" or "position". Rather than grounding responses in document context, models often default to general world knowledge (e.g., answering "city" for a location type), indicating weak comprehension of domain-specific semantics.

Implicit Information Gaps The model struggles to extract answers that are not explicitly stated but implied through document structure, such as signature blocks or formatting cues. For instance, it fails

Table 4: This table presents the performance of 16 LLMs, evaluated across Easy, Medium, and Hard question categories. The metrics include the F1 Score and the LLM-as-a-Judge rating, capturing both token-level accuracy and semantic correctness (see the appendix D.2 for extended evaluation). The results reveal a consistent decline in performance as question complexity increases, with notable model-specific strengths and weaknesses. * denotes the models evaluated on a smaller subset due to cost constraints (see Appendix D.4 for details).

Model	Size	Context Length	Easy		Medium		Hard	
			F1 Score ↑	LLM-as-a-Judge ↑	F1 Score ↑	LLM-as-a-Judge ↑	F1 Score ↑	LLM-as-a-Judge ↑
Proprietary LLMs								
Gemini-2.0-Flash*	–	1M	**0.6748**	**4.0000**	0.3294	2.4820	0.1172	**2.2838**
Gemini-2.0-Flash-Lite*	–	1M	0.5908	3.4172	0.3883	2.6497	**0.1943**	1.5135
Gemini-1.5-Pro*	–	2M	0.5410	3.2448	**0.4577**	**2.8436**	0.1791	1.6986
Gemini-1.5-Flash*	–	1M	0.5103	3.1207	0.3892	2.5130	0.1400	1.5411
GPT-4o*	–	128K	0.5272	3.2034	0.3783	2.5451	0.1178	1.4795
GPT-4o-Mini*	–	128K	0.4723	2.9690	0.2732	2.0681	0.0274	1.1370
Open-weight LLMs								
DeepSeek-R1*	685B	128K	**0.6269**	**3.6965**	0.3293	2.6354	0.2609	2.0736
DeepSeek-V3	671B	128K	0.6218	3.5668	**0.4796**	3.0210	0.0856	1.4795
GPT OSS	120B	128K	0.5885	3.5839	0.4684	**3.1661**	0.1136	1.4888
Llama-4-Maverick-Instruct*	400B	1M	0.5552	3.4482	0.3848	2.6976	0.2031	**3.1351**
Llama-4-Scout-Instruct*	109B	1M	0.4993	3.2138	0.2215	1.8772	0.0472	1.5270
Llama-3.1-Instruct-Turbo*	405B	128K	0.5472	3.4793	0.4204	2.6377	**0.3651**	2.7297
Llama-3.1-Instruct	70B	128K	0.4955	3.1373	0.4017	2.6722	0.1639	2.3425
Llama-3.1-Instruct	8B	128K	0.3891	2.6996	0.2962	2.2332	0.1169	1.4315
Qwen3-Instruct	235B	262K	0.5243	3.2693	0.2760	2.4460	0.1240	1.9578
Qwen2-Instruct	72B	128K	0.3886	3.3884	0.3100	2.9078	0.0762	1.3288

to associate a person with an organization unless the relationship is stated verbatim, ignoring visual or positional indicators that a human reader would readily infer. We consider it a potential reason for why models perform slightly better on full documents than in the oracle setting for hard questions (see Appendix D.6 for details).

Set Operation Failures To examine the impact of the number of set operations on model performance, we conducted an experiment (Figure 11a) by varying one dimension within our three-dimensional parameter space while holding the others constant. Specifically, we increased the number of set operations while keeping the plurality and hop dimensions fixed. We observe that an increase in the number of set operations consistently resulted in a decline in performance metrics. Questions involving comparisons, intersections, or exclusions across entities consistently yield low F1 scores. The model fails to isolate shared or distinct roles when multiple organizations or individuals are involved, indicating limited capacity for multi-hop or set-based reasoning.

Long-Context Retrieval Errors Despite the presence of answers in the document, models often return "`Not Found`", especially when relevant information appears in the middle of the context. This aligns with the "lost-in-the-middle" phenomenon (Liu et al., 2024b), where models disproportionately focus on the beginning or end of long texts.

Our error analysis reveals that while LLMs perform adequately on simple, single-step queries, they falter on multi-hop, set-operation, and implicit-knowledge tasks – particularly those requiring retrieval of details buried in the "middle" of the context. Addressing these shortcomings will likely involve a combination of targeted fine-tuning, improved context chunking or retrieval modules, and stronger in-context reasoning prompts.

5 Related Work

Answering questions about a long document requires, first of all, ***retrieval*** of relevant information piece(-s). Based on the nature of the questions asked, answering them might require ***reasoning*** over the retrieved piece(-s), while reasoning itself can be defined differently across benchmarks. For example, CLongEval (Qiu et al., 2024), ∞Bench (Zhang et al., 2024b), MMLongBench-DOC (Ma et al., 2024) and Michelangelo (Vodrahalli et al., 2024) treat long-context reasoning as synthesizing a conclusion from lengthy inputs, and the former one evaluates it via two task types: ab-

straction (generate content not explicitly in the source) and extraction (identify and copy from the input). MARATHON (Zhang et al., 2024a) frames long-context reasoning as choosing a single correct option from long contexts in a multiple-choice setup with human-verified, misleading distractors. DocFinQA (Reddy et al., 2024) and DocMath-Eval (Zhao et al., 2024c) evaluate long-context numerical reasoning: models find the relevant evidence and carry out multi-step numerical and logical reasoning over text and tables to produce the answer. In our study, we focus on ***multi-hop reasoning***: combining multiple evidence pieces by following multi-step relations between entities to reach an answer. It allows to directly measure a model's ability to track dependencies, integrate facts, and maintain coherence across distant spans: abilities that are central to robust long-context comprehension.

Prior long-context benchmarks typically rely on manual QA construction, which is very costly and time-consuming. RULER (Hsieh et al., 2024), mLongRR (Agrawal et al., 2024), and ∞Bench (Zhang et al., 2024b) all evaluate retrieval by inserting target facts/keys into long distractor contexts and probing whether models can recover or aggregate them at scale. M4LE (Kwan et al., 2024), Michelangelo (Vodrahalli et al., 2024), and BABILong (Kuratov et al., 2024) programmatically synthesize evaluation items, using templates/simulations or latent-structure queries to automatically create large numbers of QA/tasks without manual question writing. We ***deterministically extract*** multi-level QA directly from knowledge-graph representations of annotated documents, varying hops, set operations, and plurality via structured queries. We do not insert synthetic facts, preserving real document distribution, and we do not rely on LLM-written items that would require human verification. MuSiQue (Trivedi et al., 2022) also follows a systematic pipeline, but it first requires manually creating the base single-hop questions and then composing them; to build a new set, those easy questions must be recreated, whereas our pipeline is fully automated from document annotations.

Most existing benchmarks do not define explicit question complexity levels, treating all QA instances as uniformly difficult. Several recent datasets (Gupta et al., 2024; Zhao et al., 2024c; Zhu et al., 2024; Hsieh et al., 2024; Kuratov et al., 2024) vary task difficulty by changing task type or length, distributing supporting facts in long contexts, or altering context length and needle placement. We introduce a ***granular notion of question complexity***, defined along multiple dimensions (multi-hop reasoning, set operations, and answer plurality), allowing systematic control over difficulty and enabling fine-grained error analysis (Section 4.2) that reveals distinct model failure modes. While MuSiQue also organizes questions by reasoning depth (2–4 hops), it lacks the multi-dimensional decomposition we propose. Our formulation extends beyond hop count to capture qualitatively different reasoning types, which provides a richer understanding of model strengths and weaknesses.

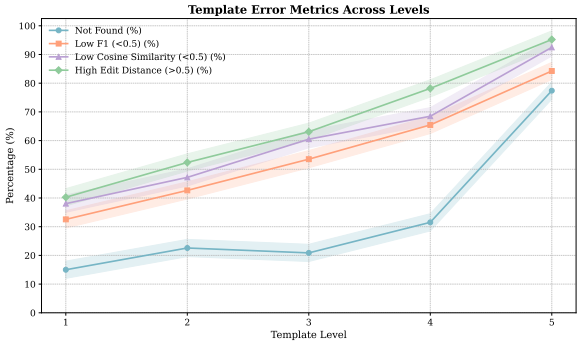


Figure 4: Trends in error types across complexity levels. The template level is given by $P + H + \#SO$.

6 Conclusion & Discussion

Our work introduces KG-MuLQA, a framework that programmatically constructs long-context QA benchmarks using knowledge-graph representations of real-world documents. We begin by annotating credit agreements with a structured schema to capture entities and relations, convert them into RDF-based knowledge graphs, and generate multi-level QA pairs via deterministic SPARQL templates that vary in hop count, set operations, and answer plurality. We then evaluate 16 long-context LLMs across these controlled complexity levels using 4 complementary metrics and perform human evaluation to validate metric alignment. Our quantitative and qualitative analyses reveal consistent weaknesses in multi-hop retrieval, set-based reasoning, and implicit relation grounding, establishing KG-MuLQA as a scalable and interpretable framework for assessing long-context comprehension.

Chunk-based Evaluation As seen in Table 5, the effect of chunking becomes noticeable as the question complexity grows. However, the observed performance increase appears to be driven primarily by a reduction in abstentions, with models more

Table 5: Evaluation of Gemini-2.0-Flash with chunking (as in the main paper) and without chunking. The context length of this model is large enough to incorporate an entire credit agreement in one prompt.

Strategy	F1 ↑	Edit ↓	Cos ↑	LLM-J ↑
		Easy		
With	0.6748	0.2679	0.6650	4.0000
Without	0.7105	0.2987	0.6949	3.9269
		Medium		
With	0.3294	0.6280	0.3121	2.4820
Without	0.5558	0.4051	0.5452	3.4344
		Hard		
With	0.1172	0.7238	0.0608	2.2838
Without	0.2565	0.6310	0.2504	2.1088

often returning explicit answers instead of "Not Found." In our experiments, we do not observe chunking to meaningfully degrade the reasoning abilities of the evaluated models, as the overall patterns of errors and difficulty across complexity levels remain largely consistent. Having said that, models with larger context lengths provide the practical advantage of incorporating entire documents within a single prompt, reducing the need for chunking altogether. Even when this does not substantially improve reasoning performance, such settings tend to encourage models to produce more confident and complete responses, suggesting that continued development of longer-context architectures remains a promising direction for improving real-world long-document QA systems.

Single-Question Evaluation We note that the absolute metric values in Table 6 are not representative of task complexity, as a sampling of 20 questions per complexity level was required due to the high cost of single-question evaluation for long-context documents. Examining relative differences, the single-question evaluation appears to produce modest improvements for easy and medium questions, while the effect is less consistent for harder ones. These differences likely reflect changes in prompting conditions rather than fundamental shifts in model capabilities. When questions are evaluated individually, the model may allocate more focused attention to a single query and retrieve supporting information more directly and reliably, whereas batch evaluation introduces a more constrained prompting context.

Table 6: Evaluation of Gemini-2.0-Flash in batches of 50 questions (as in the main paper) and with a single questions.

