Tree-of-Thought-Augmented LLMs for Automated Document Review Across Industrial Domains

Zehui Zhu[1], **Mohamad Iliya Haziq Mohamad Lokman**[1], **Meenakshi Mishra**[1], **Peng Xu**[1], **Jennifer Shin**[2], **Hua Qian**[2], **Zhifu Shu**[1]

[1]ExxonMobil Technology and Engineering Company
[2]ExxonMobil Biomedical Sciences, Inc.
{zehui.zhu, mohamad.mohamadlokman, meenakshi.mishra1, peng.xu, jennifer.shin, hua.qian, zhifu.shu}@exxonmobil.com

Abstract

We present AI Screener, an end-to-end automated document review system that integrates a 12-billion-parameter pretrained large language model with a Tree-of-Thought reasoning framework to emulate and scale expert-level decision-making. Designed for high-stakes, domain-specific analysis, AI Screener empowers subject matter experts to encode their domain knowledge and reasoning processes in a no-code, efficient manner—enabling rapid customization without technical barriers. The system has been deployed across three different and unrelated mission-critical business functions: (1) accelerating scientific literature reviews to support the development of occupational exposure limits for worker health protection, (2) streamlining patent screening to optimize intellectual property portfolio management, and (3) automating procurement contract analysis to identify value leakage and drive better commercial terms. Across these diverse deployments, subject matter experts encoded their knowledge with AI Screener to transform traditional workflows—significantly reducing manual review time while maintaining expert-grade accuracy and consistency. This work highlights how Tree-of-Thought-augmented LLMs can be pragmatically applied to reshape enterprise document intelligence at scale.

Introduction

The review and analysis of large volumes of specialized documents are essential to critical business operations across sectors. Yet, these processes often create significant bottlenecks due to their reliance on expert-driven workflows. In practice, organizations have typically adopted one of two approaches. The first—and still most common—is direct manual review by professionals with deep domain expertise: health scientists combing through scientific literature to establish occupational exposure limits, patent analysts evaluating filings to manage intellectual property portfolios, or commercial teams scrutinizing procurement contracts to uncover value leakage and improve terms. While this approach delivers high-quality insights, it is time-consuming, costly, and difficult to scale. The second approach involves building task-specific machine learning models, which requires experts to label a representative subset of documents to train classifiers. Although this can accelerate throughput, it demands substantial upfront investment in data annotation and model development for each new use case, limiting its adaptability across domains.

In this work, we propose a more efficient paradigm. We introduce a system that utilizes a relatively small 12-billion parameter large language model (LLM) augmented with a Tree-of-Thought (ToT) reasoning framework (Yao et al. 2023). This approach enables the automated screening of complex documents by directly following expert-specified logic articulated through a series of structured prompts, thereby bypassing the need for manual labeling and domain-specific model training. We demonstrate its applicability across diverse industrial use cases, including (1) scientific literature review for worker health protection, (2) patent screening for intellectual property portfolio management, and (3) procurement contract analysis for value leakage identification and recapture.

The primary contributions of this work are as follows:

1. We developed and deployed a scalable, multi-domain automated document review system. This system proves to be robust and effective across fundamentally different industrial applications, demonstrating a generalizable solution to a widespread business challenge without requiring task-specific architectural changes.
2. We demonstrated the effectiveness of the ToT framework in simulating a human expert's decision-making process using an LLM. Our results show that by structuring the model's reasoning path, even a relatively small 12B parameter model can achieve high accuracy in complex classification and analysis tasks.

Application Development and Deployment

Our system is engineered as a modular application to facilitate robust and scalable automated document review. As shown in Figure 1, the architecture consists of three core components: a data ingestion module, an LLM serving module, and a ToT reasoning engine.

