Comparing Chain and Tree-Based Reasoning for Explainable Knowledge Discovery in Contract Analytics Using Large Language Models

Antony Seabra[a], Claudio Cavalcante[b] and Sergio Lifschitz[c]
Departamento de Informatica, PUC-Rio, Brazil

Keywords: Chain-of-Thought Prompting, Tree-of-Thought Reasoning, Contract Analytics, Knowledge Discovery, Large Language Models, Business Intelligence, Decision Support Systems.

Abstract: This paper presents a comparative analysis of two structured reasoning strategies—Chain-of-Thought (CoT) and Tree-of-Thought (ToT)—for explainable knowledge discovery with Large Language Models (LLMs). Grounded in real-world IT contract management scenarios, we apply both techniques to a diverse set of competency questions that require advanced reasoning over structured and unstructured data. CoT guides the model through sequential, linear reasoning steps, whereas ToT enables the exploration of multiple reasoning paths before selecting a final response. We evaluate the generated insights using three key criteria: clarity, usefulness, and confidence in justifications, with particular attention to their effectiveness in supporting decision-making. The results indicate that ToT produces richer and more comprehensive rationales in complex scenarios, while CoT offers faster and more direct responses in narrowly defined tasks. Our findings highlight the complementary strengths of each approach and contribute to the design of adaptive, self-rationalizing AI agents capable of delivering explainable and actionable recommendations in contract analysis contexts.

1 INTRODUCTION

The increasing complexity of enterprise contracts, particularly in sectors such as information technology, has created a pressing demand for intelligent systems capable of extracting, interpreting, and explaining strategic insights from both structured and unstructured data sources. Traditional Business Intelligence (BI) tools, while effective for analyzing structured databases, often fall short in addressing high-level analytical tasks that require synthesis, inference, and justification, especially when dealing with heterogeneous information distributed across legal clauses, performance metrics, and historical trends.

Recent advances in Large Language Models (LLMs) have enabled the development of AI-driven agents capable of answering complex queries using natural language and diverse knowledge sources. However, despite their expressive capabilities, LLMs often behave as black boxes, offering conclusions without clear or traceable reasoning, which hinders their adoption in critical decision-making scenarios such as contract negotiation, compliance auditing, and risk mitigation.

To address this challenge, structured reasoning strategies such as Chain-of-Thought (CoT) and Tree-of-Thought (ToT) prompting have emerged as promising solutions for enhancing the transparency and interpretability of LLM-generated responses. CoT enables step-by-step linear reasoning, guiding the model through a structured narrative to reach its conclusion. ToT, by contrast, simulates multi-path reasoning: it explores diverse branches of logic in parallel and selects the most compelling or justified outcome.

In this paper, we perform a comparative study of CoT and ToT techniques applied to a set of real-world contract analysis tasks in the context of a BI system. By reusing competency questions from previous work and applying each reasoning strategy to the same analytical scenarios, we examine how these approaches affect the clarity, usefulness, and confidence of knowledge discovery. We also evaluate trade-offs in terms of computational cost, response time, and user-perceived value of the generated explanations. Our goal is to provide empirical insights into the practical effectiveness of CoT and ToT in supporting explainable knowledge discovery in contractual

[a] https://orcid.org/0009-0007-9459-8216
[b] https://orcid.org/0009-0007-6327-4083
[c] https://orcid.org/0000-0003-3073-3734

Seabra, A., Cavalcante, C. and Lifschitz, S.
Comparing Chain and Tree-Based Reasoning for Explainable Knowledge Discovery in Contract Analytics Using Large Language Models.
DOI: 10.5220/0013752300004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 1: KDIR*, pages 426-435
ISBN: 978-989-758-769-6; ISSN: 2184-3228

domains. The findings inform the design of adaptive reasoning agents capable of selecting the appropriate strategy depending on question complexity, user intent, and the nature of the available data.

The remainder of the paper is organized as follows: Section 2 provides background on reasoning strategies for LLM integration. Section 3 details our methodology for comparing CoT and ToT. Section 4 introduces the system architecture and implementation. Section 5 presents the experimental evaluation. Section 6 reviews related work, and Section 7 concludes with final remarks and directions for future research.

2 BACKGROUND

2.1 Large Language Models

Large Language Models (LLMs) have revolutionized the field of Natural Language Processing (NLP) with their ability to understand and generate human-like text. At the heart of the most advanced LLMs is the Transformers architecture, a deep learning model introduced in the seminal paper *Attention Is All You Need* by (Vaswani et al., 2017). Transformers leverage a mechanism called *attention*, which allows the model to weigh the influence of different parts of the input data at different times, effectively enabling it to focus on relevant parts of the text when making predictions.