Difficulty	F1 ↑	Edit ↓	Cos ↑	LLM-J ↑
		Easy		
Batch	0.5277	0.3773	0.5249	3.7500
Single	0.5719	0.3573	0.5688	3.8000
		Medium		
Batch	0.3980	0.4220	0.3958	3.2040
Single	0.4451	0.3695	0.4427	3.5510
		Hard		
Batch	0.5333	0.4223	0.5333	3.1333
Single	0.4971	0.4367	0.4938	3.0666

Human Performance Evaluation As shown in Appendix D.5, human performance remains stable across difficulty levels. Unlike LLMs, whose performance declines as question complexity increases, human annotators are able to consistently retrieve and combine the relevant information from the documents. This contrast suggests that the observed performance degradation primarily reflects limitations in current models' long-context retrieval and reasoning rather than ambiguity or flaws in the dataset itself. The human baseline therefore supports the validity of the benchmark and indicates that the increasing difficulty levels capture genuine challenges for automated systems rather than artifacts of question construction.

Applicability and Generalizability As discussed in Appendix A.1, the ability to process long documents is essential for real-world QA tasks. Our benchmark illustrates this challenge using credit agreements, where information is densely structured, scattered, and cross-referenced. Legal contracts, policy documents, patent filings, and other non-legal such as medical records or scientific articles all exhibit similarly complex entity structures, making our methodology broadly applicable. To illustrate the applicability and generalizability of our framework, we performed a full-scale case study (see Appendix F) on medical documents, including defining the annotation schema, constructing question templates, etc.

Limitations

The created QA pairs are based on financial credit agreements only, and the types of questions that can be generated are constrained by the predefined annotation schema. The benchmark focuses on relational reasoning via multi-hop retrieval and set operations, excluding other reasoning types such as temporal or numerical reasoning. Finally, as with most existing QA benchmarks, our dataset favors deductive logic because the dataset includes graph-representable questions.

Ethical Considerations

We generally avoid mentioning specific agreements and their details. Instead, we use placeholders such as `ABCDE Inc.` for organization names, `Richard Roe` for person names, and `10 Peachtree...` for locations to illustrate the structure of agreement entities and relations. However, in our annotation guide (Appendix A) and error analysis (Appendix E), we include screenshots of real document annotations and mention specific examples to highlight the complexity of these agreements. Moreover, we store the raw HTML files of the agreements in our GitHub repository, enabling other researchers to reproduce our benchmark or conduct further experiments. As these documents are publicly available through the SEC's EDGAR system, their inclusion for both illustrative and experimental purposes raises no ethical concerns.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ameeta Agrawal, Andy Dang, Sina Bagheri Nezhad, Rhitabrat Pokharel, and Russell Scheinberg. 2024. Evaluating multilingual long-context models for retrieval and reasoning. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 216–231, Miami, Florida, USA. Association for Computational Linguistics.

Meta AI. 2024a. Llama 4 maverick 17b 128e instruct fp8. https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Instruct-FP8. Accessed: 2025-05-14.

Meta AI. 2024b. Llama 4 scout 17b 16e instruct. https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct. Accessed: 2025-05-14.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411, Bangkok, Thailand. Association for Computational Linguistics.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Kathrin Blagec, Georg Dorffner, Milad Moradi, Simon Ott, and Matthias Samwald. 2022. A global analysis of metrics used for measuring performance in natural language processing. *Preprint*, arXiv:2204.11574.

Arun Tejasvi Chaganty, Stephen Mussman, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evaluation. *Preprint*, arXiv:1807.02202.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. LongRoPE: Extending LLM context window beyond 2 million tokens. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11091–11104. PMLR.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2024. BAMBOO: A comprehensive benchmark for evaluating long text modeling capacities of large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2086–2099, Torino, Italia. ELRA and ICCL.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Michael Galarnyk, Veer Kejriwal, Agam Shah, Yash Bhardwaj, Nicholas Watney Meyer, Anand Krishnan, and Sudheer Chava. 2025. Videoconviction: A multimodal benchmark for human conviction and stock

market recommendations. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 5447–5458, New York, NY, USA. Association for Computing Machinery.

Lavanya Gupta, Saket Sharma, and Yiyun Zhao. 2024. Systematic evaluation of long-context LLMs on financial concepts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1163–1175, Miami, Florida, US. Association for Computational Linguistics.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. 2024. Kvquant: Towards 10 million context length llm inference with kv cache quantization. *Preprint*, arXiv:2401.18079.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *Preprint*, arXiv:2404.06654.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024a. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H. Abdi, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024b. Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention. *Preprint*, arXiv:2407.02490.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One thousand and one pairs: A "novel" challenge for long-context language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17048–17085, Miami, Florida, USA. Association for Computational Linguistics.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *Preprint*, arXiv:2406.10149.

Wai-Chung Kwan, Xingshan Zeng, Yufei Wang, Yusen Sun, Liangyou Li, Yuxin Jiang, Lifeng Shang, Qun Liu, and Kam-Fai Wong. 2024. M4LE: A multi-ability multi-range multi-task multi-domain long-context evaluation benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15568–15592, Bangkok, Thailand. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. LooGLE: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333, Bangkok, Thailand. Association for Computational Linguistics.

Hongzhan Lin, Ang Lv, Yuhan Chen, Chen Zhu, Yang Song, Hengshu Zhu, and Rui Yan. 2024. Mixture of in-context experts enhance llms' long context awareness. *Preprint*, arXiv:2406.19598.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Preprint*, arXiv:2407.01523.

Xin Men, Mingyu Xu, Bingning Wang, Qingyu Zhang, Hongyu Lin, Xianpei Han, and Weipeng Chen. 2024. Base of rope bounds context length. *Preprint*, arXiv:2405.14591.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Zexuan Qiu, Jingjing Li, Shijue Huang, Xiaoqi Jiao, Wanjun Zhong, and Irwin King. 2024. CLongEval: A Chinese benchmark for evaluating long-context large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3985–4004, Miami, Florida, USA. Association for Computational Linguistics.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. DocFinQA: A long-context financial reasoning dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 445–458, Bangkok, Thailand. Association for Computational Linguistics.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. 2024. Quest: Query-aware sparsity for efficient long-context llm inference. *Preprint*, arXiv:2406.10774.

Nikita Tatarinov, Siddhant Sukhani, Agam Shah, and Sudheer Chava. 2025. Language modeling for the future of finance: A survey into metrics, tasks, and data opportunities. *Preprint*, arXiv:2504.07274.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024b. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Kiran Vodrahalli, Santiago Ontanon, Nilesh Tripuraneni, Kelvin Xu, Sanil Jain, Rakesh Shivanna, Jeffrey Hui, Nishanth Dikkala, Mehran Kazemi, Bahare Fatemi, Rohan Anil, Ethan Dyer, Siamak Shakeri, Roopali Vij, Harsh Mehta, Vinay Ramasesh, Quoc Le, Ed Chi, Yifeng Lu, and 5 others. 2024. Michelangelo: Long context evaluations beyond haystacks via latent structure queries. *Preprint*, arXiv:2409.12640.

Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. 2024. Leave no document behind: Benchmarking long-context LLMs with extended multi-doc QA. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5627–5646, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Li Yujian and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095.

Lei Zhang, Yunshui Li, Ziqiang Liu, Jiaxi Yang, Junhao Liu, Longze Chen, Run Luo, and Min Yang. 2024a. Marathon: A race through the realm of long context with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5201–5217, Bangkok, Thailand. Association for Computational Linguistics.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024b. ∞Bench: Extending long context evaluation beyond 100K tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277, Bangkok, Thailand. Association for Computational Linguistics.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Ö. Arik. 2024c. Chain of agents: Large language models collaborating on long-context tasks. *Preprint*, arXiv:2406.02818.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and Zhangyang Wang. 2024d. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. *Preprint*, arXiv:2403.04797.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. 2024a. Financemath: Knowledge-intensive math reasoning in finance domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12841–12858, Bangkok, Thailand. Association for Computational Linguistics.

Yilun Zhao, Yitao Long, Tintin Jiang, Chengye Wang, Weiyuan Chen, Hongjun Liu, Xiangru Tang, Yiming Zhang, Chen Zhao, and Arman Cohan. 2024b. FinDVer: Explainable claim verification over long and hybrid-content financial documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14739–14752, Miami, Florida, USA. Association for Computational Linguistics.

Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. 2024c. DocMath-eval: Evaluating math reasoning capabilities of LLMs in understanding long and specialized documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16103–16120, Bangkok, Thailand. Association for Computational Linguistics.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024. FanOutQA: A multi-hop, multi-document question answering benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 18–37, Bangkok, Thailand. Association for Computational Linguistics.

A Annotation Guide

A.1 Suitability of Credit Agreements

Credit agreements are particularly well-suited for long-context evaluation because they are ***rich in entities and relations***, involving numerous organizations, individuals, roles, and locations that appear in varied forms throughout the document, making both retrieval and reasoning challenging. They also exhibit ***ambiguous co-reference***, as entities may be referred to inconsistently (e.g., "Bank of America, N.A." vs. "Bank of America"), demanding precise resolution of abbreviations and partial mentions. While ***structured but not standardized***, these documents follow broad templates (e.g., loan terms, covenants, defaults) but vary greatly in phrasing and layout, preventing reliance on fixed patterns and requiring genuine comprehension. Finally, their ***real-world significance*** as legally binding records of multi-million-dollar transactions makes accurate understanding essential for assessing risk, ensuring compliance, and supporting financial and regulatory decisions. Supporting examples illustrating these aspects are provided in the latter subsections of this Appendix.

A.2 Labels

Our annotation framework captures key entities and relationships within credit agreements, focusing on seven labels that define organizational structure, roles, and locations.

`Organization Name` The name of a company or financial institution mentioned in the agreement. *Examples: "Goldman Sachs", "JP Morgan Chase".*

`Organization Role` The role that an organization holds in the agreement. *Examples: "Lender", "Borrower", "Guarantor".*

`Organization Sub-Role` A more specific classification within an Organization Role that provides additional detail. *Examples: "Lead Lender", "Administrative Agent".*

`Person Name` The name of an individual explicitly mentioned in the agreement. *Examples: "John Doe", "Jane Smith".*

`Person Position` The title or role held by a person within an organization. *Examples: "Vice President", "Loan Officer".*

`Location` A physical address associated with an organization. *Examples: "35 Peachtree St, Atlanta, GA".*

`Location Type` A classification of a Location, providing additional information on its function. *Example: "Headquarters", "Branch Office".*

A.3 General rules

Unique Entity Annotation Organization Names, Person Names, and Locations are annotated only *once per document*, even if they appear multiple times. Organization Roles, Organization Sub-Roles, Person Positions, and Location Types are annotated *once for each unique entity* they describe.

Example: In Figure 5a, "BANK OF AMERICA, N.A." appears multiple times but is annotated only

once. However, its roles, "Administrative Agent" and "Lender", are each annotated where they are explicitly stated. If another organization is a lender and/or an administrative agent, those role labels will be annotated as close as possible to the name label. Similarly for "Alexandra M. Knights" and her position "Authorized Signatory".

Ordered Search for Annotations To maintain consistency, we follow a structured order when searching for annotations.

- *Signature Box*: The signature box is the primary source for annotations, as it usually contains the most organizations (e.g., Parent Borrowers, Lead Lenders) and their roles. In most cases, organizations are explicitly labeled with their roles within this section, making it the most reliable source.
- *Headers*: If an entity is missing in the signature box, we check the document headers. It is common for certain organizations' roles to be absent from the signature box but present in section headings. In some cases, even company names are missing from the signature box and must be retrieved from headers.
- *Full Document*: Locations are often not explicitly mentioned in the signature box or headers, requiring a broader document search. Occasionally, some organizations (e.g., Guarantors) are not mentioned in either the signature box or the headers, making a full-document search necessary.

