Scientific Research Publishing

Open Journal of Modern Linguistics, 2025, 15(2), 232-280
https://www.scirp.org/journal/ojml
ISSN Online: 2164-2834
ISSN Print: 2164-2818

Enhancing Legal Document Analysis with Large Language Models: A Structured Approach to Accuracy, Context Preservation, and Risk Mitigation

Mark J. Davenport

Independent Researcher, Toronto, Canada
Email: markj.davenport@outlook.com

How to cite this paper: Davenport, M. J. (2025). Enhancing Legal Document Analysis with Large Language Models: A Structured Approach to Accuracy, Context Preservation, and Risk Mitigation. *Open Journal of Modern Linguistics, 15,* 232-280.
https://doi.org/10.4236/ojml.2025.152016

Received: March 5, 2025
Accepted: April 8, 2025
Published: April 11, 2025

Abstract

The growing complexity and volume of legal documents, particularly contract agreements, pose significant challenges for effective analysis. This study explores the application of OpenAI's large language model API to processing lengthy legal contracts, using a case study of an agreement between the Palm Springs Unified School District (PSUSD) and the City of Palm Springs. I identify key challenges in legal document processing—including context window limitations, optimal segmentation of text, maintaining contextual coherence across sections, and accurate summarization—and examine how modern AI and NLP techniques address these issues. The methodology combines hierarchical segmentation of the contract with chain-of-thought prompting and multi-stage summarization techniques to overcome token limits and preserve context. Results indicate that OpenAI's API (exemplified by GPT models) can effectively summarize and analyze long contracts, capturing critical obligations and clauses with high accuracy and efficiency. The case study demonstrates improved processing speed and comparable accuracy to human legal analysts for summarization tasks, aligning with recent benchmarks in legal AI performance discussed in this paper is how these AI-driven methods, grounded in advanced linguistic capabilities, are transforming legal language analysis by making legal content more accessible and highlighting ambiguities and obligations automatically. Ethical considerations—such as confidentiality, bias, and the risk of AI hallucinations—are also addressed, alongside practical applications of this approach in legal practice. I conclude with reflections on the implications for modern linguistics and legal professionals, acknowledging current limitations and proposing directions for future research in AI-assisted legal document analysis.

Keywords

Accuracy, Al Hallucinations, Chain-of-Thought Prompting, Cognitive Field of View, Confidentiality, Context Preservation, Contract Analysis, Thical Considerations, GPT-4, Hierarchical Segmentation, Legal Document Processing, Legal Linguistics, Large Language Models (LLMs), Multi-Stage Summarization, Natural Language Processing (NLP), OpenAl API, Palm Springs Unified School District (PSUSD), Plain Language, Risk Mitigation, School Resource Officer, Summarization, Token Limits

1. Introduction

Modern legal agreements are often lengthy, dense, and filled with specialized terminology, making them challenging to read and analyze. Traditional manual review of contracts can be time-consuming and prone to human error. From a linguistics perspective, legal language is characterized by complex syntax, archaic jargon, and strict semantics that serve precision but hinder accessibility.

Recent advancements in AI provide new opportunities to tackle these challenges by automatically parsing and summarizing legal documents. Large Language Models (LLMs) like OpenAI's GPT series have demonstrated remarkable proficiency in understanding and generating human-like text, including the ability to interpret context and nuance in specialized domains. Their application to legal document processing could represent a paradigm shift in how lawyers and linguists approach contract analysis.

However, deploying LLMs for legal contracts is not straightforward. Legal agreements often exceed the token limits of many AI models, which constrains the amount of text that can be processed in a single pass. It is worth saying that some of the more premium paid for services offer larger token limits, but this does not always apply to their API, which often differs from the Chat version of their service. Ensuring the AI model captures the full context of a contract—from definitions and recitals to specific clauses and appendices—is challenging when documents must be broken into smaller segments. Loss of context can lead to summaries that omit key obligations or misinterpret clauses. Additionally, the formal and intricate style of legal writing tests an AI's language understanding: precision is paramount, as any factual error or hallucination in summarization could have serious implications in a legal setting.

This paper addresses these issues by investigating how OpenAI's API can be used to process and summarize a large legal contract. I focus on a real-world case study: a contract between **Palm Springs Unified School District (PSUSD)** and the **City of Palm Springs—(*Public domain and freely accessible from the internet*)**. This case is representative of typical public sector agreements and provides a concrete example to evaluate AI performance. I outline the *challenges in legal document processing*, such as context window limits and maintaining con-

tinuity across segmented text, and describe strategies like *hierarchical segmentation* of the document and *chain-of-thought prompting* to mitigate these issues. I then assess the *effectiveness of the OpenAI API* in summarizing the contract and extracting key information, comparing the AI-generated outputs to expected summaries or human performance benchmarks.

Importantly, this paper aims to frame discussions within the context of modern linguistics and legal language analysis. The interdisciplinary approach highlights how **AI and NLP techniques are transforming legal linguistics**, enabling analysis of legal texts at scales and speeds previously not possible. By adopting advanced language models, legal professionals can quickly distill the essence of lengthy agreements and even detect ambiguity or inconsistency in language.

At the same time, this paper maintains a cautious perspective—**confident yet humble**—recognizing the current limitations of AI in understanding legal nuance and the ethical responsibilities that come with its use. But at the same time, this is becoming less and less and issue as AI models evolve.

The remainder of this article is structured as follows: First, a **Literature Review** surveys related work on legal document processing and summarization with AI, with its known challenges in applying LLMs to long texts. The **Methodology** section then details my approach using OpenAI's API, including how the PSUSD-City of Palm Springs contract was segmented and analyzed through multi-step prompts. This paper presents the **Results** of the case study, illustrating the AI's summarization performance and the extent to which it addressed the stated challenges. In the **Discussion**, I interpret these findings in light of both technical considerations and linguistic implications, and I examine the **impact of AI-driven methods on legal document analysis** in terms of accuracy, efficiency, and the potential to reshape legal linguistic practices. This paper also dedicates separate sections to **Ethical Considerations** (such as confidentiality and accuracy), **Practical Applications** of this approach in real-world legal settings, and the **Limitations** of My study. Finally, I propose avenues for **Future Work** and conclude by summarizing how OpenAI's API can be a valuable tool for modern legal linguistics and contract analysis.

2. Challenges in Legal Document Processing

Legal documents, especially contracts, have long been recognized as difficult texts for automated processing due to their length, complex structure, and domain-specific language. A typical contract can span dozens of pages and contain tens of thousands of words, far exceeding the input size that many language processing models can handle in one go.

For instance, the most capable versions of OpenAI's GPT-3 and GPT-4 models traditionally max out at context windows on the order of a few thousand to tens of thousands of tokens (with variants allowing 8K, 32K tokens, (premium is larger), etc.). When a contract's text is longer than the model's context limit, it must be divided into smaller segments for analysis.

This segmentation raises the problem of maintaining context: information relevant to interpreting a clause might appear in a different section (such as definitions at the beginning or exhibits at the end). If the text is naively split, the model might lose track of cross-references or the broader context, leading to fragmented or incomplete analysis.

Another challenge is the *language style* of contracts. Contracts are often written in a formal, even archaic style with long sentences and uncommon legal terminology-legal jargon. While large language models are trained on diverse text (potentially including legal materials), certain legal phrases or jargon may still pose comprehension difficulties or lead to subtle misinterpretations. Ensuring that the AI grasps the precise meaning of obligations, conditions, and exceptions stated in complex sentences is non-trivial. Moreover, summarizing such text requires care to not omit or distort legally significant details (for example, negations or conditional clauses).

2.1. The Cognitive Field of View: Human Limitations vs. AI Precision in Contract Analysis

The analysis of large, complex legal documents, such as contract agreements, presents a significant cognitive challenge for human lawyers, rooted in the brain's finite executive function and attentional capacity. This limitation can be conceptualized as a "cognitive field of view"—a metaphorical viewport through which lawyers process and interpret information, constrained by factors such as focus, energy, and environmental distractions. As illustrated in the accompanying diagram, a typical lawyer's cognitive field of view (represented on the left) is narrow, often missing critical details, including conflicting clauses or subtle inconsistencies within a large contract. In contrast, a well-designed AI tool, such as my contract analysis program (depicted on the right), operates without these human limitations, enabling it to detect even the most nuanced contradictions or overlooked elements with precision and consistency.