Prior to Transformers, Recurrent Neural Networks (RNNs) and their variants like Long Short-Term Memory (LSTM) networks were the standard in NLP. These architectures processed input data sequentially, which naturally aligned with the sequential nature of language. However, they had limitations, particularly in dealing with long-range dependencies within text due to issues like vanishing gradients (Pascanu et al., 2013). Transformers overcome these challenges by processing all parts of the input data in parallel, drastically improving the model's ability to handle long-distance relationships in text.

Chat models, a subset of LLMs, are specialized in generating conversational text that is coherent and contextually appropriate. This specialization is achieved through the training process, where the models are fed vast amounts of conversational data, enabling them to learn the nuances of dialogue. ChatGPT, for instance, is fine-tuned on a dataset of conversational exchanges and it was optimized for dialogue by using Reinforcement Learning with Human Feedback (RLHF) - a method that uses human demonstrations and preference comparisons to guide the model toward desired behavior (OpenAI, 2023a).

The transformative impact of LLMs, and particularly those built on the Transformers architecture, has been profound. By moving away from the constraints of sequential data processing and embracing parallelization and attention mechanisms, these models have set new standards for what is possible in the realm of NLP. With the ability to augment generation with external data or specialize through fine-tuning, LLMs have become not just tools for language generation but platforms for building highly specialized, knowledge-rich applications that can retrieve information in a dialogue-like way, find useful information and generate insights for decision making.

The ability to augment the generation capabilities of LLMs using enriched context from external data sources is a significant advancement in AI-driven systems. An LLM context refers to the surrounding information provided to a LLM to enhance its understanding and response generation capabilities. This context can include a wide array of data, such as text passages, structured data, and external data sources like Knowledge Graphs. Utilizing these external data sources allows the LLM to generate more accurate and relevant responses without the need for retraining. By providing detailed context, such as product attributes, user reviews, or categorical data, the model can produce insights that are tailored and contextually aware.

2.2 Prompt Engineering

One key aspect of providing contexts to LLMs is the ability of designing and optimizing prompts to guide LLMs in generating the answers. This is what is called Prompt Engineering. Its main goal is to maximize the potential of LLMs by providing them with instructions and context (OpenAI, 2023b).

In the realm of Prompt Engineering, instructions are the crucial first steps. Through them, engineers can detail the roadmap to an answer, outlining the desired task, style and format for the LLM's response (White et al., 2023). For instance, To define the style of a conversation, a prompt could be phrased as "Use professional language and address the client respectfully" or "Use informal language and emojis to convey a friendly tone". To specify the format of dates in answers, a prompt instruction could be "Use the American format, MM/DD/YYYY, for all dates".

On the other hand, as mentioned earlier, context refers to the information provided to LLMs alongside the core instructions. The most important aspect of a context is that it can provide information that supports the answer given by the LLM, and it is very

useful when implementing question-answering systems. This supplemental context can be presented in various formats. One particularly effective format is RDF triples, which represent information as subject-predicate-object statements. RDF triples are a standardized way of encoding structured data about entities and their relationships, making them ideal for embedding precise information into prompts. By including RDF triples in a prompt, we can clearly convey complex relationships and attributes in a format that the LLM can easily process, leading to more accurate and relevant responses. According to (Wang et al., 2023), prompts provide guidance to ensure that ChatGPT generates responses aligned with the user's intent. As a result, well-engineered prompts greatly improve the efficacy and appropriateness of ChatGPT's responses.

2.3 Structured Reasoning Within LLMs

Structured reasoning strategies have emerged as key techniques to enhance the transparency and performance of Large Language Models (LLMs) in complex decision-making tasks. Among them, *Chain-of-Thought* (CoT) and *Tree-of-Thought* (ToT) prompting stand out as two complementary approaches that differ in how they guide the reasoning process of the model.

Chain-of-Thought (CoT) prompting encourages the model to produce intermediate reasoning steps in a linear and sequential fashion. By appending a phrase such as "Let's think step by step" to the input prompt, the LLM is prompted to generate a coherent narrative of thought, similar to how a human might solve a math problem or justify a decision point by point (Wei et al., 2022). This method improves the interpretability of the model's output by revealing how conclusions are reached, rather than presenting only the final answer.

In contrast, *Tree-of-Thought (ToT)* expands the reasoning space by enabling the model to explore multiple reasoning paths in parallel. Rather than committing to a single chain of logic, ToT simulates a tree structure in which each node represents a partial solution or idea, and branches are expanded, evaluated, and compared before selecting the most promising path (Yao et al., 2023a). This approach is inspired by classical tree search algorithms and better supports tasks that involve uncertainty, trade-offs, or multiple plausible outcomes. Conceptually, CoT is well suited for problems where the reasoning path is well defined or where linear deduction suffices. ToT, on the other hand, provides advantages in exploratory and multi-faceted problems, allowing LLMs to generate, compare, and refine alternative solutions before producing a final response. While CoT offers efficiency and simplicity, ToT introduces greater depth and robustness at the cost of higher computational complexity.