Relation Annotations

- `Org Name` → `Org Role`: Each organization name is annotated once per document, while its role is annotated as close to the name as possible. If multiple organizations share the same role (e.g., Borrower) and it is mentioned collectively in the signature box, the role is annotated once and linked to all relevant organizations. Otherwise, each occurrence of a role is annotated separately and linked to the corresponding organization.
- `Org Role` → `Org Sub-Role`: If an organization has a specific sub-role (e.g., Parent Borrower), its role is annotated separately from others, and the sub-role is linked to it. The connection follows a hierarchical structure, linking the organization to its role, then the role to its sub-role.
- `Org Name` → `Person Name`: Each person's name is annotated once per document and linked to a single organization (see A.4 for special scenarios).
- `Person Name` → `Person Position`: A person's position is annotated as close to their name as possible. Individuals may hold multiple positions, which are all linked to the same person annotation. In some cases, position mentions may be missing (see Figure 5d and A.4).
- `Org Name` → `Location`: Locations are annotated once per document and linked to the relevant organization.
- `Location` → `Location Type`: Each location has exactly one type, which is annotated as close to the location name as possible. Location types are often missing, and in such cases, no annotation is applied (see Fig 5d).

Continuation Links In cases where the text to be annotated is split across multiple lines or interrupted by symbols, preventing a single coherent annotation, we annotate each segment separately and link them sequentially: the first to the second, the second to the third, and so on. This issue is particularly common for Locations, where addresses or names are fragmented (see Figure 5d). Since our annotation structure does not include relations between labels of the same type, these links are easily identifiable. During post-processing, entities connected by continuation links are automatically merged with spaces into a single coherent label.

A.4 Main Edge Cases

Parent Borrower and Related Entities Some agreements list multiple entities as borrowers and guarantors, often sharing highly similar names (e.g., Summit Hospitality JV, LP, Summit JV MR 2, LLC in Figures 5b and 5c), suggesting a structured division of what may functionally be a single entity. In such cases, one entity is designated as the `Parent Borrower`, while the others are categorized under `Borrowers` and `Guarantors`. The same individual typically signs on behalf of all these entities, but following our general rule, *we annotate a person once* and link them to the `Parent Borrower`. Additionally, since the Parent Borrower's role is often labeled simply as `Parent` in the signature box, while other borrowers explicitly receive the `Borrower` role, we annotate the closest explicit mention of `Borrower` elsewhere in the document to ensure accurate labeling.

Missing Person Position In some cases, individuals are mentioned within address sections rather than as signatories or officials, likely indicating their responsibility for a specific location. These

individuals often lack an explicitly stated position. As shown in Figure 5d, Jim Plocica appears in the address block without a corresponding title. When this occurs, we annotate the Person Name but leave the Person Position unannotated.

Additionally, this example highlights several annotation conventions:

- The Organization Name (Bright Horizons Family Solutions LLC) is omitted here because it is annotated in the signature box (once per document). The links to the Location and Person Name originate from that primary annotation.
- The Location spans multiple lines, requiring the use of Continuation Links (described in Section A.3) to maintain coherence.
- No Location Type is explicitly mentioned, which is a common occurrence in such address sections.

Shared Roles and Prefixes When multiple organizations share a role, they are often referred to collectively using prefixes like Joint or Co- (e.g., Joint Bookrunner or Co-Documentation Agent). However, our annotation framework requires capturing role information at the individual organization level. Since a single entity cannot hold a joint role independently, we do not annotate these prefixes. Instead, we annotate only the core role (Bookrunner, Documentation Agent, etc.) for each company separately, as illustrated in Figure 5e).

Additionally, in this example, the annotations are sourced from the header rather than the signature box because Merrill Lynch, Pierce, Fenner & Smith Incorporated is absent from the signature box, and for other companies, only their primary role (Lender) is mentioned there.

Smallest Document One of the extracted credit agreements involves only 2 organizations, and their roles are mentioned neither in the signature box nor in the header (Figure 5f). Moreover, the role of the lender is not mentioned in the document, while the role of the borrower is mentioned once (Figure 5g).

A.5 Annotation Process Details

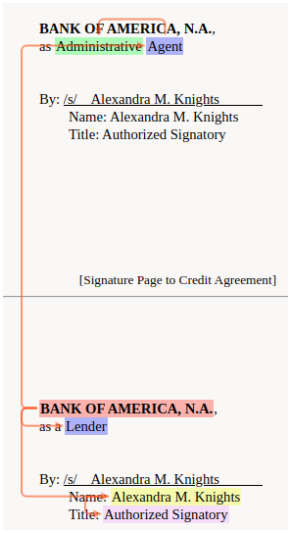
We annotated the extracted credit agreements using Label-Studio[7], supporting entity linking and relation extraction for `.html` files. The annotation was carried out by annotators with backgrounds in Computer Science or Mathematics, under the direct supervision of a Master of Finance and a PhD student advised by a Chair Professor of Finance, also a co-author. The annotation followed a structured, iterative approach to ensure consistency and resolve edge cases. The team first established label definitions and completed an initial annotation round. The results were reviewed to identify and discuss edge cases (Appendix A.4), which led to refinements in the annotation guidelines and the second round of annotations (inter-annotator agreement was not calculated at this point, as the guidelines were evolving). Following the second annotation round, we constructed knowledge-graph (KG) document representations, sampled graphs, and compared them against the ground-truth annotations. Insights from this qualitative assessment (Appendix B.1) informed a third annotation round, after which we applied an additional cleaning script to the KG representations.

A.6 Complexity vs coverage

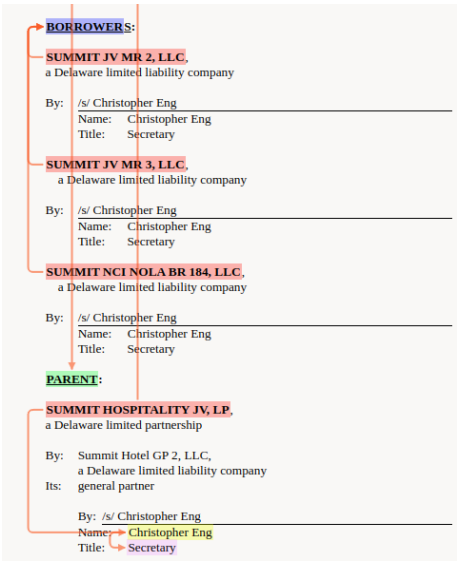
KG-MuLQA defines seven entity and six relation types, and this choice reflects a deliberate trade-off between semantic coverage and compositional complexity. Expanding the ontology would increase topical breadth but would not necessarily yield deeper reasoning requirements. Our goal is to evaluate long-context comprehension under settings where information is distributed across distant parts of a document and must be ***retrieved, composed, and reasoned over***.

Thus, the benchmark emphasizes complex inference chains (multi-hop reasoning, set operations, and answer plurality) over the schema coverage. Answering questions from our dataset requires linking dispersed mentions, disambiguating roles, and integrating partial evidence: operations that jointly test both retrieval and reasoning. This structure allows us to probe failure modes of long-context models in a controlled environment, balancing interpretability, scalability, and difficulty without relying on full document coverage.

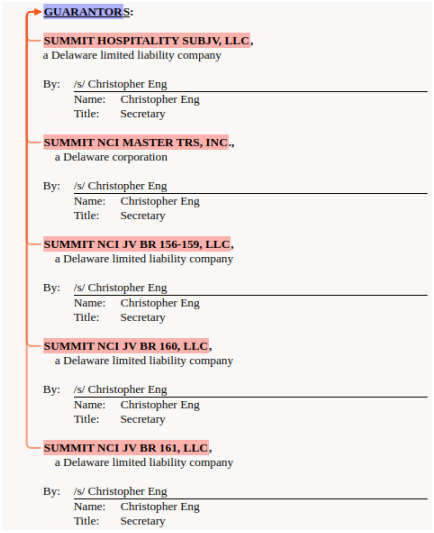
[7] https://github.com/HumanSignal/label-studio



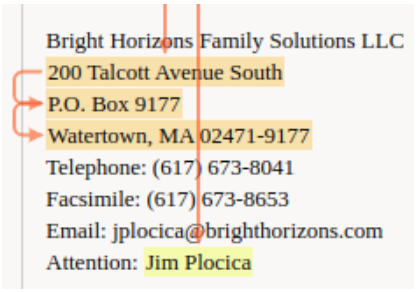
(a) General and most common annotation in the signature box.



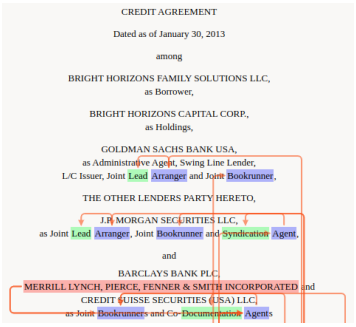
(b) Multiple organizations with similar name being borrowers.



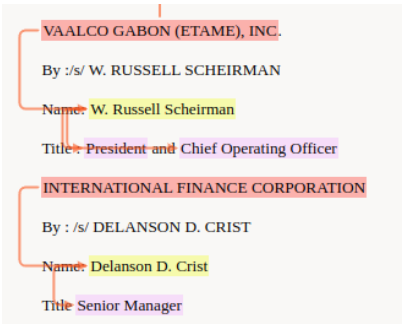
(c) Multiple organizations with similar name being guarantors in addition to being borrowers.



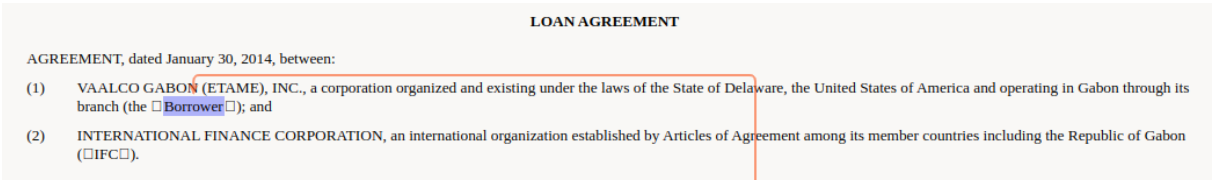
(d) Location without type, person without position and continuation links.



(e) Annotations in the header and joint roles.



(f) Only 2 organizations, not roles in the signature box or header.



(g) Only 2 organizations, only borrower's role is mentioned.

Figure 5: Representative Annotation Scenarios from Label Studio.

B Knowledge Graph Construction Details

Ontology Layer The ontology layer defines a class hierarchy that structures entities in the knowledge graph. We organize annotated entity types into three top-level classes — `Person`, `Organization`, and `Location` — based on their functional roles in financial agreements. These are connected to their respective sub-categories through explicit `Sub-class` relationships, capturing hierarchical distinctions. For example, `Organization` includes roles such as `Lender`, `Agent`, and `L/C Issuer`, while `Person` includes positions like `Vice President` and `Treasurer`. To further refine responsibilities, we introduce a second level of subclasses via `Sub-Roles`, such as dividing `Agent` into `Administrative Agent` and `Documentation Agent`. Only these role and position types appear as classes in the ontology layer; the actual named entities (e.g., "Jane Smith", "ABCDE Inc.") are treated as instances in the data layer. This abstraction supports generalization across documents and enables scalable, template-based QA pairs extraction.

Data Layer Complementing the ontology layer, the data layer grounds the class structure in real-world instances and their relationships. Entities such as `ABCDE Inc.`, `John Doe`, and `10 Peachtree` instantiate ontology classes like `Organization`, `Person`, and `Location` via `Instance Of` links. These instances are further connected through `Inter-Instance` relations (e.g., `hasEmployee`, `isLocationOf`), capturing document-specific associations such as employment and location.