Data Ingestion Module

To accommodate diverse document review use cases, the application's data ingestion module is designed to parse multiple file formats. The implementation supports three primary types:

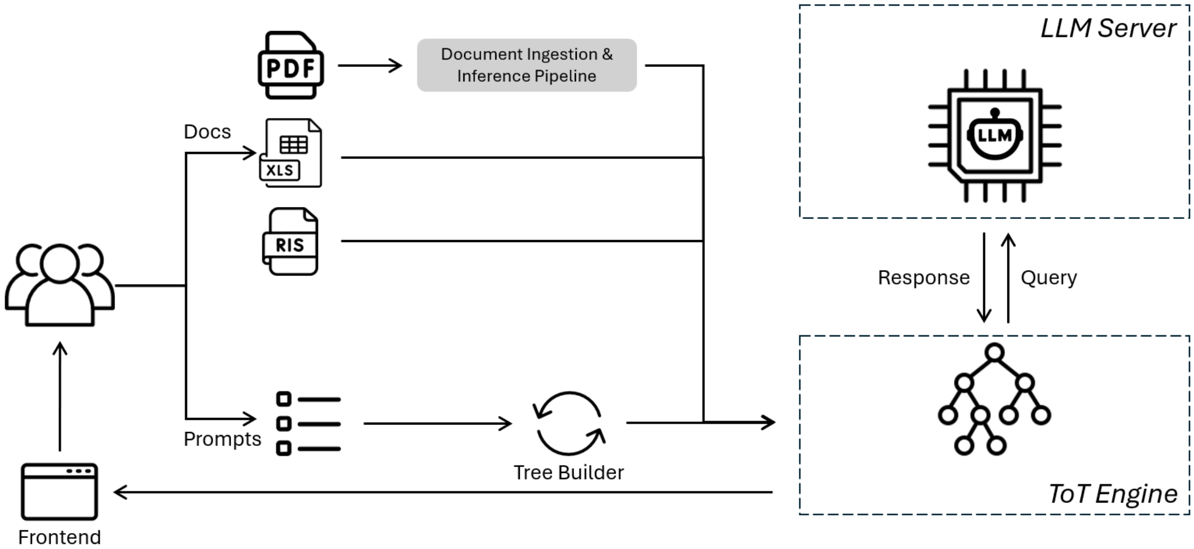

Figure 1: Modular Architecture for Automated Document Review. Our system has three core components: a data ingestion module, an LLM serving module, and a ToT reasoning engine, designed for robust and scalable automated document review.

- **RIS Files:** This standardized tag format is commonly used for bibliographic data and is central to academic literature reviews. The module ingests `.ris` files, parsing metadata fields such as title, author, publication year, journal, abstract, and keywords[1]. This structured data becomes the context for the LLM to evaluate the relevance of research articles against predefined inclusion/exclusion criteria.
- **XLS(X) Files:** For intellectual property analysis, the system accepts `.xls` or `.xlsx` files. These spreadsheets typically contain patent data exported from search databases, with columns for patent number, title, abstract, and the full text of claims. The ingestion module extracts these fields, preparing the patent's core information for substantive review by the ToT engine.
- **PDF Files:** For procurement contract analysis, the system processes `.pdf` files through a multi-stage ingestion pipeline designed to build a rich contextual foundation. As shown in Figure 2, the process begins by converting contracts into markdown-formatted text using either a basic parser or Optical Character Recognition (OCR) tools. Following conversion, an LLM performs two critical extraction tasks:
  - It generates key metadata by analyzing the most informative portions of the document (e.g., the first 50,000 tokens) to identify attributes such as the document name, effective date, and type (e.g., base document, amendment).
  - It extracts the glossary of definitions to establish a global semantic context that enhances the interpretability of the contract's terms.

  This prepared context, consisting of both structured metadata and semantic definitions, is then utilized by the inference engine. When a user poses a query, the system uses this foundation to deconstruct the query into high-level questions, extract relevant quotes from the document, and synthesize a final, contextually-aware answer that prioritizes the most current and relevant information based on the extracted metadata.