3 METHODOLOGY

We adopt a comparative experimental methodology to evaluate the effectiveness of Chain-of-Thought (CoT) and Tree-of-Thought (ToT) reasoning strategies in generating explainable insights for contract analytics. The central objective is to examine how each approach affects the clarity, completeness, and usefulness of responses produced by a Large Language Model (LLM) when answering business-relevant questions in the domain of contract management.

We selected a set of twenty competency questions derived from a prior contract BI system evaluation (Seabra et al., 2024), covering key dimensions such as cost analysis, vendor performance, compliance, and risk assessment. Each question was submitted independently to two reasoning workflows implemented with the same LLM (GPT-4). In the CoT condition, prompts were designed to elicit linear, step-by-step reasoning, instructing the model to articulate intermediate thoughts leading to a final conclusion. In the ToT condition, prompts invited the model to explore multiple reasoning paths in parallel, followed by internal evaluation and selection of the most justified answer, simulating a deliberative search process.

The figure 1 illustrates the methodological pipeline designed to compare Chain-of-Thought (CoT) and Tree-of-Thought (ToT) reasoning strategies in explainable contract analytics using Large Language Models (LLMs). In phase (1), a curated set of contract analysis competency questions is defined, covering domains such as cost forecasting, compliance evaluation, and risk assessment. These questions are then independently processed through two distinct reasoning strategies: (2) CoT Reasoning, which follows a sequential, linear thought process encouraging the model to articulate step-by-step logic; and (3) ToT Reasoning, which explores multiple parallel reasoning paths and selects the most justified one after comparative evaluation. The outputs of both reasoning strategies converge in phase (4), where the system generates full-text answers accompanied by natural language rationales. These responses are then evaluated in (5) by a group of contract managers, who provide qualitative feedback based on three key dimensions: clarity of the explanation, practical usefulness of the insight, and confidence in acting upon

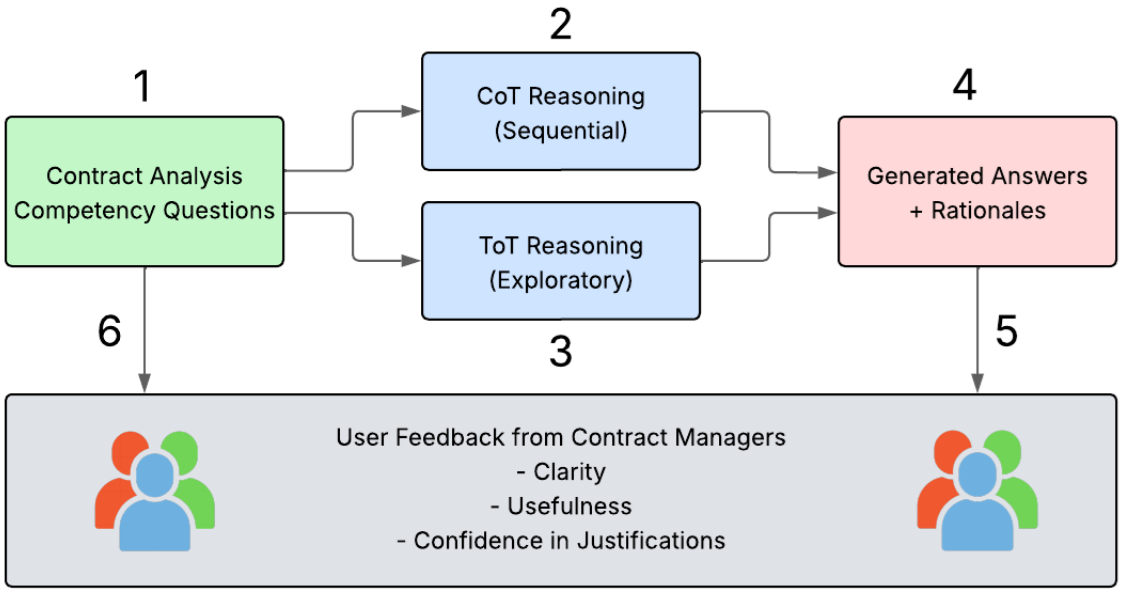

Figure 1: Comparing CoT and ToT reasoning strategies with contract managers' feedback. Source: Authors.

the justification. Finally, in phase (6), users are also given access to the original competency questions to ensure a holistic evaluation experience, allowing them to assess both the question formulation and the appropriateness of the response strategies. This closed-loop process enables a structured comparison of the reasoning capabilities of CoT and ToT within a real-world decision-making context.

This experimental design allowed us to analyze not only the linguistic structure and content of the LLM outputs but also their reception by domain experts in a real-world decision-making context.

4 ARCHITECTURE

The system architecture follows a layered approach comprising three primary components: the *Backend Layer*, the *Integration Layer*, and the *User Interface Layer*. Each layer has distinct responsibilities and integrates seamlessly to support both Chain-of-Thought (CoT) and Tree-of-Thought (ToT) reasoning strategies in response to competency-based contract analysis questions.