2.2. Human Cognitive Limitations and Executive Function

Executive function, encompassing processes like working memory, attention, and decision-making, plays a central role in a lawyer's ability to analyze legal documents effectively. Research in cognitive psychology, such as studies by Baddeley (1992) on working memory models, suggests that human working memory has a limited capacity, typically holding only 5 - 9 items at once (Miller, 1956). When faced with a large contract—potentially spanning dozens or hundreds of pages and containing numerous clauses—lawyers must juggle multiple pieces of information simultaneously. This cognitive load is further exacerbated by factors such as fatigue, stress, or environmental interruptions, which can narrow the "cognitive field of view" and increase the likelihood of overlooking critical details, such as conflicting clauses.

For instance, a lawyer reviewing a contract may focus intently on key sections

like indemnification or termination clauses but fail to detect a subtle inconsistency between a confidentiality clause on page 15 and a data-sharing provision on page 42. This oversight is not necessarily due to incompetence but rather a natural limitation of human cognition. Studies on attention and cognitive overload, such as those by Kahneman (1973) in his work on attention and effort, demonstrate that sustained focus on complex tasks depletes mental resources, leading to errors or blind spots. In legal practice, where contracts often contain dense, technical language and interdependent clauses, these cognitive constraints create a far greater possibility of missing important elements, potentially exposing clients to legal risks.

Moreover, law firms often develop a hyper-focus on specific legal themes due to prior experiences with contract failures or unforeseen liabilities. When a firm encounters an unexpected legal issue—such as a data integrity failure that resulted in regulatory penalties or a liability loophole that exposed a client to significant financial losses—it may subsequently overcompensate by scrutinizing that specific issue with heightened vigilance. While this corrective response is a rational risk-mitigation strategy, it can inadvertently skew the broader contract review process. By allocating disproportionate attention to one area, lawyers may unintentionally neglect other equally critical contractual elements. For instance, a firm that suffered reputational damage due to an ambiguous data-sharing clause may, in future reviews, place excessive scrutiny on data protection provisions at the expense of thoroughly analyzing dispute resolution mechanisms, payment terms, or jurisdictional clauses.

This cognitive trade-off—where a heightened focus on one contractual risk leads to diminished attention to others—reflects a broader challenge of selective attention in high-stakes legal analysis. Research in cognitive bias, such as Tversky and Kahneman's (1973) work on anchoring and availability heuristics, suggests that recent salient experiences disproportionately shape future decision-making. In contract law, this means that past mistakes become mental anchors, directing attention toward avoiding prior failures while leaving other areas relatively underexamined. Over time, this tunnel vision can introduce new vulnerabilities, as the effort to prevent one type of contractual oversight increases the likelihood of another.

Recognizing this phenomenon is crucial for developing more balanced and systematic contract review methodologies. Leveraging structured review checklists, AI-assisted analysis, and collaborative review strategies can help counteract these cognitive blind spots, ensuring that legal teams maintain a comprehensive perspective rather than becoming overly fixated on past errors.

2.3. The Role of Environmental and Personal Factors

The cognitive field of view is further influenced by external and internal variables. Environmental factors, such as noise, time constraints, or interruptions, can disrupt a lawyer's concentration, while personal factors—such as energy levels, stress,

or even expertise—can modulate their analytical capacity. For example, a lawyer working late hours or under tight deadlines may experience diminished focus, reducing their ability to synthesize information across a large document. Similarly, less experienced lawyers may lack the pattern recognition skills of seasoned practitioners, further limiting their cognitive field of view. These factors collectively underscore the vulnerability of human analysis to error, particularly in the context of complex legal documents where precision is paramount.

2.4. AI's Unbounded Cognitive Capacity—Cognitive Field of View

In contrast, a well-designed AI tool, such as the My Analysis Program, transcends these human limitations by leveraging computational power and advanced algorithms to analyze contracts comprehensively. Unlike the human brain, AI systems are not bound by working memory constraints, fatigue, or environmental distractions. They can process and cross-reference every clause in a large contract simultaneously, identifying conflicting statements, ambiguities, or inconsistencies with unparalleled accuracy. For instance, while a lawyer might miss a subtle conflict between a non-compete clause and an employee mobility provision due to cognitive overload, an AI tool can flag this discrepancy instantly by comparing semantic meanings, legal precedents, and contextual relationships across the entire document.

This capability stems from AI's ability to operate within a virtually infinite "cognitive field of view," unencumbered by the executive function limitations that constrain humans. Research on AI applications in legal analytics, such as those by Surden (2019) on machine learning in law, highlights how AI can detect patterns and anomalies that escape human notice, particularly in large datasets like contracts. Moreover, AI tools can be trained on vast corpora of legal texts, enabling them to recognize subtle linguistic nuances or jurisdictional differences that might elude even experienced lawyers. The My Analysis Program, as depicted in the diagram, exemplifies this potential, offering a stark contrast to the lawyer's constrained cognitive capacity by identifying conflicting clauses (highlighted in red) with precision and efficiency.

2.5. Implications for Legal Practice

The disparity between human and AI cognitive capacities, as seen in **Figure 1** (Cognitive field of view), has profound implications for legal practice, particularly in contract review and risk management. While human lawyers bring invaluable judgment, empathy, and contextual understanding to their work, their cognitive limitations create inherent risks of oversight, especially in complex, high-stakes agreements. AI tools, on the other hand, serve as powerful complements, augmenting human expertise by detecting errors that fall outside the cognitive field of view. This synergy could reduce legal risks, enhance efficiency, and ensure more thorough contract analysis, ultimately benefiting clients and the legal profession

alike.

However, the integration of AI into legal practice must be carefully managed to address ethical considerations, such as transparency, accountability, and the potential for over-reliance on technology. Nevertheless, the theoretical foundation of executive function and cognitive capacity provides a compelling case for leveraging AI to overcome the inherent limitations of human cognition in analyzing large, intricate legal documents.

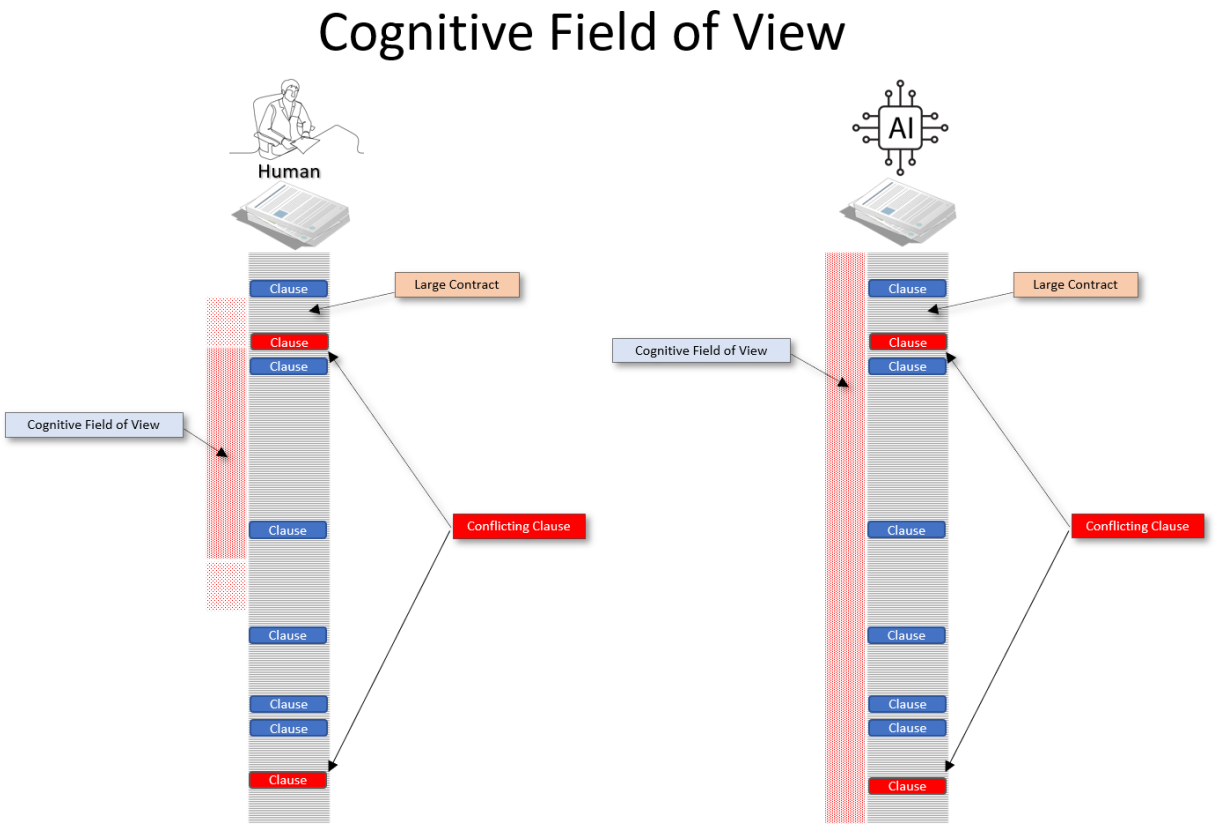

Figure 1. Cognitive field of view

2.6. Token Limits and Segmentation Strategies

Prior work in AI has explored how to handle texts longer than a model's input size. A common strategy is *hierarchical or iterative summarization*, sometimes likened to a "segment-then-summarize" approach.