LLM Serving Module

The foundation of our system's analytical capability is an instruction-tuned LLM. We employed the 12-billion parameter Mistral-Nemo-Instruct-2407 model, which offers a strong balance of performance and resource efficiency (Mistral AI and NVIDIA 2024).

The model is deployed on a dedicated node equipped with four NVIDIA A100 GPUs. To optimize inference speed and throughput, we implemented two key techniques:

1. **Half-Precision (FP16):** The model weights are loaded in 16-bit floating-point format, which halves the memory footprint and significantly accelerates computation compared to full-precision (FP32) inference.
2. **Flash Attention 2:** We leverage the Flash Attention 2 algorithm, an I/O-aware attention mechanism that avoids materializing the large attention matrix in GPU high-bandwidth memory (Dao 2023). This results in faster execution and reduced memory usage, enabling the processing of longer document contexts common in patent claims, research articles, and contracts.

ToT Engine Module

To move beyond simple keyword matching or summary generation and enable nuanced, human-like reasoning, we integrated a ToT engine. The ToT framework allows an LLM to explore multiple reasoning paths, evaluate intermediate thoughts, and backtrack when necessary, thereby improving

[1] https://pypi.org/project/rispy/

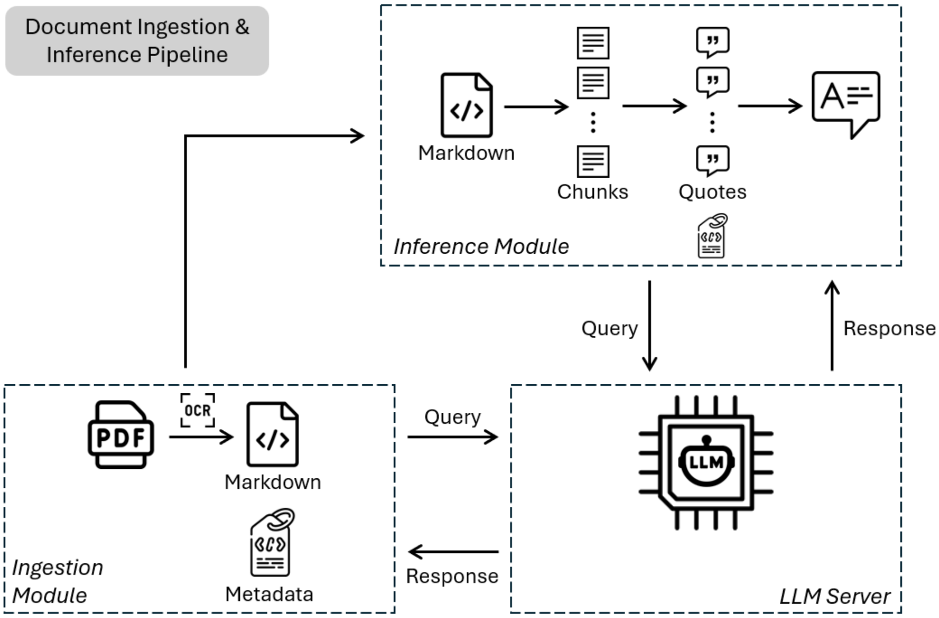

Figure 2: Multi-Stage Ingestion Pipeline for Procurement Contract Analysis. We convert `.pdf` contracts to markdown, then an LLM extracts key metadata and the glossary of definitions. This contextual foundation then enables the inference engine to process user queries, extract relevant information, and generate contextually-aware answers.

the quality and reliability of its final conclusion (Yao et al. 2023).

In our implementation, the ToT engine constructs binary decision trees to systematically address a review question (e.g., ”Does the patent mention seaweed or macroalgae as feedstock or raw material?”). At each node of the tree, the LLM is prompted to make a discrete decision. We utilize function calling to structure the model's output, requiring it to return a JSON object containing two fields (Basu 2024):

1. `answer`: A strict binary response (`true` or `false`).
2. `reasoning`: A clear and concise natural language explanation for its decision at that step.