Ontology and Data Scope While each credit agreement yields its own data layer containing document-specific instances and relationships, the ontology layer is shared across all documents. This unified ontology defines a consistent schema for entity types and roles, enabling systematic alignment across diverse agreements.

Question Construction with SPARQL SPARQL queries enable automated, structured QA extraction from our knowledge graph. This approach dynamically adapts to diverse financial agreements by leveraging ontology. The query in Figure 6 retrieves individuals and their positions using the `isInstanceOf` relationship, while `GROUP_CONCAT` consolidates multiple positions per individual for comprehensive QA pairs extraction.

```
SELECT DISTINCT ?person (GROUP_CONCAT(?
    position; separator="|") as ?
    positions)
WHERE {
    ?person a <http://example.org/base/
        Person> ;
            <http://example.org/
                isInstanceOf/> ?position
                .
    FILTER(STRSTARTS(STR(?position), "
        http://example.org/
        person_position/"))
}
GROUP BY ?person
```

Figure 6: SPARQL Query Example for Structured QA Extraction.

B.1 Qualitative Analysis of the Constructed KGs

After generating question-answer pairs from the extracted graphs, we observed two common issues affecting the quality and consistency of the QAs.

- **Duplicate entries:** Multiple questions or answers representing the same entity or relation, differing only by capitalization, whitespace, punctuation, or pluralization.
- **Noisy text:** URL-encoded characters (e.g., `%26`, `%20`) and miscellaneous special characters were present in questions and answers.

To address these issues, we applied a structured cleaning procedure ***after QA generation***, ensuring the graphs themselves were not altered but the extracted QAs were reliable:

- **Text normalization**:
  - Convert all text to lowercase.
  - Remove URL-encoded sequences and special characters, preserving commas for multiple answers.
  - Collapse multiple whitespace characters.
- **Duplicate answer removal**:
  - Split multi-answer entries by commas and clean each entry.
  - Normalize text (lowercase, remove spaces) and reduce plural forms to singular where appropriate.
  - Remove duplicates while preserving the first occurrence.
  - Log all removed duplicates, including document identifiers, for traceability.

C All templates

This appendix contains the defined question templates.

Table 7: Level 1 Question Templates with Counts.

Template	Based On	P	H	#SO	Count
What is the position of [Person Name]? [if one]		0	1	0	1252
In what organization does [Person Name] work?		0	1	0	1288
Who is the representative of [Org Name]? [if one]		0	1	0	1358
What is the role of [Org Name] in the agreement? [if one]		0	1	0	845
What company is the [Org Role (+ Sub-Role)] in the agreement? [if one]		0	1	0	444
What is the location of [Org Name]? [if one]		0	1	0	790
Which company is associated with [Location]?		0	1	0	530
What type of location is [Location] (e.g., Headquarters, Trade Operations, etc.)? [if one]		0	1	0	43

Table 8: Level 2 Question Templates with Counts.

Template	Based On	P	H	#SO	Count
Who is the [Person Position] of [Org Name]? [if one]		0	2	0	2274
What is the position held by both [Person Name 1] and [Person Name 2]? [if one]	What is the position of [Person Name]? [if one]	0	1	1	1612
What is the role in the agreement of the company where [Person Name] is employed? [if one]	What is the role of [Org Name] in the agreement? [if one]	0	2	0	613
What are the roles of [Org Name] in the agreement?	What is the role of [Org Name] in the agreement? [if one]	1	1	0	409
What companies are the [Org Role (+ Sub-Role)] in the agreement?	What company is the [Org Role (+ Sub-Role)] in the agreement? [if one]	1	1	0	304
What role do both [Org Name 1] and [Org Name 2] have in the agreement? [if one]	What is the role of [Org Name] in the agreement? [if one]	0	1	1	257
What is the role in the agreement of the company associated with [Location]? [if one]	What is the role of [Org Name] in the agreement? [if one]	0	2	0	180
Who are the representatives of [Org Name]?	Who is the representative of [Org Name]? [if one]	1	1	0	160
What are the positions of [Person Name]?	What is the position of [Person Name]? [if one]	1	1	0	136
What are the locations of [Org Name]?	What is the location of [Org Name]? [if one]	1	1	0	91
In what organizations does [Person Name] work?	In what organization does [Person Name] work?	1	1	0	90
What is the [Location Type] office of [Org Name]?		0	2	0	70
What types of location is [Location] (e.g., Headquarters, Trade Operations, etc.)?	What type of location is [Location] (e.g., Headquarters, Trade Operations, etc.)? [if one]	1	1	0	11

Table 9: Level 3 Question Templates with Counts.

Template	Based On	P	H	#SO	Count
What are the positions held by both [Person Name 1] and [Person Name 2]?	What are the positions of [Person Name]?	1	1	1	3
What is the position held by [Person Name 1] but not by [Person Name 2]? [if one]	What are the positions held by both [Person Name 1] and [Person Name 2]?	0	1	2	8
What roles do both [Org Name 1] and [Org Name 2] have in the agreement?	What are the roles of [Org Name] in the agreement?	1	1	1	92
What role does [Org Name 1] have in the agreement which is not the role of [Org Name 2]? [if one]	What roles do both [Org Name 1] and [Org Name 2] have in the agreement?	0	1	2	278
What company is the [Org Role (+ Sub-Role) 1] but not the [Org Role (+ Sub-Role) 2] in the agreement? [if one]	What companies are both the [Org Role (+ Sub-Role) 1] and [Org Role (+ Sub-Role) 2] in the agreement?	0	1	2	56
Who are the [Person Position]s of [Org Name]?	Who is the [Person Position] of [Org Name]? [if one]	1	2	0	57
Who is the [Person Position 1] and [Person Position 2] of [Org Name]? [if one]	Who is the [Person Position] of [Org Name]? [if one]	0	2	1	1346
What are the roles in the agreement of the company where [Person Name] is employed?	What is the role in the agreement of the company where [Person Name] is employed? [if one]	1	2	0	318
What are the roles in the agreement of the company associated with [Location]?	What is the role in the agreement of the company associated with [Location]? [if one]	1	2	0	169
Who is the [Person Position] of the company which is the [Org Role(-s) (+ Sub-Role(-s))] in the agreement? [if one, and the company should be uniquely identifiable]	Who is the [Person Position] of [Org Name]? [if one]	0	3	0	262
Who is the [Person Position] of the company associated with [Location]? [if one]	Who is the [Person Position] of [Org Name]? [if one]	0	3	0	616
Who is the [Person Position] of the company where [Person Name] is employed? [if one]	Who is the [Person Position] of [Org Name]? [if one]	0	3	0	289
What is the address of [Location Type] of the company which is the [Org Role(-s) (+ Sub-Role(-s))] in the agreement? [if one, and the company should be uniquely identifiable]	What is the [Location Type] office of [Org Name]?	0	3	0	13
What is the address of [Location Type] of the company where [Person Name] is employed?	What is the [Location Type] office of [Org Name]?	0	3	0	75
Who is the [Person Position] of the company where [Person Name] is employed? [if one]	Who is the [Person Position] of [Org Name]? [if one]	0	3	0	289

Table 10: Level 4 Question Templates, part 1.

Template	Based On	P	H	#SO	Count
What are the positions held by [Person Name 1] but not by [Person Name 2]?	What are the positions held by both [Person Name 1] and [Person Name 2]? What is the position held by [Person Name 1] but not by [Person Name 2]? [if one]	1	1	2	2
What roles does [Org Name 1] have in the agreement which are not the roles of [Org Name 2]?	What roles do both [Org Name 1] and [Org Name 2] have in the agreement? What role does [Org Name 1] have in the agreement which is not the role of [Org Name 2]? [if one]	1	1	2	134
What companies are the [Org Role (+ Sub-Role) 1] but not the [Org Role (+ Sub-Role) 2] in the agreement?	What companies are both the [Org Role (+ Sub-Role) 1] and [Org Role (+ Sub-Role) 2] in the agreement? What company is the [Org Role (+ Sub-Role) 1] but not the [Org Role (+ Sub-Role) 2] in the agreement? [if one]	1	1	2	124
What is the position held by [Person Name 1] but not by [Person Name 2] or [Person Name 3]? [if one]	What is the position held by [Person Name 1] but not by [Person Name 2]? [if one]	0	1	3	50
What is the position held by [Person Name 1] and [Person Name 2] but not by [Person Name 3]? [if one]	What is the position held by [Person Name 1] but not by [Person Name 2]? [if one]	0	1	3	4
What role do [Org Name 1] and [Org Name 2] have in the agreement which is not the role of [Org Name 3]? [if one]	What role does [Org Name 1] have in the agreement which is not the role of [Org Name 2]? [if one]	0	1	3	0
What role does [Org Name 1] have in the agreement which is not the role of [Org Name 2] or [Org Name 3]? [if one]	What role does [Org Name 1] have in the agreement which is not the role of [Org Name 2]? [if one]	0	1	3	0
What company is the [Org Role (+ Sub-Role) 1] and [Org Role (+ Sub-Role) 2] but not the [Org Role (+ Sub-Role) 3] in the agreement? [if one]	What company is the [Org Role (+ Sub-Role) 1] but not the [Org Role (+ Sub-Role) 2] in the agreement? [if one]	0	1	3	1553
What company is the [Org Role (+ Sub-Role) 1] but not the [Org Role (+ Sub-Role) 2] [Org Role (+ Sub-Role) 3] in the agreement? [if one]	What company is the [Org Role (+ Sub-Role) 1] but not the [Org Role (+ Sub-Role) 2] in the agreement? [if one]	0	1	3	816

Table 11: Level 4 Question Templates, part 2.

Template	Based On	P	H	#SO	Count
Who are the both [Person Position 1]s and [Person Position 2]s of [Org Name]?	Who are the [Person Position]s of [Org Name]? Who is the [Person Position 1] and [Person Position 2] of [Org Name]? [if one]	1	2	1	0
Who are the [Person Position]s of the company which is the [Org Role(-s) (+ Sub-Role(-s))] in the agreement? [the company should be uniquely identifiable]	Who is the [Person Position] of the company which is the [Org Role(-s) (+ Sub-Role(-s))] in the agreement? [if one, and the company should be uniquely identifiable]	1	3	0	40
Who are the [Person Position]s of the company associated with [Location]?	Who is the [Person Position] of the company associated with [Location]? [if one]	1	3	0	22
Who are the [Person Position]s of the company associated where [Person Name] is employed?	Who is the [Person Position] of the company associated where [Person Name] is employed? [if one]	1	3	0	5
Who is the [Person Position] of the company which is both the [Org Role(-s) (+ Sub-Role(-s)) 1] and the [Org Role(-s) (+ Sub-Role(-s)) 2] in the agreement? [if one, and the company should be uniquely identifiable]	Who is the [Person Position] of the company which is the [Org Role(-s) (+ Sub-Role(-s))] in the agreement? [if one, and the company should be uniquely identifiable]	0	3	1	187
Who is both the [Person Position 1] and [Person Position 2] of the company which is the [Org Role(-s) (+ Sub-Role(-s))] in the agreement? [if one, and the company should be uniquely identifiable]	Who is the [Person Position] of the company which is the [Org Role(-s) (+ Sub-Role(-s))] in the agreement? [if one, and the company should be uniquely identifiable]	0	3	1	61
Who is both the [Person Position 1] and [Person Position 2] of the company associated with [Location]? [if one]	Who is the [Person Position] of the company associated with [Location]? [if one]	0	3	1	59
Who is both the [Person Position 1] and [Person Position 2] of the company associated where [Person Name] is employed? [if one]	Who is the [Person Position] of the company associated where [Person Name] is employed? [if one]	0	3	1	23
What is the address of the [Location Type] office of the company which is both the [Org Role(-s) (+ Sub-Role(-s)) 1] and the [Org Role(-s) (+ Sub-Role(-s)) 2] in the agreement? [if one, and the company should be uniquely identifiable]	What is the [Location Type] office of the company which is the [Org Role(-s) (+ Sub-Role(-s))] in the agreement? [if one, and the company should be uniquely identifiable]	0	3	1	14

Table 12: Level 5 Question Templates, part 1.