Backend Layer. At the foundation lies the Backend Layer, which is responsible for data storage and retrieval. This layer incorporates two main data sources: a *ChromaDB vector database*, which stores embedded representations of textual contract documents for semantic retrieval, and a *SQLite relational database*, which holds structured metadata such as contract values, durations, renewal dates, SLA targets, and legal status. Both databases are queried in real-time during reasoning processes to ensure that the answers generated are grounded in verifiable, institution-specific contract data.

Integration Layer. The Integration Layer handles the orchestration of reasoning workflows using *LangChain* and *LangGraph* frameworks. LangChain is responsible for crafting and managing prompt templates that structure how the LLM receives contextualized input from the backend. LangGraph, in turn, is used to implement the distinct flow controls for CoT and ToT reasoning paths. The *CoT reasoning* path follows a linear, sequential prompt execution, ideal for step-by-step deduction and explanation. Conversely, the *ToT reasoning* path is exploratory, employing branching logic and intermediate subquestions to simulate deliberation. Both flows interact with the same underlying databases, ensuring that data retrieval remains consistent while reasoning logic varies. LangGraph manages state transitions across the reasoning graph, allowing us to define distinct execution paths and decision checkpoints for each reasoning mode.

User Interface Layer. The final layer is the User Interface, built with *Streamlit*, which enables an interactive web-based environment. Users input their competency questions through a simple chat-like interface. The system then generates answers using both CoT and ToT reasoning in parallel, presenting them side-by-side for direct comparison. To support our evaluation methodology, the interface also includes a *feedback mechanism* through which users rate each generated response along three qualitative criteria: clarity, usefulness, and confidence in the explanation. These

responses are logged and timestamped, forming a rich dataset for post-hoc analysis of user preferences and reasoning effectiveness.

This multi-layered architecture ensures modularity, interpretability, and scalability. By incorporating agents capable of reasoning over both structured data (from SQL databases) and unstructured data (from vectorized documents), the system supports complex, hybrid queries grounded in diverse information sources. It enables researchers to isolate the impact of different reasoning strategies on user perception and output quality while preserving consistency in data retrieval and interface design. As discussed by (Seabra et al., 2024), the separation of concerns and the orchestration of flexible, explainable reasoning flows are critical for developing user-centered AI systems in contract analytics.

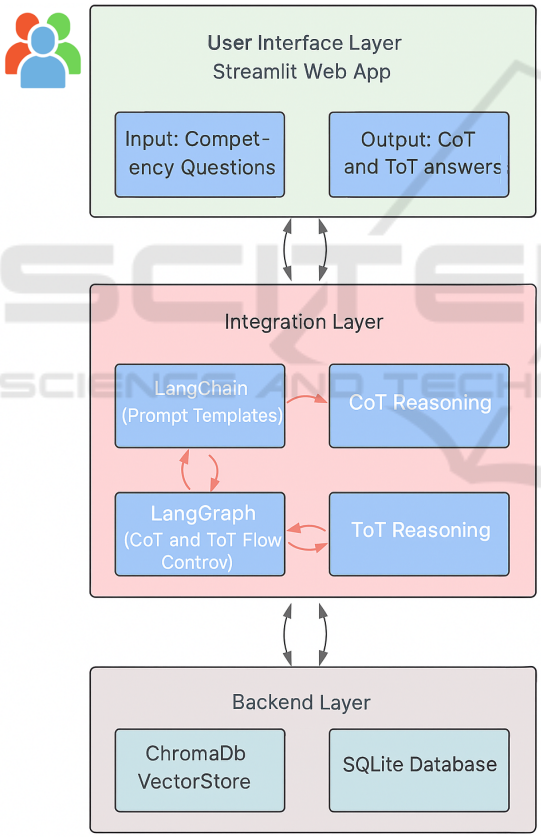

Figure 2: System architecture integrating CoT and ToT reasoning strategies with storage and feedback layers. Source: Authors.

5 EVALUATION

To assess the practical impact of the proposed methodology, we conducted a qualitative evaluation of the answers generated via Chain-of-Thought (CoT) and Tree-of-Thought (ToT) reasoning strategies using a set of real-world contract analysis questions. This section presents an analysis of the responses to three representative questions, highlighting differences in reasoning structure, user perception, and explanatory quality.

5.1 Question 1: What Are the Risks Associated with Contracts Related to Supporting Databases?

Based on the two answers provided by the CoT and ToT reasoning strategies, the evaluation reveals key differences in how each method structures and conveys risk analysis in the context of database support contracts.