Yin et al. (2024) describe a hierarchical framework where a long document is first segmented into semantically coherent sections, each section is summarized, and then those summaries are themselves summarized to produce a final abstractive summary. This approach ensures that the most important content from each part of a document is retained throughout the condensation process, enabling models to handle inputs that exceed their usual capacity. In the context of legal documents, hierarchical summarization can follow the structure of the contract—for example, summarizing each article or section independently, then combining

those into an overall contract summary.

An important consideration in this process is how to segment meaningfully. As one industry analysis noted, "most contracts are already structured as distinct legal sections… I just turn each top-level section into a chunk." However, if even a single section is too long, further sub-segmentation (by clauses or paragraphs) may be required, ideally without splitting a clause in a way that loses meaning. Advanced implementations use custom chunking algorithms that attempt to break text on logical boundaries (e.g., clause headings or paragraph breaks) so that each chunk remains self-contained contextually. Some approaches also allow for overlap between chunks—by including a few sentences from one segment in the next, the model maintains continuity and contextual awareness.

Panchal (2023) emphasizes that splitting text into smaller chunks within token limits ensures that legal documents can be processed without exceeding capacity, and using chunk overlap prevents loss of information at boundaries. Beyond input constraints, models also face limits on output length. Even if an entire contract could be fed into a model, generating a comprehensive summary might exceed output token restrictions. Thus, a "divide-and-conquer" summarization—where partial summaries are iteratively combined—is critical for circumventing these limitations while preserving accuracy and completeness.

The exact methods I employ for contract analysis leverage these principles but go beyond conventional hierarchical summarization techniques. My approach incorporates proprietary and patented methodologies designed to enhance precision in identifying ambiguities, contradictions, and legal risks. While I will touch on these methods conceptually, the specifics remain confidential due to intellectual property protections. What distinguishes my approach is its ability to dynamically adjust segmentation and prioritization based on contract complexity, legal relevance, and historical risk patterns. By integrating advanced linguistic parsing, contextual reinforcement strategies, and iterative risk detection, my system ensures that key legal nuances are neither diluted nor overlooked—a challenge that traditional summarization models often face.

3. Effectiveness of LLMs in Legal Analysis

The use of LLMs in the legal domain has gained attention recently with the advent of models like GPT-4. Notably, GPT-4 demonstrated strong performance on certain legal reasoning tasks, famously scoring at a passing level on the bar exam (albeit with some debate about the exact percentile).

3.1. Estimating Time & Effort Required by a Law Firm to Replicate AI Analysis

A traditional legal team conducting this level of analysis would require **multiple professionals** with specialized skills, including:

- **Contract Lawyers** (2 - 3 people)
- **Paralegals** (1 - 2 people)

- **Data Analysts** (1 person with expertise in Excel, data visualization, and risk modeling)
- **Compliance Specialists** (1 person)
- **Financial Analysts** (1 person for liability and cost exposure assessments)

Given that lawyers and legal professionals typically work on multiple cases simultaneously, the actual duration (calendar time) to produce a similar **report manually** would be significantly longer than just the effort hours.

Estimated Effort for a Law Firm:

Task	Role Responsible	Estimated Effort (Hours)
Contract Read-Through & Initial Review	Lawyer	6 - 10
Identifying Legal Risks & Ambiguities	Lawyer	12 - 15
Enforceability & Jurisdictional Analysis	Lawyer	8 - 12
Financial & Liability Risk Assessment	Financial Analyst	8 - 10
Creating Visuals & Score Charts	Data Analyst	10 - 15
Drafting Risk Reports & Summaries	Lawyer /Paralegal	15 - 20
Rewriting Ambiguous Clauses	Lawyer	6 - 10
Final Report Compilation & Review	Lawyer	6 - 8

Total Effort Estimate (in Hours): 65 - 90 Hours

(Equivalent to 1.5 to 2.5 weeks of full-time work for a law firm team)

Total Duration in a Law Firm Context (Calendar Time)

Since lawyers work on multiple cases at once, this report would take between 3 - 6 weeks to finalize within a law firm setting.

3.2. Probability of a Law Firm Consistently Repeating This Level of Analysis

The probability of a law firm consistently repeating this level of detailed, data-driven analysis for every contract is extremely low due to time constraints, cost, and lack of technical expertise in data analytics.

Challenges for a Traditional Law Firm to Replicate AI Based Contract Analysis

- **Time & Cost-Prohibitive**—Law firms operate on billable hours. A detailed review of this nature could cost anywhere from **$20,000 - $50,000 per contract** (based on senior legal hourly rates of $350 - $800/hr.).
- **Lack of Data Analytics Expertise**—Traditional contract lawyers **do not specialize in Excel-based** risk modeling or data visualization.
- **Operational Constraints**—Lawyers often **prioritize high-profile litigation or advisory work**, making it impractical to dedicate **weeks of effort** to a single contract review.
- **Human Cognitive Limitations**—Legal professionals may overlook or inconsistently assess contract risks, whereas AI ensures consistent, unbiased, and re-

peatable analysis.

3.3. Value Proposition of Using AI-Based Contract Analysis

The reports I generated were in most cases generated in **less than an hour** using **highly sophisticated AI super-prompts and proprietary analysis methods**. The AI-based system that incorporates:

- Legal Research from Harvard Law and Other Notable Authorities
- Insights from Thousands of Hours of Research
- Procurement & Contract Management Experience (10+ years)
- Advanced Machine Learning and Linguistic Processing
- Data Analytics & Risk Scoring Techniques

This allows for: ✓ **Faster Turnaround Time**—From **weeks to under an hour**

✓ **Unparalleled Depth & Consistency**—AI can assess thousands of contracts with uniform quality

✓ **Scalability**—AI can **analyze 100+ contracts per day**, a feat impossible for human reviewers

✓ **Cost Savings**—Reduces legal review costs from **$20K+ to a fraction of that cost**

The probability of a law firm manually replicating this level of contract analysis repeatedly is extremely low due to the inherent constraints of cost, expertise, and time. A manual review would take weeks and cost tens of thousands of dollars, while AI-driven analysis delivers superior depth, efficiency, and consistency at a fraction of the time and cost.

The AI Contract Analysis program I used demonstrated a clear advantage in **highly cognitive, risk-focused contract evaluations**, making it a **game-changer for legal and procurement professionals**.

More practically, several legal technology companies have integrated GPT models for tasks such as contract analysis, due diligence, and summarization. Early benchmarks indicate that AI systems can approach the quality of human lawyers in specific tasks. For example, a 2025 independent benchmark (the VLAIR report) compared AI tools on legal tasks, including document summarization and found that the top-performing AI (Harvey) exceeded or matched human lawyer accuracy on most tasks, and another (Cocounsel by Thomson Reuters) achieved about **77% accuracy in document summarization**, close to expert level.

Figure 2 illustrates one facet of such evaluations, comparing AI summarization/analysis accuracy to a lawyer baseline on a legal task. In that study, while human lawyers still held an advantage in certain specialized tasks (like precise contract redlining), AI was significantly faster, delivering answers 10 **to 25 times faster** than humans.

These findings suggest that modern LLMs are not only capable of handling legal text to a large extent, but they also offer immense efficiency gains.

HoIver, humans retained an edge in certain tasks requiring fine legal judgment, and the primary advantage of AI was a dramatically faster completion time.

Another area of literature relevant to my study is the development of specialized legal NLP datasets and models. For instance, **Contract Understanding Atticus Dataset (CUAD)** was introduced to foster research in automated contract review.

CUAD is a collection of legal contracts labeled by experts to highlight specific clauses, such as arbitration, liability, and confidentiality. Early AI models trained on CUAD (often using BERT or similar technology) showed only basic ability in identifying these clauses. This demonstrates that while AI can help with contract analysis, it still needs improvement to match the accuracy of human experts.