This approach forces the model to externalize its thought process, creating a transparent and auditable reasoning chain. By decomposing a complex review into a series of verifiable binary decisions, the ToT engine enhances the accuracy of the final judgment and provides human reviewers with a clear rationale for validation.

Application Use

This section details the application of our Tree-of-Thought-augmented LLM system in three distinct, high-value industrial use cases: (1) scientific literature review for the development of occupational exposure limits (OELs), (2) patent screening for intellectual property portfolio management, and (3) procurement contract analysis for value leakage identification.

Scientific Literature Review for OEL Development

Introduction An OEL refers to the airborne concentration under which it is believed that nearly all workers may be repeatedly exposed, day after day, over a working lifetime, without adverse health effects (American Conference of Governmental Industrial Hygienists (ACGIH) 2025). Establishing science-based OELs is critical for worker health protection. The process requires an exhaustive review of scientific literature to identify all relevant evidence, a task that is traditionally a time-consuming bottleneck performed by subject matter experts. A typical keyword-based search in scholarly databases can yield thousands of articles, the vast majority of which do not contain information relevant for OEL setting.

Application We deployed AI Screener to accelerate the initial relevance assessment of scientific articles for an OEL for hydrogen sulfide. A cohort of three health science experts manually reviewed the titles and abstracts of 1,649 articles, a process that required over 40 work hours. Concurrently, we encoded the experts' screening logic into the AI system. The core logic for classifying an article as ”Relevant” required it to meet all of the following criteria:

1. **Publication Type:** Must be a primary research study.
2. **Language:** Must be originally published in English.
3. **Study Focus:** Must not be primarily a methods-development paper.
4. **Exposure Context:** Must pertain to occupational or workplace exposure to hydrogen sulfide.
5. **Endpoint:** Must evaluate health effects or hazards associated with hydrogen sulfide exposure.

Quantitative Results The AI Screener processed all 1,649 articles in under three hours. The performance, when compared against the expert consensus labels, is summarized in Table 1.

Table 1: Confusion Matrix for Scientific Literature Screening.

	Human: Relevant	**Human: Non-Relevant**
AI: Relevant	34 (TP)	81 (FP)
AI: Non-Relevant	12 (FN)	1522 (TN)

The system achieved an accuracy of 94.4% and a recall of 73.9% for the ”Relevant” class. Precision is less critical in this context, since a limited number of false positives only result in minimal additional human review effort. The OEL development team’s primary concern was the 12 false negatives (articles deemed relevant by humans but non-relevant by the AI). A senior expert panel conducted a detailed review of these 12 articles and found that:

- The root cause was the overly narrow classification and exclusion criteria—specifically, the occupational exposure question was too narrowly defined. This inadvertently led the AI Screener to exclude non-occupational but toxicologically relevant studies that had been included by human reviewers.

The application demonstrated that the AI Screener can substantially reduce manual effort—from weeks to hours—while maintaining high accuracy, thereby resolving a longstanding bottleneck in the OEL development process and enabling health scientists to redirect their expertise toward higher-value tasks. The analysis of discrepancies highlighted the critical importance of iterative prompt engineering to fully capture the complex logic of domain experts.

Patent Screening for Intellectual Property Portfolio Management

Introduction Effective management of intellectual property risk is essential for protecting corporate assets and avoiding reputational or financial harm. Traditional keyword-based patent searches often yield large volumes of irrelevant results due to the complexity of patent language and chemical nomenclature. Manual screening—reading titles, abstracts, and claims—is time-consuming, while commercial machine learning tools trained on small labeled datasets often fail to generalize, especially in chemically dense domains. The organization screens over 1.8 million patents annually, consuming an estimated 15,000 work hours.