Template	Based On	P	H	#SO	Count
What are the positions held by [Person Name 1] but not by [Person Name 2] or [Person Name 3]?	What are the positions held by [Person Name 1] but not by [Person Name 2]? What is the position held by [Person Name 1] but not by [Person Name 2] or [Person Name 3]? [if one]	1	1	3	8
What are the positions held by [Person Name 1] and [Person Name 2] but not by [Person Name 3]?	What are the positions held by [Person Name 1] but not by [Person Name 2]? What is the position held by [Person Name 1] and [Person Name 2] but not by [Person Name 3]? [if one]	1	1	3	49
What roles do [Org Name 1] and [Org Name 2] have in the agreement which are not the roles of [Org Name 3]?	What roles does [Org Name 1] have in the agreement which are not the roles of [Org Name 2]? What role do [Org Name 1] and [Org Name 2] have in the agreement which is not the role of [Org Name 3]? [if one]	1	1	3	0
What roles does [Org Name 1] have in the agreement which are not the roles of [Org Name 2] or [Org Name 3]?	What roles does [Org Name 1] have in the agreement which are not the roles of [Org Name 2]? What role does [Org Name 1] have in the agreement which is not the role of [Org Name 2] or [Org Name 3]? [if one]	1	1	3	0
What companies are the [Org Role (+ Sub-Role) 1] and [Org Role (+ Sub-Role) 2] but not the [Org Role (+ Sub-Role) 3] in the agreement?	What companies are the [Org Role (+ Sub-Role) 1] but not the [Org Role (+ Sub-Role) 2] in the agreement? What company is the [Org Role (+ Sub-Role) 1] and [Org Role (+ Sub-Role) 2] but not the [Org Role (+ Sub-Role) 3] in the agreement? [if one]	1	1	3	122
What companies are the [Org Role (+ Sub-Role) 1] but not the [Org Role (+ Sub-Role) 2] [Org Role (+ Sub-Role) 3] in the agreement? [if one]	What companies are the [Org Role (+ Sub-Role) 1] but not the [Org Role (+ Sub-Role) 2] in the agreement? What company is the [Org Role (+ Sub-Role) 1] but not the [Org Role (+ Sub-Role) 2] [Org Role (+ Sub-Role) 3] in the agreement? [if one]	1	1	3	0

Table 13: Level 5 Question Templates, part 2.

Template	Based On	P	H	#SO	Count
Who are the [Person Position]s of the company which is both the [Org Role(-s) (+ Sub-Role(-s)) 1] and the [Org Role(-s) (+ Sub-Role(-s)) 2] in the agreement? [the company should be uniquely identifiable]	Who are the [Person Position]s of the company which is the [Org Role(-s) (+ Sub-Role(-s))] in the agreement? [the company should be uniquely identifiable] Who is the [Person Position] of the company which is both the [Org Role(-s) (+ Sub-Role(-s)) 1] and the [Org Role(-s) (+ Sub-Role(-s)) 2] in the agreement? [if one, and the company should be uniquely identifiable]	1	3	1	0
Who are both the [Person Position 1] and [Person Position 2] of the company which is the [Org Role(-s) (+ Sub-Role(-s))] in the agreement? [the company should be uniquely identifiable]	Who are the [Person Position]s of the company which is the [Org Role(-s) (+ Sub-Role(-s))] in the agreement? [the company should be uniquely identifiable] Who is both the [Person Position 1] and [Person Position 2] of the company which is the [Org Role(-s) (+ Sub-Role(-s))] in the agreement? [if one, and the company should be uniquely identifiable]	1	3	1	0
Who are both the [Person Position 1] and [Person Position 2] of the company associated with [Location]?	Who are the [Person Position]s of the company associated with [Location]? Who is both the [Person Position 1] and [Person Position 2] of the company associated with [Location]? [if one]	1	3	1	0
Who are both the [Person Position 1] and [Person Position 2] of the company associated where [Person Name] is employed?	Who are the [Person Position]s of the company associated where [Person Name] is employed? Who is both the [Person Position 1] and [Person Position 2] of the company associated where [Person Name] is employed? [if one]	1	3	1	0
Who is the [Person Position 1] but not [Person Position 2] of the company which is the [Org Role(-s) (+ Sub-Role(-s))] in the agreement? [if one, and the company should be uniquely identifiable]	Who is both the [Person Position 1] and [Person Position 2] of the company which is the [Org Role(-s) (+ Sub-Role(-s))] in the agreement? [if one, and the company should be uniquely identifiable]	0	3	2	2
Who is the [Person Position 1] but not [Person Position 2] of the company associated with [Location]? [if one]	Who is both the [Person Position 1] and [Person Position 2] of the company associated with [Location]? [if one]	0	3	2	1
Who is the [Person Position 1] but not [Person Position 2] of the company associated where [Person Name] is employed? [if one]	Who is both the [Person Position 1] and [Person Position 2] of the company associated where [Person Name] is employed? [if one]	0	3	2	3
Who is the [Person Position] of the company which is the [Org Role(-s) (+ Sub-Role(-s)) 1] but not the [Org Role(-s) (+ Sub-Role(-s)) 2] in the agreement? [if one, and the company should be uniquely identifiable]	Who is the [Person Position] of the company which is both the [Org Role(-s) (+ Sub-Role(-s)) 1] and the [Org Role(-s) (+ Sub-Role(-s)) 2] in the agreement? [if one, and the company should be uniquely identifiable]	0	3	2	508
What is the [Location Type] office of the company which is both the [Org Role(-s) (+ Sub-Role(-s)) 1] but not the [Org Role(-s) (+ Sub-Role(-s)) 2] in the agreement? [if one, and the company should be uniquely identifiable]	What is the [Location Type] office of the company which is both the [Org Role(-s) (+ Sub-Role(-s)) 1] and the [Org Role(-s) (+ Sub-Role(-s)) 2] in the agreement? [if one, and the company should be uniquely identifiable]	0	3	2	13

D Evaluation Details

D.1 Prompting Strategy

We implement a two-stage prompting pipeline for QA inference over long documents. In the first stage, the document is either processed **in full or split into chunks** (depending on the context limit of the LLM), and **each chunk is paired with a batch of up to 50 questions**. We then apply a one-shot, instruction-based prompt (Figure 7a) that instructs the LLM to act as a financial expert and return an array of answers – either grounded in the document or explicitly stating "Not found" – in valid JSON format. The effect of chunk-based evaluation is discussed in Section 6.

The expected format is enforced both through the prompt instructions and through structured parsing logic. The parser first attempts to **extract a JSON object from the model's response**, supporting both raw JSON and code-fenced output. If parsing fails, the system falls back to extracting answers from list-formatted outputs or reattempts the query, up to a maximum number of retries. All models are queried with temperature=0.0.

In the second stage, for questions where answers may be distributed across multiple chunks, a merging prompt is issued (Figure 7b). This prompt **consolidates partial answers from earlier responses** and asks the model to output a single final answer per question, again using the same strict JSON format. This strategy allows the model to incorporate information across multiple views of the document.

Regarding resources, as mentioned in Section 3.1, for open-weight models we are using Together.AI for some of the models. For the other open-weight models, we used 1 H200 GPU.

D.2 Evaluation Metrics

We employ four complementary metrics to evaluate model responses (Table 15).

- **Word-Level F1 Score** Token-level precision and recall against the gold answer, ensures strict factual correctness in question answering.
- **Normalized Levenshtein Distance** Character-level string distance, indicating surface-level divergence (described by Yujian and Bo (2007)).
- **Cosine Similarity** Evaluates the semantic alignment, with robustness to minor changes.
- **LLM-as-a-Judge** Gemini-2.0-Pro is prompted to score each prediction from 1 to 5 based on its semantic correctness, given the question and reference answer.

We exclude ROUGE, BLEU, and similar metrics because they are known to perform poorly on short factual answers, tend to over-penalize surface-level variations and are not well-aligned with human judgment in our setting (Blagec et al., 2022; Chaganty et al., 2018).

D.3 Human Evaluation Metrics

We conducted a human evaluation of the LLM-generated responses using three independent evaluators. Two of these evaluators were annotators familiar with credit agreements, while the third evaluator had no prior exposure to the documents. Each evaluator was requested to rate the LLM responses on a scale from 1 to 5, according to the criteria below. A total of 100 questions were sampled from multiple templates to ensure maximal diversity in the human evaluation task.

- 1 response is irrelevant or does not match the actual answer;
- 2 response is somewhat relevant but does not match the actual answer;
- 3 response shows slight similarity but lacks accuracy;
- 4 response is largely relevant and accurate;
- 5 response is a perfect match to the actual answer.

We assess the relationship between human evaluation scores and other metrics using both Pearson and Kendall's Tau correlation coefficients (see Figure 8 and Table 16). Pearson correlation captures linear relationships, while Kendall's Tau assesses rank-based similarity, making them complementary for evaluating metric alignment. Across both measures, F1 Score and LLM-as-a-Judge consistently show the highest positive correlations with average human ratings—indicating not only strong agreement in scale but also in ranking. This indicates that both F1 Score and LLM-as-a-Judge closely align with human evaluations, supporting their use as strong performance metrics.

Metric prioritization and interpretation Based on the observed correlations with human judgments, we recommend interpreting LLM-as-a-Judge as the primary indicator of semantic correctness, particularly for higher-complexity questions where answers may be paraphrased, partially implicit, or expressed with variable surface forms. F1 score should be viewed as a complementary metric that captures extractive fidelity and coverage, and is most informative when exact grounding in the doc-

Table 14: Mean human performance across question levels.

Level	F1	Edit Dist.	Cosine Sim.
1	0.846	0.142	0.834
2	0.778	0.264	0.750
3	0.780	0.235	0.756
4	0.803	0.244	0.795
5	0.935	0.101	0.928

ument is required. Discrepancies between the two metrics are therefore not treated as inconsistencies but as signals of different model behaviors (e.g., cautious but semantically accurate answers versus broader but partially grounded ones), as discussed in Section 3.3. For practical use, we recommend reporting both metrics jointly: LLM-as-a-Judge to assess answer correctness from a human-aligned perspective, and F1 to diagnose retrieval completeness and over-/under-generation. The remaining metrics (Edit Distance and Cosine Similarity) provide additional diagnostic insight into surface-level mismatch and semantic drift but are not intended to be used in isolation for model ranking.

D.4 Cost-Constrained Models

For the cost-constrained models, we used stratified sampling over our question templates. Because each template corresponds to a unique point in the (H, P, #SO) space, sampling proportionally over templates preserves the difficulty distribution, the multi-hop vs single-hop ratio, and the relative frequency of set-operation queries. Using DeepSeek-V3, the strongest open-weight model evaluated on the full dataset, we have verified that performance on the sampled subset correlates strongly with full-set performance: 0.87 for F-score, 0.84 for edit distance, and 0.88 for cosine similarity.