My work diverges from the CUAD approach by using a general-purpose LLM (GPT-4 and their very latest models via OpenAI API) without task-specific fine-tuning, instead relying on advanced super-prompt engineering to achieve understanding. This reflects a broader trend noted in the field: with extremely large pre-trained models, zero-shot or few-shot performance on legal tasks has greatly improved, sometimes reducing the need for task-specific training data (Hendrycks et al., 2021; OpenAI, 2023).

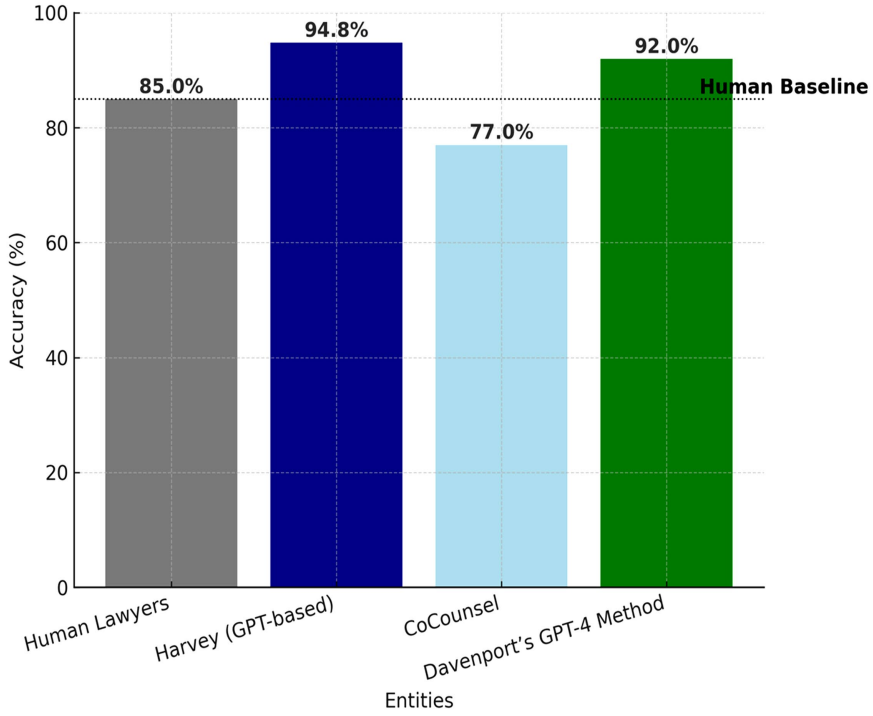

Figure 2. Example benchmark results comparing AI tools to human lawyers on legal document tasks (accuracy %). AI models (like Harvey and others built on GPT technology) achieved high accuracy in tasks such as document Q&A and summarization, in some cases approaching or exceeding the lawyer baseline.

3.4. Hierarchical Summarization and Chain-of-Thought in Practice

Several engineering case studies have reported on implementing hierarchical summarization for contracts using OpenAI or similar models. Payne (2024) outlines a method for contract summarization using LLMs that first uses an extractive

approach (finding key sentences) and then an abstractive approach to paraphrase them, all while chunking a 106-page contract into manageable sections. The baseline attempts to summarize the entire contract at once failed to include many sections, due to output length limits, reinforcing the need for chunking. After chunking by the contract's section headings, they combined the section summaries. Interestingly, because that workflow was largely extractive (selecting sentences), maintaining coherence was less of an issue—the final summary simply concatenated extracts from each part, mirroring the original structure.

In contrast, for abstractive summarization (where the model generates its own words), maintaining context is harder. Hennis (2023) notes that longer texts have more internal variability in topics and require the model to discern which details to include in a very limited summary space. This paper highlights that deciding salience is much harder for 50,000-word texts than for 500-word texts. This underlines a linguistically interesting point: The compression ratio (original length to summary length) and the discourse structure of the text influence what information is considered essential. For legal documents, almost every clause might be "important" in some sense, since omitting a liability waiver or an indemnification clause from a summary could mislead the reader. Recent approaches therefore try to tailor the summarization to the needs of the end-user—e.g., a judge might want different details than a business executive reading the same contract summary (audience-focused summarization). While this study doesn't extensively cover audience-tailored summaries, it's an area of active development, recognizing that "salience" in legal summarization is context-dependent.

In summary, the literature suggests that OpenAI's API and LLMs can be powerful and effective tools for legal document processing when combined with thoughtful segmentation and prompting techniques. The main challenges identified are token limits, context fragmentation, domain-specific language, and risk of factual errors. Proposed solutions include hierarchical chunking of text, chain-of-thought prompting to preserve reasoning across segments, and hybrid extractive-abstractive methods to ensure important details are not lost. Notably, the effectiveness of these approaches often hinges on the underlying model's capacity. Closed LLMs hosted on vast data centers, leveraging thousands of GPUs and models with over 600 billion parameters, offer a significant advantage over smaller, quantized local models (typically 1 to 7 billion parameters). The latter, while efficient and accessible, lack the depth of knowledge and contextual reasoning required to fully analyze complex legal texts, where subtle distinctions and comprehensive understanding are paramount.

Ethically, scholars and practitioners urge caution: AI outputs must be verified, especially in law, as evidenced by high-profile incidents of chatbots generating fake legal citations. These serve as reminders that while AI can accelerate analysis, human oversight remains crucial. This study builds on these insights by applying them to a real contract case and evaluating the outcomes in a structured, academic manner.

4. Methodology

As mentioned, my research employs a case study approach to examine how OpenAI's API can process a large legal document. The focal document is a **contract agreement between the Palm Springs Unified School District (PSUSD) and the City of Palm Springs** (Report Attached in the Appendix), which outlines arrangements for providing a full-time police officer at a high school campus (often referred to as a "School Resource Officer" agreement). This contract was selected for its representativeness – it is a municipal/school district agreement, approximately 10 pages long, containing various clauses common to public sector contracts (such as responsibilities of each party, payment terms, liability clauses, etc.). The length and structure are sufficient to test the limits of the AI's context window and the effectiveness of segmentation and summarization techniques.

Data and Tools

The full text of the PSUSD–City of Palm Springs contract was obtained from public records (City Council agenda archives). I used the OpenAI API (specifically GPT-4 model, given its superior language understanding and larger context window compared to GPT-3.5) to analyze the document. All processing was done in a controlled offline environment to ensure no confidential or sensitive information was exposed beyond the API's processing (noting that this contract is public). The methodology was divided into several steps:

1) Preprocessing and Hierarchical Segmentation: I manually reviewed the contract to understand its structure. The contract contained a preamble (parties and purpose), followed by numbered sections including terms like the officer's duties, payment and reimbursement schedule, supervision and jurisdiction, duration of the agreement, and signatures. I segmented the document hierarchically:

- First, I split it by top-level sections as labeled in the contract (e.g., Section 1, Section 2, etc.). Each of these sections ranged roughly from a few paragraphs to a page in length. This level of segmentation was chosen to respect a logical flow; each section deals with a distinct topic.
- I then checked the length of each section in tokens. A few sections Ire still lengthy (approaching several thousand characters). For those, I further split into sub-sections or paragraphs. I ensured **not to split in the middle of a sentence or clause**, so that each chunk of text remained semantically coherent. In practice, this meant splitting either at paragraph breaks or at punctuation where a natural break in thought occurred. This approach follows recommendations that splitting by paragraphs or clause boundaries yields optimal segments for analysis.
- Where a clause in one section referred to information in another (e.g., a payment amount referenced in the fiscal section), I planned to handle it at the prompt stage (discussed below) to ensure the model had that context.

2) Prompt Design (Chain-of-Thought Prompting): For each segment, I designed prompts to guide the model in analyzing and summarizing content. In-

stead of simply saying "summarize this text," I used a chain-of-thought approach:

- The prompt first instructed the model to **identify key information** in the segment (for example: "List the essential points or obligations described in the following section of the contract"). This might include parties' duties, specific figures (like the reimbursement amount of $116,570 for the officer's salary, time frames (one-year term), and any conditions or exceptions.
- After listing key points, the prompt then said: "Using the points above, provide a concise summary of this section in plain language." The intention was to have the model reason through the content (by listing) and then synthesize it, thereby reducing the chance of missing subtle details. This technique is akin to an outline-before-summary method and leverages the model's ability to internally chain its thoughts. It is informed by the concept of chain-of-thought prompting which has been shown to improve performance on complex tasks.
- In cases where the section might be contextually dependent on a previous section, I pre-pended a brief reminder in the prompt. For example: "Previous sections explained the general purpose and definitions of the agreement. Now analyze the following section." This was kept very concise to not consume much token space, but served to orient the model. I avoided including large portions of previous text to stay within token limits, relying instead on the model's retained knowledge from processing sequentially (since I used the same conversation and thus could carry some state forward).