Application We deployed AI Screener to automate relevance classification of patents related to metallocene linear low-density polyethylene (mLLDPE). The model was prompted with domain-specific reasoning trees to assess patent relevance based on implicit chemical relationships, structural variants, and manufacturing processes. Due to confidentiality constraints, the specific reasoning logic and prompt structure are not disclosed. Two test sets were used, each manually labeled by senior chemists. For comparison, we also evaluated a leading commercial classification tool on the same datasets.

Quantitative Results Table 2 summarizes the performance of the AI Screener and the commercial tool against human-labeled ground truth. The AI Screener demonstrated substantial improvements across both evaluation sets. In Set 1, accuracy increased from 86.6% to 92.4%, and recall rose dramatically from 66.7% to 100%, effectively eliminating false negatives—critical for intellectual property risk mitigation. In Set 2, accuracy improved from 42.3% to 71.5%, while recall increased from 89.2% to 98.5%, further reinforcing the AI Screener’s superior ability to identify relevant patents.

A detailed error analysis revealed that the AI Screener correctly identified patents involving synonymous chemical names and structurally related compounds that were missed by the commercial tool. The intellectual property management team validated that the AI Screener not only significantly reduced manual review time while maintaining high confidence in coverage of relevant patents, but also enabled broader intellectual property monitoring and strategic portfolio management—capabilities that were previously constrained by limited resources.

Procurement Contract Analysis for Risk Mitigation

Introduction Large procurement organizations manage tens of thousands of supplier contracts. Manually reviewing these documents to ensure favorable terms, mitigate risk, and prevent value leakage is an intractable problem due to the sheer volume and complexity of the data. This creates significant operational risks and missed financial opportunities.

Application Unlike the previous two use cases, contract analysis involves a significantly more complex data ingestion pipeline, as illustrated in Figure 2. In this case, we implemented a Retrieval-Augmented Generation (RAG) system with several enhancements discussed in the previous section and depicted in Figure 2. We applied AI Screener to automate the extraction and scoring of key commercial terms from procurement contracts. Procurement experts defined a set of critical terms and a corresponding scoring logic to rate them on a scale of 1 (significant improvement needed), 3 (needs improvement), or 5 (best-in-class). The terms included tenure discount, payment terms, order termination, termination notice period, third-party claims, subcontractor clauses, governing law, audit rights, and order assignment.

The scoring logic can be complex and hierarchical. For example, the ”Payment Terms” score was determined by the following logic:

1. Locate and extract clauses related to financial obligations from the document.
2. Determine if a designated override section exists that stipulates primary financial terms; if found, prioritize its contents.
3. In the absence of the override section, examine standard contractual provisions for relevant financial terms, treating them as authoritative.

Table 2: Confusion Matrix for Patent Screening.

	Set 1		Set 2	
	Human: Relevant	**Human: Non-Relevant**	**Human: Relevant**	**Human: Non-Relevant**
AI: Relevant	12 (TP)	34 (FP)	64 (TP)	284 (FP)
AI: Non-Relevant	0 (FN)	403 (TN)	1 (FN)	651 (TN)
Commercial Software: Relevant	8 (TP)	56 (FP)	58 (TP)	570 (FP)
Commercial Software: Non-Relevant	4 (FN)	381 (TN)	7 (FN)	365 (TN)

4. If neither of the above are present, identify alternative settlement mechanisms that may govern payment procedures.
5. Assign a score based on the final, authoritative payment term duration (D):
   - If $D < X$ days, score = 1.
   - If $D \geq Y$ days, score = 5.
   - If $X \leq D < Y$ days, score = 3.

   (Note: X and Y are confidential thresholds set by the procurement organization).

Quantitative Results We compared the AI's scoring against senior expert labels across a large set of contracts. The accuracy for each analyzed term is presented in Table 3.

Table 3: AI Scoring Accuracy for Key Procurement Contract Terms.