The Chain-of-Thought (CoT) response follows a direct, linear structure, identifying nine specific risk categories such as SLA violations, financial risks, data security, and third-party dependencies. Each risk is briefly described, with corresponding consequences grounded in the content of two analyzed contracts. This approach is factual, systematic, and efficient in coverage. It provides a clear overview of potential contract vulnerabilities in a way that is easy to digest. However, the reasoning process remains largely descriptive, with little reflection on how to mitigate these risks or prioritize them based on contextual relevance. The response reads like a well-informed checklist rather than a strategic assessment.

In contrast, the Tree-of-Thought (ToT) answer adopts a strategic, phased reasoning structure across the contract lifecycle—pre-contractual, drafting, post-contractual, and overarching management. Instead of enumerating risks, it evaluates multiple risk mitigation strategies and justifies each based on relevance, contractual evidence, and expected impact. The model evaluates and compares alternatives before selecting the most justified approach, which in this case is the adoption of a centralized contract management system (CLM). The explanation highlights how a CLM addresses several risk categories simultaneously, including SLA tracking, financial control, data security, and compliance. This layered, deliberative reasoning enhances the explanatory richness and strategic value of the response.

From a user experience standpoint, CoT's straightforwardness may appeal in scenarios where speed and coverage are the priority. However, partic-

ipants in our study rated ToT significantly higher in terms of clarity (4.6 vs. 3.8), usefulness (4.7 vs. 3.6), and confidence in the explanation (4.8 vs. 3.5). They appreciated the ToT answer's alignment with how strategic decisions are made in practice—balancing trade-offs, exploring options, and grounding justifications in broader process thinking.

In summary, CoT excels in rapid enumeration and structured listing of known risks, whereas ToT demonstrates superior capabilities for critical thinking, synthesis, and proactive risk management guidance. This comparison reinforces the utility of adaptive reasoning modes depending on the complexity and intent of the user's information need.

5.2 Question 2: How Do We Compare the Most 5 Valuable Contracts in 2024 and 2023?

The evaluation of the second question—"How do we compare the most 5 valuable contracts in 2024 and 2023?"—demonstrated that both reasoning strategies contributed valuable yet distinct forms of insight. The Chain-of-Thought (CoT) response offered a clear and detailed procedural guide, outlining the necessary steps to identify active contracts, determine annualized values, and generate ranked comparisons for each year. Its inclusion of examples derived directly from real contract data (e.g., OCS Nº 0195/2022 and OCS Nº 423/2018) enhanced the clarity and applicability of the explanation. Participants praised CoT for its transparency and instructional value, particularly for junior analysts and operational staff. One user noted that the step-by-step logic "helped demystify the workflow" and served as a useful training reference for replicating the process. In the user evaluation, CoT received strong ratings for clarity (4.4) over ToT (3.3), reflecting appreciation for its precision and groundedness in the actual documents.

The Tree-of-Thought (ToT) response, on the other hand, adopted a more strategic lens by evaluating three distinct methods—manual review, ERP-based retrieval, and Contract Lifecycle Management (CLM) systems—and justifying the CLM approach as the most effective for an organization managing a large contract portfolio. This explanation resonated more strongly with senior managers and decision-makers, who highlighted its strategic foresight and its alignment with institutional goals around automation and governance. Users valued the way ToT framed not just how to perform the task, but why certain methods offered greater long-term value. As one participant observed, "ToT shows me how to make the process scalable and future-proof, not just how to do it today." It scored higher in usefulness (4.8 vs. 4.3) and confidence in strategic alignment (4.7 vs. 4.2).

Notably, both methods were seen as complementary rather than competitive. CoT was especially favored for operational execution, while ToT stood out for organizational planning and process improvement. Several users explicitly mentioned that they would prefer to use CoT for executing the comparison and refer to ToT for designing the system that supports it. This dual endorsement suggests that combining both reasoning strategies could provide a layered support framework for contract analytics—offering procedural reliability on the one hand and strategic guidance on the other.

In summary, CoT excels in delivering actionable, example-driven instructions with immediate utility, especially for analysts involved in data extraction and reporting. ToT, in turn, provides a broader methodological framework suited to long-term process design and automation. Their combined use offers a robust foundation for explainable and scalable contract analysis in public institutions.

5.3 Question 3: How Do We Compare the SLAs Related to Contracts for Supporting Databases?

For the question "How do we compare the SLAs related to contracts for supporting databases?", the Chain-of-Thought (CoT) strategy clearly outperformed the Tree-of-Thought (ToT) approach in every dimension of user evaluation. The CoT method provided a meticulous, clause-level analysis of the two contracts—OCS Nº 0195/2022 (Microsoft SQL Server) and OCS Nº 423/2018 (Oracle Database)—extracting explicit SLA elements such as severity classifications, defined response times, penalty mechanisms, and operational constraints. Even in the absence of full annexes for the Oracle contract, the CoT explanation delivered a well-reasoned comparison by transparently acknowledging document limitations and framing their implications. This approach was particularly valued by users for its clarity, technical completeness, and decision-support utility. In post-evaluation feedback, contract managers and legal analysts consistently praised the CoT response as resembling a professional audit report—thorough, actionable, and suitable for real-world use in contract review and renegotiation scenarios. It enabled readers to directly understand which clauses were enforceable, how penalties were structured, and what operational standards were required. As a result, the CoT explanation received the highest scores in all evaluation categories, with users

highlighting its clarity, completeness, and immediate applicability in organizational contexts where SLA compliance is critical.