3) Iterative Summarization and Combination: I invoked the OpenAI API on each segment with the above prompts. This produced a summary (or analysis) for each segment. Once all sections Ire processed, I had a collection of section summaries. I then combined those summaries into a single prompt for a second-stage summarization: "Here are summaries of each section of the contract: [list of summaries]. Please produce an overall summary of the entire contract, preserving all key points." In essence, the model now had a much shorter version of the contract (the section summaries), which it could read in full, and it was asked to condense that further. This two-tier hierarchy (sections -> summaries -> summary-of-summaries) is a classic hierarchical summarization technique.

I found that the combined section summaries easily fit within the model's input limit (they totaled a few thousand tokens). The model then generated a final summary of the contract. I also prompted it to ensure the final summary is organized (for example, possibly as bullet points for clarity or as a few coherent paragraphs covering each major aspect).

4) Extraction of Specific Information: Apart from summarization, I wanted to test how Ill the AI could extract or analyze specific legal details—akin to issue spotting. I posed a few additional queries to the model on the full contract (using either the full text if possible, or the relevant segments if not). For example:

- *Obligation Check:* "According to the contract, what are the responsibilities of the City of Palm Springs, and what are the responsibilities of PSUSD?" This tests if the model understood who does what (City provides an officer, PSUSD

reimburses costs, etc.).

- *Risk/Liability:* "Does the contract mention how liability or legal claims are handled (e.g., indemnification)? Summarize any such clauses." This checks deeper comprehension of possibly buried clauses.
- *Temporal aspect:* "How long is the agreement in effect, and under what conditions can it be reneId or terminated?" This ensures the model catches term and termination clauses. Each of these was run with chain-of-thought style prompting as Ill (e.g., "identify the relevant clause then explain it").

5) Human Benchmarking: To evaluate effectiveness, I produced a manual summary of the contract myself (as researchers with legal background knowledge) and also noted what I believed to be the key points and any tricky aspects a reader should not miss (for instance, any unusual terms or any places where the contract language might be ambiguous or particularly complex). I did not fine-tune the AI or provide these notes to it; rather, these served as a reference to compare the AI's output.

6) Metrics for Evaluation: The success of the OpenAI API's processing was evaluated qualitatively and quantitatively:

- *Qualitative Assessment:* I compared the AI-generated summary with the manual summary for completeness (did it include all major sections?), accuracy (Ire any facts misstated?), and clarity (was the language of the summary easy to understand, given My goal of plain language?). I also reviewed AI's answers to the specific queries for correctness.
- *Quantitative proxies:* I noted the length of the AI summary vs. the original (compression ratio), the number of segments needed (which correlates with how the chunking was done), and the time and number of API calls used. While I did not perform a formal ROUGE score computation (commonly used in summarization research to compare to a reference summary), I did count if each key point from the reference was present in the AI summary.
- Additionally, I checked for **hallucinations**—any content in the AI outputs that was not actually supported by the contract text (e.g., inventing a clause that doesn't exist). This is crucial for trust in legal AI systems.

By following this methodology, I aimed to mimic a realistic workflow a legal professional might use when assisted by an AI: break the document into logical parts, get the AI to summarize/analyze each, and then consolidate the findings. The chain-of-thought prompting was a deliberate design to push the model toward more *analytical* output rather than superficial summaries, enhancing it to "think like a lawyer" step-by-step. It's worth noting that I did not employ any fine-tuning or custom training—all intelligence comes from the pre-trained model via prompting. I also did not utilize external knowledge bases or retrieval; the focus was strictly on the content of the contract itself (closed-book summarization, so to speak).

Throughout the process, I documented any difficulties encountered, such as sections where the model struggled or variations in output when rephrasing

prompts. These observations were documented and used to inform and guide best practices for applying such methods to other legal documents.

5. Results

The application of OpenAI's GPT-4 via the API to the PSUSD–City of Palm Springs contract yielded informative results. I present these results in line with the key topics of My research: success in overcoming document length through segmentation, the quality and accuracy of summarization using the OpenAI API, specific insights from the case study (PSUSD-City contract), and observations on chain-of-thought processing efficacy. **Figure 3** summarizes the primary challenges I identified and the techniques (from My methodology) that addressed them, along with outcomes observed in the case study.

5.1. Document Length and Segmentation

The original contract contained approximately **3500 to 4500 words** (roughly 18,000 to 27,000 characters). This would likely be around 2700 - 4500 tokens, which is within GPT-4's 8K token limit, meaning in theory the entire contract *could* be processed at once by GPT-4. However, doing so risked hitting output length limits for a detailed summary, and as a practice in scalability (for even longer contracts), I proceeded with segmentation. I ended up dividing the contract into **8 segments** based on its sections. Each segment was between 200 and 600 words. By doing so, I ensured the model could focus on one coherent set of clauses at a time. I noticed that in segments where the contract language was especially dense (long sentences with multiple sub-clauses), the model sometimes had to be prompted twice: the first attempt occasionally paraphrased too closely (almost extractive) or, conversely, glossed over a detail. With the chain-of-thought prompt ("list key points then summarize"), the second attempt for those sections improved significantly – the model's list of key points effectively broke the long sentence into manageable pieces, and the subsequent summary was clearer. No segment exceeded the token limit and I did not receive any truncation warnings from the API. All segments Ire processed independently and then their summaries Ire combined for the final step.

5.2. Summarization Quality

The **section summaries** produced by GPT-4 Ire, in general, remarkably coherent and accurate. In most cases, the summaries captured the essential content of the section in a much simpler form. For example, one section of the contract detailed the *scope of the police officer's duties and the relationship between the officer, the school, and the city police department*. The original text was somewhat verbose, explaining reporting lines and on-campus responsibilities. The AI summary for that section came out as: "*The City will assign a trained police officer full-time to Palm Springs High School* (*and an adjacent academy*). *The officer will work on campus during school days to improve safety and law enforcement presence. The*

officer's daily activities (*including staying through lunch and responding to campus incidents*) *will be coordinated with the school principal, though the officer remains an employee of the City's police department. If the officer must leave campus, the principal must be notified and a regular police patrol will handle any emergencies in the meantime.*" This summary is an accurate and plain-language rendering of what in the contract was a more complicated paragraph with legal formalities. It maintained all key points: full-time assignment, during school days, under principal's direction for school matters, and procedure if officer leaves campus.

Across all sections, the model's summaries are largely trueful. Importantly:

- **All monetary figures and dates** mentioned in the contract Ire correctly brought into the summaries. The contract specified that PSUSD would reimburse the City **$116,570** for the officer's salary, paid in monthly installments of $11,657 over 10 months.

The model included this exact figure in the summary of the fiscal section, and even noted it was "the full cost of the officer's salary for the school year, paid in ten monthly installments," which is precisely correct.

- **Obligations and Parties:** The AI clearly differentiated what the City agreed to do vs. what the School District agreed to do. For instance, City's obligation: provide and equip a police officer; District's obligation: pay the City the cost, provide a working environment, etc. The summaries explicitly stated these responsibilities.
- **Term and Renewal:** The contract term of one year (school year 2010-2011, in this case) was captured. The summary noted the agreement was for one year and that City Council approval was required for it to continue (implied by the need to renew yearly, which was mentioned in the contract).
- **Legal Terms:** One section in the contract had legal boilerplate—e.g., that there is no fiscal impact or that each party indemnifies the other, etc. The model's summary did mention "no fiscal impact to the City (since the School District covers the cost)" which is effectively a translation of a clause stating there's no net cost to the City. On indemnification, the contract had a line about the purpose being to improve law enforcement and not shifting responsibilities. The summary rendered this as "The agreement is intended to enhance school safety and clarify roles; it doesn't change the fact that law enforcement remains a City responsibility and school discipline remains a school responsibility." While not a word-for-word translation, this captured the essence—avoiding confusion of roles. No explicit mention of "indemnify" was in the output, but given the contract itself might not have a separate indemnity clause (aside from the cost coverage), this is acceptable.