D.5 Human Performance Evaluation

To establish a reference point for model comparison, we conducted a human evaluation on a balanced subset of 100 questions (20 per difficulty level). Three independent human evaluators participated in this study: one annotator who was fully familiar with the documents, and two participants with general financial knowledge but no prior exposure to the documents. Their responses were normalized and compared against ground truth answers using three similarity metrics: word level F1, edit distance and cosine similarity. Unlike model performance, which typically declines as task complexity increases, human performance remained relatively stable across levels, as seen in Table 14. This suggests that humans effectively leverage contextual reasoning and background knowledge to consistently maintain accuracy even in multi-hop or compositional settings, capabilities that current models still struggle with.

Batch QA Prompt

[SYSTEM INPUT]
You are a financial expert, and your task is to answer the questions given to you in batches about the provided credit agreement. If you believe the answer is not present in the agreement, say 'Not found'.

[EXPECTED OUTPUT]
Respond ONLY with valid JSON, nothing else. See the example below.

The given document:
Tesla, Inc. is an American electric vehicle and clean energy company founded in 2003 by Martin Eberhard and Marc Tarpenning. Elon Musk became the largest investor and later CEO.

The given questions:
Q1: Who founded Tesla?
Q2: What year was Tesla founded?

The expected output:

```
{
   "answers": [
         {"question_index": 1,
         "answer": "Martin Eberhard,
                    Marc Tarpenning"},
         {"question_index": 2,
         "answer": "2003"}
     ]
}
```

[USER INPUT]
{Document Context}

[QUESTIONS]
Q1: {question}
Q2: {question}
...

(a) Example of a batch prompt combining up to 50 questions with a single document chunk. The LLM is instructed to return only JSON-formatted answers, each linked to a question index.

Merging Prompt

[SYSTEM INPUT]
You are a financial expert, and your task is to combine or merge the provided partial answers, coming from different chunks of a credit agreement, into a single final answer for each of the questions given to you. If you believe the answer is not present in the agreement, say 'Not found'.

[EXPECTED OUTPUT]
The given document:
Tesla, Inc. is an American electric vehicle and clean energy company founded in 2003 by Martin Eberhard and Marc Tarpenning. Elon Musk became the largest investor and later CEO.

The given questions:
Q1: Who founded Tesla?
Q2: What year was Tesla founded?

The expected output:

```
{
   "answers": [
         {"question_index": 1,
         "answer": "Martin Eberhard,
                    Marc Tarpenning"},
         {"question_index": 2,
         "answer": "2003"}
     ]
}
```

[USER INPUT]
Chunk 1 partial answer JSON: {ans 1}
Chunk 2 partial answer JSON: {ans 2}
...
[QUESTIONS]
Q1: {question}
Q2: {question}
...

(b) Example of a merging prompt that consolidates partial answers from multiple document chunks into one final JSON-formatted answer per question.

Figure 7: Two-stage prompting pipeline. Long documents are chunked if needed and paired with question batches ($\leqslant 50$). A structured one-shot prompt is used per chunk. For multi-chunk documents, a merging prompt consolidates partial answers.

Table 15: Comparison of model performance across difficulty levels, including F1 Score, Normalized Edit Distance, Cosine Similarity, and LLM-as-a-Judge. * denotes the models evaluated on a smaller subset due to resources constraints.

Model	Easy				Medium				Hard			
	F1 Score ↑	Edit Dist. ↓	Cosine Sim. ↑	LLM-as-a-Judge ↑	F1 Score ↑	Edit Dist. ↓	Cosine Sim. ↑	LLM-as-a-Judge ↑	F1 Score ↑	Edit Dist. ↓	Cosine Sim. ↑	LLM-as-a-Judge ↑
Proprietary LLMs												
Gemini-2.0-Flash*	**0.6748**	**0.2679**	**0.6650**	**4.0000**	0.3294	0.6280	0.3121	2.4820	0.1172	**0.7238**	0.0608	**2.2838**
Gemini-2.0-Flash-Lite*	0.5908	0.3763	0.5829	3.4172	0.3883	0.5985	0.3715	2.6497	**0.1943**	0.7956	0.1255	1.5135
Gemini-1.5-Pro*	0.5410	0.3903	0.5331	3.2448	**0.4577**	**0.5340**	**0.4415**	**2.8436**	0.1791	0.7786	**0.1780**	1.6986
Gemini-1.5-Flash*	0.5103	0.4285	0.5032	3.1207	0.3892	0.6101	0.3657	2.5130	0.1400	0.8389	0.1255	1.5411
GPT-4o*	0.5272	0.4177	0.5213	3.2034	0.3783	0.6098	0.3690	2.5451	0.1178	0.7738	0.1144	1.4795
GPT-4o-Mini*	0.4723	0.4614	0.4654	2.9690	0.2732	0.6756	0.2740	2.0681	0.0274	0.8294	0.0336	1.1370
Open-source LLMs												
DeepSeek-R1*	**0.6269**	**0.3510**	**0.6208**	**3.6965**	0.3293	0.5191	0.3179	2.6354	0.2609	0.7733	0.1451	2.0736
DeepSeek-V3	0.6218	0.3722	0.6077	3.5668	**0.4796**	0.4912	0.4678	3.0210	0.0856	0.7975	0.0737	1.4795
GPT OSS 120B	0.5885	0.4187	0.5776	3.5839	0.4684	**0.4765**	**0.4740**	**3.1661**	0.1136	**0.7067**	0.1202	1.4888
Llama-4-Maverick-Instruct*	0.5552	0.3585	0.5494	3.4482	0.3848	0.5606	0.3669	2.6976	0.2031	0.7231	0.1172	**3.1351**
Llama-4-Scout-Instruct*	0.4993	0.4114	0.4920	3.2138	0.2215	0.6765	0.2099	1.8772	0.0472	0.8225	0.0303	1.5270
Llama-3.1-Instruct-Turbo*	0.5472	0.3901	0.5389	3.4793	0.4204	0.5489	0.4018	2.6377	**0.3651**	0.7197	**0.2849**	2.7297
Llama-3.1-Instruct	0.4955	0.4707	0.4847	3.1373	0.4017	0.5481	0.3917	2.6722	0.1639	0.7667	0.1106	2.3425
Llama-3.1-Instruct	0.3891	0.5617	0.3793	2.6996	0.2962	0.6255	0.2856	2.2332	0.1169	0.7980	0.0915	1.4315
Qwen3-Instruct	0.5243	0.4243	0.5133	3.2693	0.2760	0.5795	0.2670	2.4460	0.1240	0.7473	0.1129	1.9578
Qwen2-Instruct	0.3886	0.6369	0.3710	3.3884	0.3100	0.7031	0.2986	2.9078	0.0762	0.8200	0.0690	1.3288

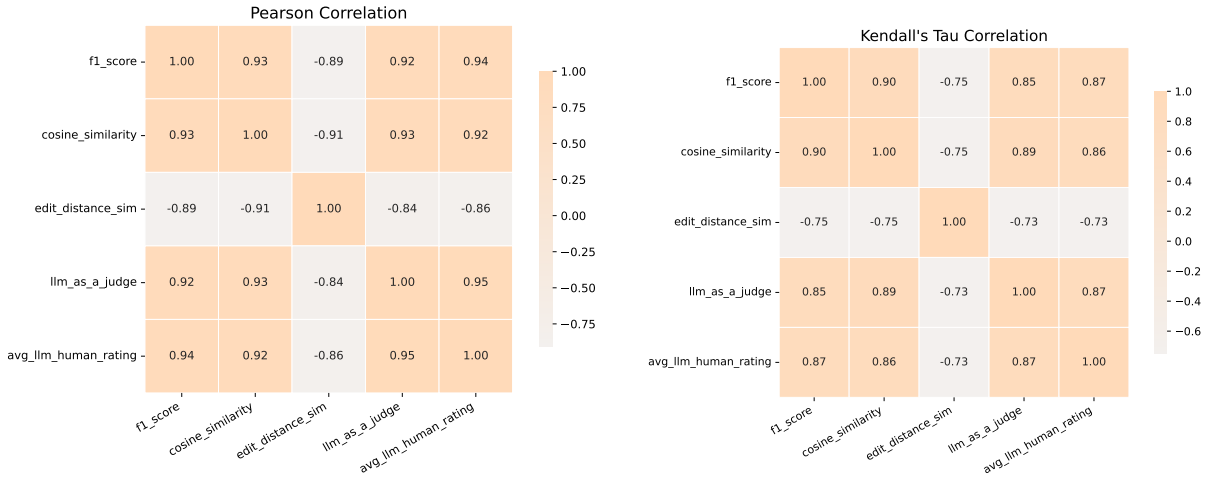
Table 16: Pearson and Kendall's Tau Correlation Tables.

Pearson Correlation Matrix

	F1 Score	Cosine Similarity	Edit Distance Sim	LLM-as-a-Judge	Avg LLM-Human Rating
F1 Score	1.000	0.933	-0.894	0.922	0.941
Cosine Similarity	0.933	1.000	-0.910	0.935	0.922
Edit Distance Sim	-0.894	-0.910	1.000	-0.844	-0.862
LLM-as-a-Judge	0.922	0.935	-0.844	1.000	0.947
Avg LLM-Human Rating	0.941	0.922	-0.862	0.947	1.000

Kendall's Tau Correlation Matrix

	F1 Score	Cosine Similarity	Edit Distance Sim	LLM-as-a-Judge	Avg LLM-Human Rating
F1 Score	1.000	0.902	-0.745	0.854	0.870
Cosine Similarity	0.902	1.000	-0.746	0.894	0.858
Edit Distance Sim	-0.745	-0.746	1.000	-0.734	-0.733
LLM-as-a-Judge	0.854	0.894	-0.734	1.000	0.867
Avg LLM-Human Rating	0.870	0.858	-0.733	0.867	1.000



(a) **Pearson correlation between various metrics.** F1 Score and Avg LLM-Human Rating (r = 0.941), and LLM-as-a-Judge and Avg LLM-Human Rating (r = 0.947) show the highest correlations in the matrix.

(b) **Kendall-Tau correlation between various metrics.** The highest correlations are observed between F1 Score and Avg LLM-Human Rating (τ = 0.870), and between LLM-as-a-Judge and Avg LLM-Human Rating (τ = 0.867).

Figure 8: Pearson and Kendall's Tau Correlation Matrices.

D.6 Comparison with Oracle and RAG

We compare four retrieval settings for answering questions from credit agreements: (1) *Full*, where the entire agreement is provided to the model, (2) *Oracle*, which uses the pieces of text containing the answer; (3) *RAG* (Retrieval-Augmented Generation), which selects top passages using a retriever; and (4) *Dynamic RAG*, which iteratively decomposes the question into sub-queries, retrieves relevant passages across multiple steps, and aggregates the retrieved evidence before generating the final answer. In Table 17, we observe that RAG, both simple and dynamic, performs drastically worse than the other two settings across all difficulty levels, which highlights the incompatibility of standard retrieval techniques with dense, cross-referenced financial documents, often requiring the aggregation of logically related but spatially distant elements.

Interestingly, the comparison between Oracle and Full Document settings reveals a non-trivial pattern. For easy look-up and for medium questions, the Oracle setting performs better, as expected, since such questions benefit from isolating the minimal, most relevant snippet. However, for hard questions, the Full Document consistently outperforms the Oracle setting. This suggests that isolating the "correct" snippet removes necessary contextual cues that models need for reasoning, ***which is detailed in the Implicit Information Gaps paragraph of our Error Analysis*** (Section 4.2). These findings reinforce the value of credit agreements as a testbed for long-context LLMs, and demonstrate how our framework's question complexity help uncover the precise scenarios where models struggle, even when given the exact ground-truth passage (see Section 4 and E for the detailed analysis).