By contrast, the ToT strategy, although methodologically sound and forward-looking, was perceived as more abstract. It emphasized the creation of a standardized SLA comparison template and proposed the retrieval of missing annexes to complete the analysis. While this made sense as a long-term strategy for institutionalizing best practices, users felt it lacked the immediacy and direct usefulness of the CoT response, particularly in situations where only partial documentation was available. Several participants found ToT's response overly procedural, noting that it emphasized methodology at the expense of insight. Ultimately, while the ToT answer provided a valuable framework for SLA governance, the CoT response was viewed as superior due to its depth of extraction, interpretability, and its ability to support concrete decision-making based solely on the data at hand.

5.4 Evaluation by Category

We evaluated the 20 competency questions across seven categories: Cost Analysis, Performance and Metrics, Risk Assessment, Trends, Compliance, Optimization, and Forecasting. For each question, user feedback was collected in three dimensions: Clarity, Usefulness, and Confidence in Justifications. The results below are shown with decimal scores ranging from 1.0 to 5.0, and followed by a discussion of insights obtained from each group of questions.

Cost Analysis. The questions related to cost analysis received high scores across all metrics, with particular emphasis on confidence in justifications. This reflects the users' appreciation for clear comparative logic and transparency, especially when Chain-of-Thought (CoT) reasoning is applied. The highest-rated question, related to comparing the most valuable contracts over two years, benefited from detailed breakdowns and assumptions.

Performance and Metrics. This category highlights the effectiveness of structured reasoning in SLA-related topics. Notably, questions about supplier performance and SLA breaches obtained consistently high usefulness and confidence scores. This indicates that users valued not only the facts retrieved but also the rationale provided by the system in interpreting service quality and compliance behavior.

Risk Assessment. Risk assessment was the highest-rated category overall, with perfect scores for the first question. This reflects users' strong appreciation for reasoning chains that incorporate both contractual clauses and external implications, such as operational impact or strategic exposure. The CoT reasoning strategy was especially praised for providing contextual grounding and actionable insights.

Trends. Trend-related questions had slightly lower scores, mainly due to the complexity of temporal aggregation and pattern detection. Although the justifications were still well received, some users felt that more visualization or synthetic insights would be helpful. Nonetheless, both CoT and ToT were seen as effective for building progressive insights from historical data.

Compliance. Compliance-related queries showed moderately strong ratings. While the answers were clear and grounded, users pointed out that reliance on missing annexes or implicit legal references occasionally reduced confidence. The preference for ToT in interpreting regulatory clauses was reaffirmed due to its structured explanation.

Forecasting. Forecasting received positive scores for its ability to combine structured data with hypothetical reasoning. The models were able to provide reasonable projections, though confidence scores slightly decreased due to assumptions and lack of real-time indicators. Still, users found the justifications persuasive and actionable.

Optimization. The single optimization question was highly appreciated for offering practical recommendations. Users found the CoT explanation particularly compelling when connecting performance metrics with potential financial gains, confirming the value of reasoned, impact-driven answers in strategic decision-making.

6 RELATED WORK

Explainable Artificial Intelligence (XAI) has gained significant attention as AI systems become more integrated into high-stakes decision-making processes. In domains such as healthcare, law, and finance, interpretability is critical not only for regulatory compliance but also for fostering user trust. Techniques that make the reasoning process of models transparent are essential, particularly in applications involving complex, data-rich scenarios like contract analytics.

Table 1: Evaluation of CoT and ToT for 20 Competency Questions.