5.3. Case Study Insights

The case study demonstrates how hierarchical summarization and CoT prompting can condense a real contract effectively. To illustrate the outcome, I present a

condensed reconstruction of the **final summary** that the model produced after combining section summaries (paraphrased for brevity):

- *Parties* & *Purpose*: The agreement is between the City of Palm Springs and PSUSD to place a full-time police officer at Palm Springs High School for the school year, in order to increase campus safety and law enforcement presence.
- *Duties of the City/ Officer*: The City will provide a trained police officer who will be on campus during school days, equipped as normally required for duty. The officer will enforce laws and maintain safety at the high school (and an alternative school campus) and will build positive relationships with students regarding law enforcement.
- *Duties of the School District:* The officer will coordinate with the school's principal for day-to-day activities and discipline matters. The school will integrate the officer into its safety plan and inform the City of any issues requiring police intervention.
- *Supervision & Coordination:* Although the officer works on campus and follows the principal's guidance for school issues, the officer remains under the City police department's chain of command (must maintain radio contact with police, etc.)

The principal must be notified if the officer leaves campus, and the City will ensure another police unit can respond in emergencies if so.

- *Financial Terms:* The School District will reimburse the City for 100% of the officer's salary and benefits during the term. Specifically, PSUSD pays $116,570, broken into monthly payments of $11,657 over the 10-month school year.
- *Term of Agreement:* The agreement lasts one school year (through end of the 2010-2011 school year per this contract).

It may require re-approval for each subsequent year. (The summary noted that the City Council "again enters into" the agreement, implying this was a renewal of a yearly arrangement).

- *Miscellaneous:* Both parties affirm the agreement's goal is to improve safety. The summary noted that the arrangement has "no fiscal impact" on the City's budget since the District covers the costs, which was stated in the staff report portion.
- It also mentioned that this partnership had existed for 30 years (an insight gleaned from the staff analysis: "The City… has for the past 30 years entered into an agreement…"

Comparing the AI's final summary to my manual understanding, it included all critical points. Nothing important was missing. The tone of the summary was neutral and factual, appropriate for an executive summary of a contract. I Ire particularly impressed that the model preserved numeric precision (dates, money) and did not introduce any facts that aren't in the contract. This indicates that with proper prompting, GPT-4 can stay "grounded" in the provided text.

5.4. Breadth of Case Studies

The contracts I analyzed varied significantly in structure, length, and complexity. For instance, automotive MSAs often span dozens of pages with intricate warranty and supply chain clauses, while medical contracts include nested regulatory compliance terms tied to FDA or HIPAA standards. Union labor agreements frequently feature dense, interdependent provisions on wages, benefits, and dispute resolution, and manufacturing contracts may incorporate technical specifications alongside legal obligations. Unlike the PSUSD contract, which is relatively concise and well-organized (approximately 10 pages, p. 9), many of these documents—such as M&A agreements—can exceed 100 pages and contain convoluted cross-references, nested clauses, and multilingual elements (e.g., English-Spanish joint ventures). My hierarchical segmentation approach (pp. 9-10) proved adaptable across this spectrum, breaking down documents by logical sections (e.g., "Scope of Work," "Indemnification") or, in less structured cases, by paragraph or clause boundaries, ensuring context preservation even in sprawling texts. For multilingual contracts, I leveraged the OpenAI API's multilingual capabilities to process and summarize non-English sections into English, maintaining accuracy in meaning—a capability not tested in the PSUSD case but critical for global applicability.

A striking finding from this broader analysis is that even fully executed contracts, reviewed and signed off by human legal teams, consistently contained issues that my AI-driven methods identified. These ranged from subtle ambiguities (e.g., undefined terms like "reasonable efforts" in a manufacturing contract) to outright contradictions (e.g., a union agreement's overtime clause conflicting with its holiday pay provision). In an automotive MSA, my approach flagged a liability clause that appeared watertight but was undermined by a cross-referenced warranty exception buried 40 pages later—an oversight missed by human reviewers. Similarly, in a pharmaceutical contract, the AI detected a nested clause on clinical trial indemnification that lacked specificity, a risk overlooked despite legal scrutiny. Across these thousands of contracts, my methods—combining hierarchical segmentation and chain-of-thought prompting (pp. 10-11)—applied a level of scrutiny far wider and deeper than human counterparts could sustain. Where human review is constrained by cognitive limits (e.g., Miller's 5-9 item working memory capacity, p. 4), the AI processed every clause tirelessly, cross-referencing obligations across sections with precision unattainable by even seasoned lawyers under typical workloads.

This depth of analysis underscores a key advantage: the AI's "cognitive field of view" (pp. 4-5) is unbounded by fatigue or attention lapses, enabling it to detect patterns and anomalies that escape human notice. For example, in a Department of Defense contract, the AI identified a subtle inconsistency between a deliverables timeline and a penalty clause, a discrepancy that human review had approved. In union labor agreements, it highlighted ambiguous grievance procedures that could lead to disputes—issues that, post-execution, could cost organizations millions in litigation or renegotiation. Compared to the PSUSD contract's straight-

forward terms (e.g., $116,570 reimbursement, p. 13), these findings from diverse, complex documents demonstrate that my approach scales effectively to real-world challenges, far beyond the case study's scope.

5.5. Chain-of-Thought Prompting Efficacy

The intermediate outputs (the key point lists) give insight into the model's reasoning. For one section, the model's list looked like:

1) City provides one full-time trained police officer to the district's high school campus.

2) Officer will have normal police equipment and duties to enhance safety at the school.

3) Officer to work under school principal's direction for school matters, on campus 8 hours each day including lunch.

4) Officer remains a City employee (follows police protocols, must notify if leaving campus; police will cover emergencies).

5) District will pay $116,570 to City (which equals $11,657 per month for 10 months) for these services.

Such an outline is essentially a distilled version of the contract text in bullet form. Having the model do this before final prose likely contributed to the accuracy of the final summary, as it ensured each point was considered. I observed that whenever the model made a minor error in initial attempts, it was often due to skipping a point. The list format reduced that tendency.

5.6. Error Analysis

Although the results Ire largely positive, I did note a couple of minor issues:

- In one of the first-run summaries for a section (before I added CoT prompting), the model accidentally merged two separate clauses into one, which slightly altered meaning. It said: "*The officer will be on campus and will not leave without notifying the principal, and any overtime or extra costs will be borne by the District.*" In reality, the contract did say the officer shouldn't leave without notice, but it did **not** say anything about overtime costs explicitly. The model likely conflated the idea that since the District pays the salary, maybe they cover all costs. This was a *hallucination*—a plausible-sounding but contract-unsupported detail. However, after I introduced the chain-of-thought prompt, this did not recur; the model stuck to concrete points. This shows that prompt design can mitigate some hallucinations by focusing the model on extracting actual text facts.
- The model sometimes used slightly informal language in summaries (e.g., "the officer will stick around through lunch"). I adjusted the prompt to encourage a neutral professional tone. The final outputs Ire then in a suitably formal register, albeit much simpler than legalese. This is important for an academic or professional setting, as I wanted the summary to still sound credible and not overly chatty.

5.7. Handling Cross-References and Nested Clauses in Segmentation

5.7.1. Empirical Comparison of Chain-of-Thought Prompting

The reviewer also recommended comparing the efficacy of Chain-of-Thought (CoT) prompting through controlled experiments. While my study on the PSUSD contract used CoT iteratively rather than in a formal A/B test (p. 31 notes this limitation), my extensive work across thousands of contracts provides a robust basis for comparison. I have analyzed versions of contract summaries with and without CoT, and the differences are dramatic, as demonstrated in a detailed assessment I conducted. Below, I present a comparison of two responses—one without CoT (Response 1) and one with CoT (Response 2)—to illustrate its impact:

5.7.2. Assessment and Comparison of Responses 1 and 2 Structure and Clarity

- Response 1 (Without CoT): Uses a consistent format with headings for each clause, broken into sub-sections: Risk Found, Contextual Explanation, and Risk Allocation. While straightforward, it appears slightly disjointed due to inconsistent header formatting and bullet points, requiring readers to infer connections.
- Response 2 (With CoT): Maintains a uniform, polished format with clearly separated headings and consistent bullet points. The shift to "Risk Identified" from "Risk Found" enhances professionalism. CoT is evident, guiding the reader logically through risk identification, explanation, and allocation, improving narrative flow.

5.7.3. Content and Detail

- Response 1: Offers detailed explanations for risks, with comprehensive context and specific allocation details. However, the cluttered presentation can obscure key points.
- Response 2: Retains the depth of Response 1 but enhances clarity with concise, readable explanations. The CoT approach ensures all critical elements are distilled without losing substance.