E Error Analysis

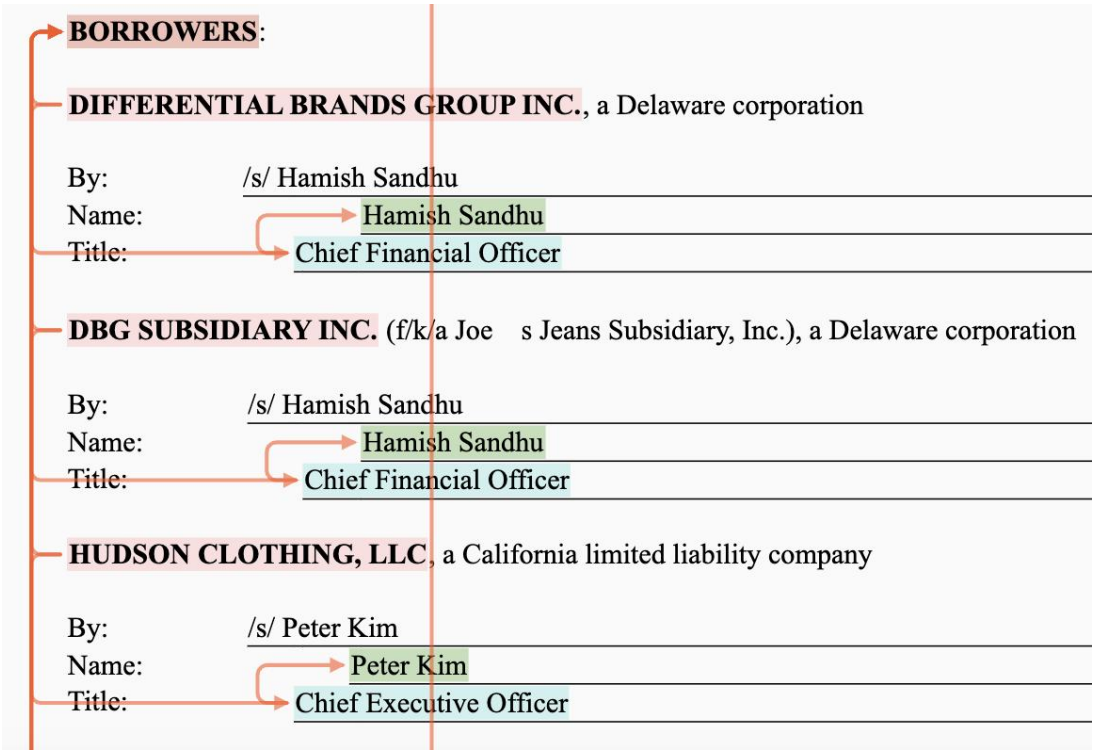
Detailed examples of errors from each of the aforementioned categories are shown in Table 18 and Figure 9. Aggregate trends across error types and complexity are shown in Figures 10 and 11.

Table 17: Performance comparison across four retrieval settings – Full (entire credit agreement), Oracle (gold snippet only), RAG and dynamic RAG (retriever-selected passages) – for each question complexity group.

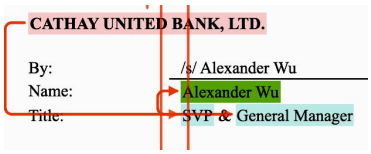
Setting	Easy				Medium				Hard			
	F1 Score ↑	Edit Dist. ↓	Cosine Sim. ↑	LLM-as-a-Judge ↑	F1 Score ↑	Edit Dist. ↓	Cosine Sim. ↑	LLM-as-a-Judge ↑	F1 Score ↑	Edit Dist. ↓	Cosine Sim. ↑	LLM-as-a-Judge ↑
Full	0.6747	0.2678	0.6650	**4.000**	**0.3294**	0.6279	0.3121	2.4820	**0.1172**	**0.7237**	0.0607	**2.2837**
Oracle	**0.7134**	**0.2588**	**0.7021**	3.9862	**0.3294**	**0.5986**	**0.3201**	**2.6497**	0.1127	0.8066	**0.0735**	1.9189
RAG	0.1579	0.7364	0.1546	1.6931	0.0695	0.8070	0.0651	1.4131	0.0190	0.8449	0.0107	1.2027
D-RAG	0.1671	0.7304	0.1631	1.7185	0.1118	0.7693	0.1087	1.5566	0.0107	0.8431	0.0069	1.1088

Table 18: Error analysis with examples.

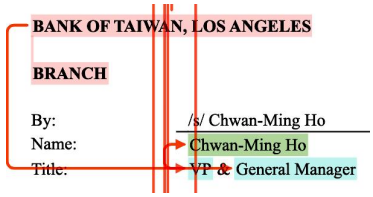
Question	*What type of location is Dubai, UAE?*
Expected Answer	*Zonal Office*
LLM Answer	*City*
Error Explanation	**Misinterpretation of Semantics.** The LLM misinterprets the semantic intent of the question. While *City* is factually accurate, the question requires identifying the Location Type of "Dubai, UAE" in the context of the credit agreement. The expected answer, *zonal office,* reflects a more precise understanding.
Question	*In what organizations does Hamish Sandhu work?*
Expected Answer	*DBG Holdings Subsidiary Inc, Differential Brands Group Inc*
LLM Answer	*RG Parent LLC*
Error Explanation	**Implicit Information Gaps.** The LLM fails to infer implicit information due to structural complexity. Human readers can associate individuals with their organizations based on layout and context. The LLM incorrectly attributes Hamish Sandhu to the wrong organization.
Question	*What is the position held by Alexander Wu but not by Chwan Ming Ho or Edward Chen?*
Expected Answer	*SVP*
LLM Answer	*SVP and General Manager*
Error Explanation	**Set Operation Failures.** The LLM struggles with multiple set operations. Alexander Wu is SVP and General Manager, while Chwan-Ming Ho and Edward Chen are both VPs and General Managers. The extra "General Manager" indicates a failure in subtracting shared titles.
Question	*What is the role of Celanese Chemicals, Inc. in the agreement?*
Expected Answer	*Subsidiary*
LLM Answer	*Not found*
Error Explanation	**Long-Context Retrieval Limitations.** The model fails because "subsidiary" is not explicitly stated in a natural sentence. Instead, Celanese Chemicals, Inc. appears under headings or clauses indicating subsidiaries, requiring structural and hierarchical reasoning.



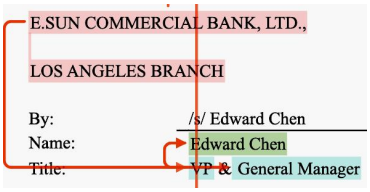
(a) Multiple entities and roles are listed in close proximity, making it difficult for the model to associate roles accurately.



(b) Alexander Wu



(c) Chwan Ming Ho



(d) Edward Chen

Figure 9: (a) In cases involving multiple signatories, the LLM cannot infer implicit information like Person Positions. (b), (c) and (d) illustrate the LLM's inability to deal with multiple set operations.

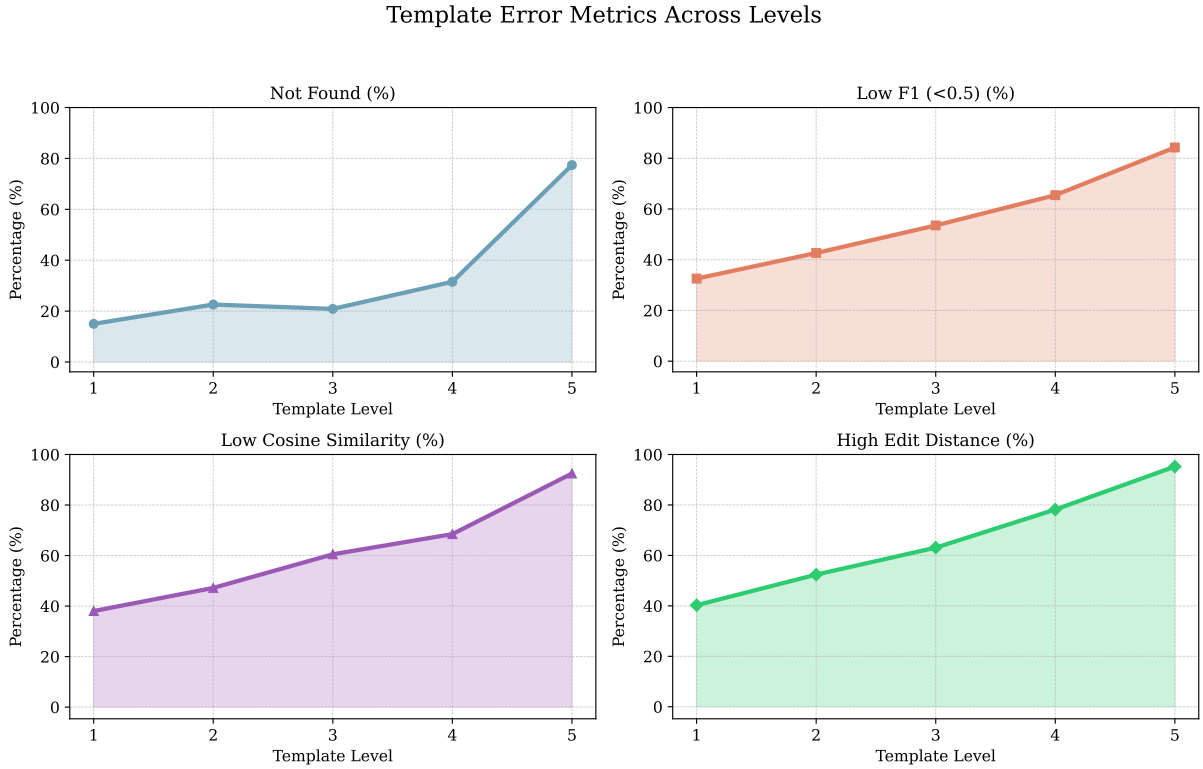
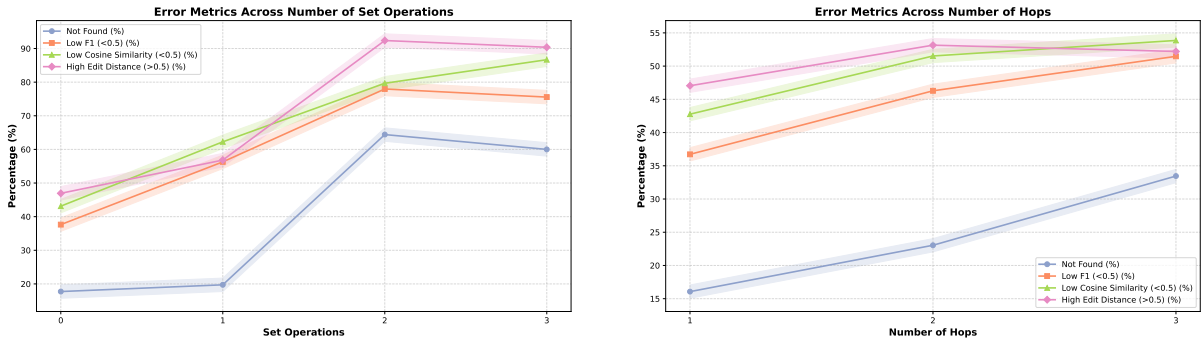


Figure 10: **Trends in Error Metrics Across Template Levels.** The plots illustrate a clear upward trend across all error metrics as template complexity increases. This includes the percentage of LLM responses marked as "Not Found", as well as those exhibiting low F1 scores, low cosine similarity, and high edit distance – indicating a decline in accuracy with higher-level templates.