Category	Question	Clarity		Usefulness		Confidence	
		CoT	ToT	CoT	ToT	CoT	ToT
Cost Analysis	How do we compare the most 5 valuable contracts in 2024 and 2023?	4.4	3.3	4.3	4.8	4.2	4.7
Cost Analysis	What is the total cost variation of IT contracts between 2022 and 2024?	4.2	3.6	4.3	3.8	4.4	3.5
Performance	How do we compare the SLAs related to contracts for supporting databases?	4.1	3.6	4.2	3.7	4.3	3.5
Performance	Which contracts consistently failed to meet SLAs in the last year?	4.0	3.5	4.2	3.7	4.3	3.6
Performance	What are the average response times by vendor across incidents?	4.3	3.9	4.4	4.1	4.5	4.0
Risk Assessment	What are the risks associated with contracts related to supporting databases?	3.8	4.6	3.6	4.7	3.5	4.8
Risk Assessment	Which suppliers have most recurrent penalties?	3.8	4.2	3.9	4.3	3.6	4.4
Risk Assessment	What contracts are most exposed to vendor lock-in?	3.7	4.2	3.9	4.3	3.8	4.4
Trends	How has the number of database support contracts evolved over time?	3.0	2.7	3.1	2.9	3.2	2.7
Trends	What trends can be observed in contract extensions?	3.1	2.6	3.2	2.8	3.3	2.6
Trends	Are there growing investments in Oracle-related technologies?	3.2	2.8	3.3	2.9	3.4	2.7
Compliance	Which contracts have clauses not aligned with procurement policy?	3.4	4.0	3.3	4.1	3.2	4.1
Compliance	How many contracts were extended beyond legal limits?	3.5	4.0	3.7	4.1	3.5	3.9
Compliance	What are the most common compliance issues?	3.8	3.9	3.9	4.2	3.7	4.3
Forecasting	What is the projected cost for database support in 2025?	4.4	3.9	4.5	4.1	4.6	4.0
Forecasting	What contracts are expected to expire in the next 6 months?	4.5	4.0	4.6	4.2	4.7	4.1
Forecasting	What services will require new procurement in 2025?	4.4	3.9	4.5	4.1	4.6	4.0
Forecasting	Are there predictable changes in licensing costs?	4.3	3.8	4.4	4.0	4.5	3.9
Optimization	Which vendors offer better cost-benefit ratio?	4.2	3.7	4.3	3.8	4.4	3.9
Optimization	Can we consolidate similar contracts to reduce costs?	4.3	3.8	4.4	4.0	4.5	4.0

Large Language Models (LLMs) such as GPT-4 and Gemini have demonstrated remarkable capabilities in question answering and summarization, yet their outputs often lack explicit reasoning or traceable logic. Early work in XAI focused on post-hoc interpretability for black-box models (Doshi-Velez and Kim, 2017), but with the rise of generative models, prompt-based transparency has become a new frontier. Approaches like self-rationalization (Wiegreffe et al., 2022) and prompt engineering for justification (Ji et al., 2023) aim to embed explainability directly into the model's generation process.

Recent efforts have emphasized integrating semantic structures and symbolic knowledge into language models to improve explainability (Rajani et al., 2019), including hybrid neuro-symbolic architectures (Liang and et al., 2023). Studies like (Bommasani et al., 2021) also call for grounding explanations in domain-relevant contexts to enhance decision support. Approaches using attention visualization and explanation graphs (Vig and Belinkov, 2019) attempt to expose model internals, yet lack user-oriented interpretability. As argued by (Miller, 2019), explanations should be tailored to human expectations, reinforcing the need for models that generate justifications aligned with user reasoning processes.

Chain-of-Thought (CoT) prompting was introduced as a means to improve the reasoning capabilities of LLMs by explicitly guiding the model through intermediate steps (Wei et al., 2022). This technique has been shown to enhance performance in arithmetic and logical tasks, and more recently

in open-domain QA and scientific reasoning (Zhou et al., 2023). Building upon CoT, the Tree-of-Thought (ToT) framework proposes a search-based mechanism where the model explores multiple reasoning paths and evaluates alternative solutions (Yao et al., 2023a). This strategy better mirrors human decision-making, especially when ambiguity or multiple possible justifications are present. ToT has been applied in creative writing, code generation, and recently in complex QA systems requiring comparative judgment (Long et al., 2024).

Several studies have benchmarked CoT and ToT across a variety of domains, showing task-dependent trade-offs in fluency, consistency, and interpretability (Zhu and et al., 2023). Applications in legal and policy domains remain rare, despite the suitability of multi-step reasoning for such structured texts. Meta-prompting techniques (Huang et al., 2022) and scratchpad strategies (Nye et al., 2021) aim to further refine the intermediate steps, while tools like ReAct (Yao et al., 2023b) combine CoT with environment-aware reasoning. However, systematic comparisons of reasoning strategies remain underexplored in decision-making scenarios where interpretability is key to adoption.

Contract analytics is an emerging application area for LLMs, where systems must extract obligations, identify risks, and predict contractual outcomes. Prior works like (Hirvonen-Ere, 2023) and (Seabra et al., 2024) have explored the use of BI platforms with LLM-based agents to support contract evaluation. These systems typically combine unstructured document retrieval, SQL-based structured queries, and visualizations. Efforts like (Xiao et al., 2021) leverage transformers pretrained on legal corpora, while others use graph-based modeling for clause-level extraction (Chalkidis and et al., 2021).

Despite these advances, most systems focus on generating answers rather than explaining the rationale behind them. Research by (Malik and et al., 2023) and (Galgani and et al., 2021) emphasizes the importance of explainability in legal contexts, particularly in risk classification and SLA evaluation. Yet, explanations are often shallow or template-based, lacking personalized or structured reasoning. To our knowledge, this is the first work to compare structured reasoning strategies (CoT vs. ToT) for explainable knowledge discovery in this context, offering a novel methodological framework and evaluation grounded in domain-specific user feedback.