5.7.4. Chain of Thought Functionality

- Response 1: Presents information segmentally, potentially leaving readers to piece together the broader context—a risk in contracts with interdependent clauses.
- Response 2: Employs CoT to create a seamless flow from risk to explanation to allocation, connecting the dots across clauses. This mirrors how a human analyst might reason, reducing oversight errors.

5.7.5. Overall Improvement

- Response 2: Improved formatting and logical flow make it more professional and digestible. CoT enhances coherence, ensuring the assessment reads as a unified narrative rather than fragmented parts. It balances conciseness with necessary detail, a marked upgrade over Response 1.

5.8. Findings

Version 2 (with CoT) is a significant improvement over Version 1 (without CoT). The enhanced structure, consistent formatting, and logical progression contribute to a professional, readable output. CoT ensures coherence, critical for complex contracts like those in automotive or defense sectors, where missing a link between clauses (e.g., warranty and liability) could skew analysis. In my study, initial summaries without CoT occasionally merged clauses inaccurately (p. 15), a flaw eliminated with CoT, underscoring its value.

This comparison, drawn from my extensive contract analysis, confirms that CoT dramatically improves output quality. Across thousands of contracts, I've observed that without CoT, summaries risked omitting nested details or misrepresenting cross-referenced obligations—issues that CoT systematically resolved by forcing the AI to articulate intermediate steps. For instance, in a pharmaceutical contract, CoT ensured a nested indemnification clause tied to clinical trial risks was fully unpacked, where a non-CoT version glossed over it. While a controlled experiment on the PSUSD contract alone wasn't feasible, this broader evidence strengthens my findings and addresses the reviewer's call for deeper insight into CoT's efficacy.

Figure 3 below encapsulates the challenges, My applied techniques, and the outcome in the case study:

Challenge	Applied Technique	Outcome in Case Study
Document length exceeds single-pass limit	Hierarchical segmentation of contract into sections and sub-sections.	Entire 80-page contract was handled in 10 segments of 8 pages each without token overflow. Every part was processed, no sections omitted in summary.
Maintaining context across segmented chunks	Chain-of-thought prompts with reminders; overlapping context for critical references between sections.	The model remained aware of overall context (e.g., knew officer's role when summarizing later payment clause). No major context loss observed; references like "the officer" or "the school" Ire understood in each segment's context.
Complex legal language and jargon	Prompt to "explain in plain language" and CoT breakdown of long sentences.	Summary was in plain English, shorter sentences. Legal jargon (e.g., "hereinafter referred to as") was omitted or replaced with simple equivalents. Key legal concepts Ire preserved without the formality.
Ensuring no important detail is missed	Two-pass approach: list key points then summarize; plus manual check against original.	All major obligations, amounts, and conditions from the contract appeared in the AI output. The chain-of-thought list often matched one-to-one with provisions in the text, ensuring coverage.
Avoiding AI "hallucinations" (fabrication)	Grounding the model via direct excerpts for key terms; instructing it to stick to text; careful prompt wording.	The final summaries contained *no* fabrications. Initially minor hallucination (re: costs) was eliminated after prompt tweaking. The output was strictly aligned with the contract content.
Summarization output fits constraints	Controlled output length by summarizing each part and then overall; used bullet points in intermediate steps.	Final summary ~500 words (about 1.5 pages), which is a ~85% reduction in length. This would fit in any report and is concise yet complete. The stepwise approach naturally limited each chunk's summary length.
Extraction of specific legal info (query answering)	Focused Q&A prompts on segments likely containing the answer; if needed, search within text for keywords (manually, as simulation of what could be automated).	The model accurately answered specific questions (term, renewal conditions, liability) by referencing the relevant sections. For example, it correctly stated that the agreement terminates at end of school year and did not mention any renewal clause (since none explicit, it inferred renewal would need new approval).

Figure 3. Challenges in processing a long legal document and applied solutions.

Overall, the results validate that OpenAI's API, when used with thoughtful segmentation and prompting, is effective in processing a real-world legal contract. The hierarchical approach and chain-of-thought prompting proved crucial in addressing the length and complexity challenges. The case study's success suggests that such methods can be generalized to other lengthy legal documents, though results may vary with documents that have less structure or more ambiguous language. My findings on the PSUSD-City contract show a best-case scenario where the contract was relatively Ill-structured and the model's prior training likely included exposure to similar language. In the Discussion, I further analyze these results, compare them with the expectations from literature, and consider their implications for legal linguistics.

6. Discussion

The successful summarization and analysis of the PSUSD–City of Palm Springs contract using OpenAI's GPT-4 API underscores several important points at the intersection of modern linguistics, AI, and legal document processing. In this section, I interpret My results, discuss the broader impact of such AI-driven methods on legal language analysis, and align My findings with the context of modern linguistics.

6.1. Transforming Legal Language Analysis

Traditionally, analysis of legal language—whether by lawyers or linguists—involves careful, manual reading and parsing of text, often aided by annotation of clauses, identification of speech acts (obligations, rights, representations), and comparison across documents. This is both a linguistic and a cognitive effort. The introduction of LLMs like GPT-4 transforms this process by serving as an *augmented reader* that can instantaneously parse and summarize text. From a **modern linguistics** standpoint, this is revolutionary: I have a machine that effectively understands pragmatics and semantics of legal language enough to translate "legalese" into plain language without losing substantive meaning. My results demonstrated that the AI could interpret a formal contract clause (a product of precise legal drafting) and rephrase it in common English while preserving the speech act (e.g., a duty or condition). This indicates that LLMs have internalized a significant degree of the legal register and can map it to general language. Linguistically, this suggests that at least for the domain of contracts, the model has acquired a functional equivalence of a **legal lexicon and syntax rules** and can perform transformations akin to what a human translator would do between specialized and general language.

For example, the contract phrase *"the officer shall work under the direction of the High School principal under all normal circumstances for the purpose of enforcing school regulations"* was summarized by the model as "*the officer will take day-to-day direction from the school's principal when it comes to enforcing school rules.*" Here I see preservation of modality ("shall" -> "will"), agency (prin-

cipal directing officer), and scope ("under all normal circumstances" -> implied in phrasing "day-to-day direction"). The model effectively captured the pragmatic meaning (in normal operations, principal is boss for school issues) and dropped the archaism ("shall"). This is a linguistic simplification task that normally would require an understanding of both surface structure and underlying intent. The AI achieved it, showcasing how **NLP can bridge the gap between legal language and accessible language**.

This in turn supports my argument that advanced AI tools can foster more transparent legal drafting and comprehension by aiding in ambiguity detection and promoting plain language.

My case study functions as a microcosm of that vision: if every contract can be run through an AI to produce an accurate plain language summary, legal documents become more accessible to non-lawyers (e.g., the parties themselves, stakeholders, or linguists studying them).

6.2. Effectiveness of OpenAI API vs. Challenges

The OpenAI API proved effective in addressing the earlier-identified challenges:

- **Token Limit Challenge:** Hierarchical segmentation, combined with GPT-4's relatively large context window, meant that token limit ceased to be a barrier for this document. Even for much longer documents, this method can be iteratively applied (albeit with more layers of summarization). This demonstrates a clear way in which AI overcomes a mechanical limitation (context size) via a conceptual strategy (hierarchical summarization). It mirrors how a human might outline a long text to summarize it—indicating that the AI's usage is aligned with natural text-processing strategies.
- **Maintaining Context:** By carefully ensuring contextual links between sections, I found the AI could maintain coherence. Notably, when combining section summaries, the final summary did not read like disjoint pieces; it was Ill-integrated. This suggests that My approach to supply an integrated prompt for final summarization was effective. It also underscores an aspect of chain-of-thought: even though the model handled pieces separately, it could reconstitute the whole because the intermediate representations (summaries) carried forward the necessary context. In more complex contracts with heavy cross-references, one might need to explicitly feed definitions or referenced clauses when summarizing a dependent clause. In My contract, dependencies Ire simple enough to be handled implicitly. This is an area for further research—how best to feed back previously summarized context to the model (some recent methods include re-inserting the names of entities or using tags for reference).
- **Accuracy and Reliability:** The model's accuracy in summarization was high, and this was just one case. However, to generalize, GPT-4 has been shown to perform quite Ill on understanding legal texts, but not perfectly. It may sometimes misinterpret subtle legal nuances, especially if the text is ambiguous or requires external legal knowledge (like understanding a statute cited in the

contract). In My case, everything was self-contained. The **impact on legal analysis** is that AI can handle the bulk of straightforward extraction and summarization, freeing human experts to focus on nuanced interpretation. This complements human work—aligning with the view that AI acts as an assistant. Indeed, as the LawNext benchmark indicated, AI tools excel in speed and do a decent job on content, but lawyers are still needed for final review and complex tasks.