(a) Error metrics across set operations.

(b) Error metrics across number of hops.

Figure 11: **Comparison of error metrics for Deepseek-V3.** (a) Trends by Set Operations. (b) Trends by Number of Hops. With increase in number of hops or set operations, the percentage distribution of each error type increases.

F Case Study on Medical Documents

To illustrate the generalizability and applicability of our framework to other domain, we conducted a Case study on the publicly available[8] "Guidelines for the Prevention and Treatment of Opportunistic Infections in Adults and Adolescents With HIV". We extracted the introductory part, the first 10 descriptions of diseases, and the conclusive part, which results in a 225-page document. We annotated the document according to the schema in Figure 12, which allowed us to extract questions of various complexity levels over the same 3 dimensions as in the main study (see Tables 20-22 for the full list of question templates).

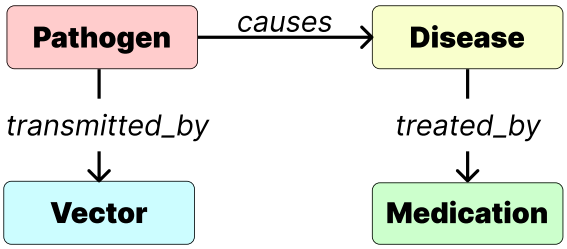


Figure 12: Annotation schema of our case study.

We ran our experiments with the best-performing open-weight model, DeepSeek-V3. In Table 19, we observe a clear performance decrease as the question complexity increases, which illustrates that our framework is generalizable and applicable to other domains.

Table 19: Evaluation of DeepSeek-V3 on the case study dataset.

Difficulty	F1 ↑	Edit ↓	Cos ↑	LLM-J ↑
Easy	0.2585	0.6169	0.2522	2.2249
Medium	0.0753	0.7512	0.0662	1.5587
Hard	0.0	0.8417	0.0	1.0681

[8] https://clinicalinfo.hiv.gov/en/guidelines/hiv-clinical-guidelines-adult-and-adolescent-opportunistic-infections/whats-new

Table 20: Case study 1-hop Question Templates

Template	P	H	#SO
What pathogen is transmitted by [Vector]? [if one]	0	1	0
What vector transmits [Pathogen]? [if one]	0	1	0
What pathogen causes [Disease]? [if one]	0	1	0
What disease is caused by [Pathogen]? [if one]	0	1	0
What disease is treated by [Medication]? [if one]	0	1	0
What medication treats [Disease]? [if one]	0	1	0
What pathogens are transmitted by [Vector]?	1	1	0
What vectors transmit [Pathogen]?	1	1	0
What pathogens cause [Disease]?	1	1	0
What diseases are caused by [Pathogen]?	1	1	0
What diseases are treated by [Medication]?	1	1	0
What medications treat [Disease]?	1	1	0
What pathogen is transmitted by [Vector 1] and [Vector 2]? [if one]	0	1	1
What vector transmits [Pathogen 1] and [Pathogen 2]? [if one]	0	1	1
What pathogen causes [Disease 1] and [Disease 2]? [if one]	0	1	1
What disease is caused by [Pathogen 1] and [Pathogen 2]? [if one]	0	1	1
What disease is treated by [Medication 1] and [Medication 2]? [if one]	0	1	1
What medication treats [Disease 1] and [Disease 2]? [if one]	0	1	1
What pathogens are transmitted by [Vector 1] and [Vector 2]?	1	1	1
What vectors transmit [Pathogen 1] and [Pathogen 2]?	1	1	1
What pathogens cause [Disease 1] and [Disease 2]?	1	1	1
What diseases are caused by [Pathogen 1] and [Pathogen 2]?	1	1	1
What diseases are treated by [Medication 1] and [Medication 2]?	1	1	1
What medications treat [Disease 1] and [Disease 2]?	1	1	1
What pathogen is transmitted by [Vector 1] but not [Vector 2]? [if one]	0	1	2
What vector transmits [Pathogen 1] but not [Pathogen 2]? [if one]	0	1	2
What pathogen causes [Disease 1] but not [Disease 2]? [if one]	0	1	2
What disease is caused by [Pathogen 1] but not [Pathogen 2]? [if one]	0	1	2
What disease is treated by [Medication 1] but not [Medication 2]? [if one]	0	1	2
What medication treats [Disease 1] but not [Disease 2]? [if one]	0	1	2
What pathogens are transmitted by [Vector 1] but not [Vector 2]?	1	1	2
What vectors transmit [Pathogen 1] but not [Pathogen 2]?	1	1	2
What pathogens cause [Disease 1] but not [Disease 2]?	1	1	2
What diseases are caused by [Pathogen 1] but not [Pathogen 2]?	1	1	2
What diseases are treated by [Medication 1] but not [Medication 2]?	1	1	2
What medications treat [Disease 1] but not [Disease 2]?	1	1	2

Table 21: Case study 2-hop Question Templates

Template	P	H	#SO
What vector transmits a pathogen which causes [Disease]? [if one]	0	2	0
What disease is caused by a pathogen which is transmitted by [Vector]? [if one]	0	2	0
What medication treats a disease which is caused by [Pathogen]? [if one]	0	2	0
What pathogen causes a disease which is treated by [Medication]? [if one]	0	2	0
What vectors transmit a pathogen which causes [Disease]?	1	2	0
What diseases are caused by a pathogen which is transmitted by [Vector]?	1	2	0
What medications treat a disease which is caused by [Pathogen]?	1	2	0
What pathogens cause a disease which is treated by [Medication]?	1	2	0
What vector transmits a pathogen which causes [Disease 1] and [Disease 2]? [if one]	0	2	1
What disease is caused by a pathogen which is transmitted by [Vector 1] and [Vector 2]? [if one]	0	2	1
What medication treats a disease which is caused by [Pathogen 1] and [Pathogen 2]? [if one]	0	2	1
What pathogen causes a disease which is treated by [Medication 1] and [Medication 2]? [if one]	0	2	1
What vectors transmit a pathogen which causes [Disease 1] and [Disease 2]?	1	2	1
What diseases are caused by a pathogen which is transmitted by [Vector 1] and [Vector 2]?	1	2	1
What medications treat a disease which is caused by [Pathogen 1] and [Pathogen 2]?	1	2	1
What pathogens cause a disease which is treated by [Medication 1] and [Medication 2]?	1	2	1
What vector transmits a pathogen which causes [Disease 1] but not [Disease 2]? [if one]	0	2	2
What disease is caused by a pathogen which is transmitted by [Vector 1] but not [Vector 2]? [if one]	0	2	2
What medication treats a disease which is caused by [Pathogen 1] but not [Pathogen 2]? [if one]	0	2	2
What pathogen causes a disease which is treated by [Medication 1] but not [Medication 2]? [if one]	0	2	2
What vectors transmit a pathogen which causes [Disease 1] but not [Disease 2]?	1	2	2
What diseases are caused by a pathogen which is transmitted by [Vector 1] but not [Vector 2]?	1	2	2
What medications treat a disease which is caused by [Pathogen 1] but not [Pathogen 2]?	1	2	2
What pathogens cause a disease which is treated by [Medication 1] but not [Medication 2]?	1	2	2

Table 22: Case study 3-hop Question Templates

Template	P	H	#SO
What vector transmits a pathogen which causes a disease which is treated by [Medication]? [if one]	0	3	0
What medication treats a disease which is caused by a pathogen which is transmitted by [Vector]? [if one]	0	3	0
What vectors transmit a pathogen which causes a disease which is treated by [Medication]?	1	3	0
What medications treat a disease which is caused by a pathogen which is transmitted by [Vector]?	1	3	0
What vector transmits a pathogen which causes a disease which is treated by [Medication 1] and [Medication 2]? [if one]	0	3	1
What medication treats a disease which is caused by a pathogen which is transmitted by [Vector 1] and [Vector 2]? [if one]	0	3	1
What vectors transmit a pathogen which causes a disease which is treated by [Medication 1] and [Medication 2]?	1	3	1
What medications treat a disease which is caused by a pathogen which is transmitted by [Vector 1] and [Vector 2]?	1	3	1
What vector transmits a pathogen which causes a disease which is treated by [Medication 1] but not [Medication 2]? [if one]	0	3	2
What medication treats a disease which is caused by a pathogen which is transmitted by [Vector 1] but not [Vector 2]? [if one]	0	3	2
What vectors transmit a pathogen which causes a disease which is treated by [Medication 1] but not [Medication 2]?	1	3	2
What medications treat a disease which is caused by a pathogen which is transmitted by [Vector 1] but not [Vector 2]?	1	3	2

G Dataset Release and Leaderboard Setup

To facilitate reproducibility and future research (Tatarinov et al., 2025), we are publicly releasing (under CC-BY NC ND 4.0 license) QA pairs for 40 documents out of the total 170 as the development set, allowing researchers to validate model performance under controlled conditions. The remaining QA pairs serve as the test set, which is not publicly available to prevent data contamination (Sainz et al., 2023). For this test set, we are making the questions public, without the ground-truth answers. The distribution of the test and dev sets is provided in Table 23. In addition, we are hosting a leaderboard[9], following standard practices in recent LLM evaluation benchmarks (Galarnyk et al., 2025; Yue et al., 2024; Zhao et al., 2024a; Lu et al., 2023).

Table 23: Dataset statistics of the constructed KG-MuLQA-D dataset.

Stats	Dev	Test	Total
# Documents	40	130	170
# Question per doc (min)	1	1	1
# Question per doc (avg)	14.75	23.49	21.44
# Question per doc (max)	83	428	428
# Easy Questions (count)	1,499	5,051	6,550
# Medium Questions (count)	2,680	10,203	12,883
# Hard Questions (count)	239	467	706
# Questions (count)	4,418	15,721	20,139

H Author Contribution

NT led the data acquisition by extracting credit agreements from SEC EDGAR, developed the annotation guidelines specifying the entity and relation schemas, conceptualized the RDF-based knowledge graph framework, devised multi-dimensional QA templates, and benchmarked long-context LLMs in both RAG and oracle settings. VK extensively annotated and cleaned documents, identified edge cases and was involved in developing the KG construction and question-answer generation modules. She performed error analysis, led human evaluation, and constructed visualizations. HS contributed to the document annotation process, developed the pipelines for KG construction and SPARQL QA generation and assisted in LLM response analysis. AR implemented the initial end-to-end benchmarking pipeline and supported the human evaluation. He experimented with various prompting techniques and evaluation metrics and created one-shot prompts used in experiments. HSA contributed to all annotation stages and SPARQL-based question generation. VS, AL, and RL contributed to annotation, data cleaning, and manuscript refinement. AS supervised all iterations of the annotation process, guided edge-case incorporation, shaped research direction for document-level question answering, advised on design of additional studies, and contributed to error analysis and writing. SC provided overall project supervision and guidance.

I Acknowledgments

We would like to thank Anvita Mahajan, Archishman VB, Arvind K R, Daksh Jain, Deng Li, Jorge Navarro Gracia, Kaustubh Gayadhankar, Krish Shah, Kritika Bansal, Meher Bhardwaj, Nilesh Gupta, Piyush Mohapatra, Pranav Krishna, Rushi Glasswala, Shashwat Bajpai, Soumyajit Basu, Tvisha Shah, Vishnu Varma, Xu Huang and Udish Jangid for contributing to the initial round of annotations and early-stage coding experiments.

[9] https://huggingface.co/spaces/gtfintechlab/KG-MuLQA-D-Leaderboard