Contract Term	Number of Contracts	Accuracy
Tenure Discount	74	97%
Payment Terms	308	92%
Order Termination	35	90%
Termination Notice Period	35	80%
Third Party Claims	35	97%
Subcontractor	35	91%
Governing Law	35	80%
Audit	35	94%
Order Assignment	35	88%

A detailed error analysis reveals that accuracy varies depending on the complexity of the term and the scoring logic. Common sources of error include cases where multiple sections of a contract govern a specific term, and the ingestion/inference pipeline fails to retrieve all relevant information. Additionally, the pipeline may retrieve non-relevant content that confuses the AI Screener; for example, it may conflate purchase order termination with agreement termination, failing to distinguish between the two.

The automated analysis unlocked significant, quantifiable business value. A focused case study on contract payment terms revealed that aligning all agreements to a best-in-class standard could yield a significant boost in cash flow and ongoing financial savings. This underscores the system's ability to convert static contract documents into dynamic, actionable financial insights at scale. Supporting this, a Bank of America study highlights that a strategic approach to payments can unlock hidden cash—optimizing the cash conversion cycle can improve both the income statement and the balance sheet (Bank of America 2025).

Related Work

The application of LLMs to automate and augment document analysis has seen a rapid expansion of research and development (Scherbakov et al. 2024). In their nascent stages, LLMs were demonstrated to have significant capabilities in a wide range of natural language tasks (Minaee et al. 2024). This has naturally led to their application in specialized, high-stakes domains where expert human resources are scarce and the volume of documentation is overwhelming.

Our work builds upon three distinct, yet related, streams of research: the use of LLMs for domain-specific document review, the challenges of applying general-purpose LLMs to specialized corpora, and the development of advanced reasoning frameworks for LLMs.

In the intellectual property domain, LLMs are being explored for tasks such as patent screening. The unique linguistic and structural characteristics of patents, such as their length and use of highly technical and legalistic language, present significant challenges for general-purpose LLMs (Jiang and Goetz 2025). Our work on patent portfolio management contributes to this area by demonstrating how advanced reasoning techniques can overcome some of these challenges.

In the occupational health fields, LLMs are being investigated as tools to accelerate scientific literature reviews. Studies have shown that LLMs can be effective in summarizing and extracting information from medical research, potentially speeding up the development of evidence-based guidelines for occupational health and safety (Shah and Mishra 2024). However, ensuring the accuracy and reliability of LLM-generated summaries remains a key challenge (Bhimani et al. 2025).

The analysis of procurement contracts represents a significant industrial application of LLM technology. The goal is to automate the identification of risks, obligations, and non-standard clauses. While there is significant commercial interest in this area, the academic literature is still emerging. The primary challenges are the need for high precision and the ability to understand complex contractual language, which often requires domain-specific knowledge.

To address the limitations of standard LLM prompting, which often follows a linear, ”chain-of-thought” process, researchers have proposed more sophisticated reasoning frameworks (Wei et al. 2022; Wang et al. 2022) . The ToT paradigm, introduced by Yao et al., allows an LLM to explore multiple reasoning paths in parallel, akin to a human exploring different lines of thought. This approach has been shown to be more effective for complex problems that require planning and strategic lookahead (Long 2023). Our work leverages a ToT-augmented approach to enhance the reliability and depth of automated document review across the use cases we have described.

Lessons Learned

Applying ToT LLMs to automated document review surfaced key insights with broad relevance for enterprise AI adoption. While much attention has been given to chatbots, we found that batch processing of documents—such as compliance checks and report reviews—offers greater immediate value. These workflows benefit from LLMs' ability to perform deep, multi-step analysis at scale, freeing domain experts to focus on higher-level tasks.

Crucially, transparency emerged as a prerequisite for user trust. The ToT architecture naturally exposes the model's reasoning, allowing users to follow its logic and validate outcomes. This not only demystifies the process but also supports effective debugging when errors occur.