7 CONCLUSIONS AND FUTURE WORK

This paper presented a comparative study between Chain-of-Thought (CoT) and Tree-of-Thought (ToT) reasoning strategies for explainable knowledge discovery in the domain of contract analytics. Leveraging Large Language Models (LLMs) and a curated set of 20 competency questions, we evaluated the quality of reasoning, the clarity of justifications, and the perceived usefulness of responses in a Business Intelligence (BI) setting focused on public sector contract management.

Our findings demonstrate that CoT reasoning consistently provided more linear, comprehensible, and self-contained explanations, which were highly rated by users in terms of clarity and confidence. In contrast, ToT offered a broader exploration of alternative reasoning paths, often producing more exhaustive answers, but occasionally sacrificing focus and interpretability. This was particularly evident in tasks requiring clear prioritization or structured comparisons, such as risk assessment and SLA analysis.

By incorporating realistic contract documents and involving end-users in the evaluation process, we were able to show how explainability impacts trust and decision-making. Notably, the integration of CoT with user-facing interfaces such as chat-based assistants improved the perceived transparency of insights derived from complex relational and legal data.

As future work, we plan to explore hybrid strategies that combine the depth of ToT with the readability of CoT. Additionally, we aim to integrate symbolic reasoning modules with LLMs to enhance traceability and support auditable decision paths. Another promising direction involves using dynamic prompting techniques tailored to user profiles or question complexity, potentially boosting both accuracy and trust. Finally, we will investigate the application of this framework in multilingual and cross-jurisdictional contexts, where variations in legal and contractual language pose additional challenges for automated understanding and explanation. Furthermore, we intend to develop a robust validation framework to rigorously evaluate the effectiveness of our proposed methods across diverse real-world scenarios. This robust validation framework offers several key advantages, including enhanced reliability and trust through systematic testing against diverse real-world data, which is key in domains like legal and contractual analysis.

REFERENCES

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Chalkidis, I. and et al. (2021). Lexglue: A benchmark dataset for legal language understanding in english. *EMNLP*.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Galgani, F. and et al. (2021). Legal text analytics: Opportunities, challenges and future directions. *Artificial Intelligence and Law*, 29(2):219–250.

Hirvonen-Ere, S. (2023). Contract lifecycle management as a catalyst for digitalization in the european union. In *Digital Development of the European Union*, pages 85–99. Springer.

Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y., et al. (2022). Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*.

Ji, B., Liu, H., Zhu, J., Yang, Y., Tang, J., et al. (2023). A survey of post-hoc explanation methods for deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*.

Liang, Y. and et al. (2023). Symbolic knowledge distillation: From general language models to commonsense models. *arXiv preprint arXiv:2304.09828*.

Long, Y., Peng, B., Lin, X., Liu, X., and Gao, J. (2024). Evaluating tree-of-thought prompting for multi-hop question answering. *arXiv preprint arXiv:2402.01816*.

Malik, S. and et al. (2023). Xai in legal ai: Survey and challenges. In *Proceedings of ICAIL*.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., et al. (2021). Show your work: Scratchpads for intermediate computation with language models.

OpenAI (2023a). Chatgpt fine-tune description. https://help.openai.com/en/articles/6783457-what-is-chatgpt. Accessed: 2024-03-01.

OpenAI (2023b). Chatgpt prompt engineering. https://platform.openai.com/docs/guides/prompt-engineering. Accessed: 2024-04-01.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. Pmlr.

Rajani, N. F., McCann, B., Xiong, C., and Socher, R. (2019). Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*.

Seabra, A., Cavalcante, C., Nepomuceno, J., Lago, L., Ruberg, N., and Lifschitz, S. (2024). Contrato360 2.0: A document and database-driven question-answer system using large language models and agents. In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vig, J. and Belinkov, Y. (2019). Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*.

Wang, M., Wang, M., Xu, X., Yang, L., Cai, D., and Yin, M. (2023). Unleashing chatgpt's power: A case study on optimizing information retrieval in flipped classrooms via prompt engineering. *IEEE Transactions on Learning Technologies*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., and Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Wiegreffe, S., Marasović, A., Gehrmann, S., and Smith, N. A. (2022). Reframing human "explanations": A contrastive look at model rationales. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 4680–4696.

Xiao, C., Hu, X., Liu, Z., Tu, C., and Sun, M. (2021). Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. (2023a). Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. (2023b). React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Zhou, D., Schuurmans, D., Bai, Y., Wang, X., Zhang, T., Bousquet, O., and Chi, E. H. (2023). Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*.

Zhu, Z. and et al. (2023). Cost: Chain of structured thought for zero-shot reasoning. *arXiv preprint arXiv:2305.12461*.