Another major challenge was the semantic gap between organizational jargon and the LLM's general training data. Users often assumed shared understanding of internal terms, leading to misinterpretations. We addressed this through targeted prompt training, helping users explicitly define specialized language and context—dramatically improving model accuracy and user confidence.

Finally, we found that our approach is broadly generalizable. Many departments—finance, engineering, research—require structured analysis of domain-specific documents. By decoupling reasoning from content and leveraging prompt-based customization, our framework enables scalable, cross-functional deployment. This modularity turns a one-off solution into a reusable asset, accelerating enterprise-wide AI integration.

Conclusions

This work demonstrates the practical viability and transformative potential of ToT-augmented LLMs for automated document review in high-stakes industrial contexts. By integrating structured reasoning with a 12-billion-parameter LLM, we developed and deployed AI Screener—a scalable, domain-agnostic system capable of emulating expert-level judgment across diverse applications.

Through deployments in occupational health, intellectual property, and procurement, AI Screener has shown that ToT reasoning not only enhances interpretability and decision quality but also enables smaller LLMs to perform complex classification tasks with high accuracy. The system has delivered measurable business value, including significant reductions in manual review time, improved consistency, and actionable insights that directly impact financial and operational outcomes.

Our deployments revealed several key lessons for enterprise AI adoption. First, batch document processing—rather than conversational interfaces—delivers greater immediate value in domains requiring deep, multi-step analysis. Second, transparency is essential: the ToT framework exposes model reasoning, fostering trust and enabling effective error diagnosis. Third, bridging the semantic gap between organizational jargon and general LLM training requires targeted prompt engineering. Finally, the modularity of our approach supports scalable, cross-functional deployment, turning a bespoke solution into a reusable enterprise asset.

Looking ahead, we envision extending this approach to support interactive human-AI collaboration, continuous learning from expert feedback, and integration with structured knowledge bases to further enhance reasoning depth and reliability. This work lays a foundation for the next generation of intelligent document processing systems that are not only automated but also aligned with human expertise and enterprise goals.

References

American Conference of Governmental Industrial Hygienists (ACGIH). 2025. *Threshold Limit Values for Chemical Substances and Physical Agents & Biological Exposure Indices*. American Conference of Governmental Industrial Hygienists, Cincinnati, OH. Documentation for the 2025 TLVs and BEIs.

Bank of America. 2025. Payments as a Working Capital Tool. Accessed: 2025-07-18.

Basu, K. 2024. Bridging knowledge gaps in llms via function calls. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 5556–5557.

Bhimani, M.; Miller, A.; Agnew, J. D.; Ausin, M. S.; Raglow-Defranco, M.; Mangat, H.; Voisard, M.; Taylor, M.; Bierman-Lytle, S.; Parikh, V.; et al. 2025. Real-world evaluation of large language models in healthcare (RWE-LLM): a new realm of AI safety & validation. *medRxiv*, 2025–03.

Dao, T. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Jiang, L.; and Goetz, S. M. 2025. Natural language processing in the patent domain: a survey. *Artificial Intelligence Review*, 58(7): 214.

Long, J. 2023. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*.

Minaee, S.; Mikolov, T.; Nikzad, N.; Chenaghlu, M.; Socher, R.; Amatriain, X.; and Gao, J. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Mistral AI; and NVIDIA. 2024. Mistral-Nemo-Instruct-2407 Large Language Model. https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407. Accessed: July 14, 2025.

Scherbakov, D.; Hubig, N.; Jansari, V.; Bakumenko, A.; and Lenert, L. A. 2024. The emergence of large language models (llm) as a tool in literature reviews: an llm automated systematic review. *arXiv preprint arXiv:2409.04600*.

Shah, I. A.; and Mishra, S. 2024. Artificial intelligence in advancing occupational health and safety: an encapsulation of developments. *Journal of Occupational Health*, 66(1): uiad017.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

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