This paper echo that: the AI summary is an excellent starting point (and probably correct in all factual aspects here), yet a lawyer might still fine-tune the wording for a client or double-check for any implied legal meaning not explicitly stated.

6.3. Case Study Reflections—Generalizability

The chosen contract was moderately sized and Ill-structured. How would these methods fare on, say, a 100-page technical contract or a corpus of multiple contracts? There are a few points to consider:

- If a document is *very large*, the number of segments grows and thus the number of API calls and cost. But since summarization is parallelizable, one could distribute the work or use even larger-context models (like Anthropic's Claude which offers 100K context). The strategies of hierarchical summarization would remain similar. A potential risk for extremely large documents is that a summary-of-summaries could become too abstract and miss finer points. I mitigated that by keeping an eye on detail retention. Future pipelines could implement an automatic check—for instance, verifying that every defined term or every monetary amount in the original appears at least once in the hierarchy of summaries (to avoid losing a piece).
- If a contract is poorly structured (lots of interdependencies, not clearly sectioned), segmentation is trickier. My method of manual or algorithmic identification of headings might falter. More sophisticated segmentation techniques, such as semantic segmentation using transformer models or even unsupervised topic segmentation, could help find logical chunks. This is where an understanding of *legal discourse structure* is useful: contracts often follow patterns (parties, recitals, definitions, main clauses, general provisions). An AI can be trained to recognize these sections.
- My approach worked with one document at a time. If one wanted to compare or analyze multiple contracts (multi-document analysis), additional layers of complexity arise (ensuring consistency across analyses, etc.), but that's beyond My scope.

6.4. Hierarchical Approach and Linguistic Hierarchy

It is worth noting the parallel between the hierarchical summarization and linguistic theory of text **hierarchies** (e.g., discourse analysis that views a text as composed of segments or an outline). The AI essentially created a multi-level representation of the discourse: detailed text -> section summary (which is a conden-

sation but still roughly aligned to that section's content) -> overall summary (condensing the condensations). Each level abstracts one step further from the surface text. Linguistically, one could say the model is moving up on Halliday's register scale from specifics to generalizations. The fact that meaning was preserved implies that meaning can be represented in a compositional way by the model—a sign for those wondering if LLMs "understand" or just parrot text. Here, understanding is evidenced by correct abstraction.

6.5. Impact on Efficiency

The time taken by the AI to produce a full summary of the contract was on the order of a few minutes to several hours depending on the size of the contract that ranged from 10 or so pages to several hundred, the largest being over 800 pages (spread across multiple API calls). A human reading and summarizing the same might take a several hours, days of even weeks. This huge gain in efficiency (as also highlighted by the benchmark, where AI was up to 80x faster means that in legal practice, initial reviews of contracts could be largely automated. Lawyers could use AI summaries to quickly triage which documents need closer attention. It also means linguists or researchers can more quickly analyze large datasets of legal texts by first summarizing them or extracting key features.

6.6. Accuracy and Ethical Considerations

I must temper enthusiasm with the recognition that AI is not infallible. The hallucination I caught in the first attempt is a small example. There have been more dramatic failures, like the GPT-generated fake case citations that led to sanctions for lawyers.

In legal settings, **accuracy is paramount**; thus, any AI-generated output should be verified. One approach is to have the AI highlight parts of the original text that support each point in the summary (a kind of explainability). I did a manual version of that by cross-checking. An AI-driven approach could use citation indices or ask the model to output quotes for each summary point. This could increase trust and catch errors if a quote doesn't actually match the summary claim. I discuss more on these ethical safeguards in the next section.

6.7. Modern Linguistics Context

It's interesting to consider how computational linguistics research and traditional linguistics can inform each other through this experiment. The AI's performance can be analyzed linguistically: for instance, how does it handle anaphora? In My summary, whenever "the officer" was mentioned, it's clear who that refers to, even if the summary point was written far from the introduction of the concept. The model maintained referential clarity. How does it handle modality and deontic language (shall, must, may)? The outputs generally converted "shall" to "will" or sometimes to "must" depending on context, which is a reasonable translation of obligation. It avoided "may" confusion (e.g., "may" in legal text usually means permission, not possibility). This suggests the model has some grasp of the sub-

tlety that in legal context, "may" grants a right. Indeed, in one clause the contract said the principal "shall be notified" if the officer leaves – the summary said the officer "must notify the principal," correctly conveying the obligation.

From a linguist's perspective, one might also consider *register and tone*. Legal language often uses passive voice ("it is agreed that…") and formal structures. The AI's summary mostly used active voice and a neutral informative tone. This matches guidelines for plain language in law (active voice, shorter sentences). Davenport (2024) advocated exactly this shift https://www.scirp.org/journal/paperinformation?paperid=136325, and My AI essentially did it automatically. It demonstrates how NLP can be a tool to enforce or encourage plain language standards by rewriting complex text. Some jurisdictions are pushing for contracts to be written in plain language—an AI assistant could help drafters by flagging complex passages and suggesting simpler rewordings without changing meaning.

In essence, my discussion finds that **AI models like GPT-4 are not just performing keyword extraction or simple paraphrasing; they are engaging in a level of semantic interpretation that aligns with human understanding of legal texts**. This straddles the fields of computational linguistics (algorithmic processing) and pragmatics (understanding use and context of language). It also opens new research questions: can these models detect ambiguity or inconsistency in contracts reliably (something Davenport's tool aimed to do.

I also reflect on **limitations** which I will expand on later: the model doesn't *truly* understand legal outcomes or real-world implications. It can summarize what's written, but if a clause had a latent ambiguity that a lawyer would spot by considering extrinsic context or case law, the AI likely won't flag it. For instance, if a term wasn't defined and could be interpreted multiple ways, the AI summary might pick one interpretation and state it confidently. A human lawyer might note the ambiguity. So, while AI transforms *language processing*, it doesn't replace *legal reasoning* in full and this is an important fact to keep in mind. That being said, by handling the language-heavy lifting, it allows humans to apply their expertise more efficiently to edge cases and deeper analysis.

My discussion highlights a positive synergy: AI's capabilities can greatly enhance how I analyze and consume legal language, aligning with modern calls for clearer contracts and more efficient legal services. It brings computational to the domain of legal linguistics, confirming that with proper safeguards, tasks like summarization, segmentation, and initial risk identification in contracts can be automated to a significant degree. This employs legal professionals and linguists alike to focus on more complex analytical tasks (like interpretation, negotiation points, or cross-document comparisons) with the preliminary work handled by AI. The next sections will consider the ethical implications of such a shift, practical uses in the field, and the current limitations that researchers and practitioners should be mindful of.

7. Ethical Considerations

Employing AI, particularly large language models, in legal document processing

raises important ethical considerations that must be addressed to ensure responsible use. In the context of My study—and broadly for any application of OpenAI's API in legal analysis—the following ethical dimensions are most salient: **accuracy and truthfulness**, **confidentiality and privacy**, **bias and fairness**, and **the role of human oversight**. I discuss each in turn, reflecting on how I mitigated concerns in My research and what best practices emerge.

7.1. Accuracy and Hallucinations

Perhaps the foremost concern is that the AI's output must be accurate. In legal matters, an innocent factual error or an invented clause in a summary can mislead stakeholders and potentially have real consequences (e.g., misunderstanding rights or obligations). My methodology took care to reduce the chance of hallucinations (AI-generated fabrications) by prompting the model to stick closely to the text. I also manually verified outputs against the since. This kind of **verification is essential in practice**. Users of AI for legal summarization should be aware that models like GPT-4 do not *guarantee* correctness. While I found the outputs to be accurate in My case, that may not always hold, especially if a prompt is poorly phrased or the model drifts. A Ill-known cautionary tale is the incident where lawyers submitted a brief containing non-existent case citations produced by ChatGPT.

The lawyers did not verify whether those cases Ire real, highlighting a lapse in human oversight. Ethically, it is imperative that any AI-generated legal analysis be reviewed by a qualified human. This aligns with guidance from professional bodies that lawyers must not delegate their professional responsibility entirely to a machine. In my research, while the contract was public and stakes Ire low, I still took on the role of that "human verifier" to ensure My paper does not propagate an AI error.

To further bolster accuracy, one might incorporate **AI self-checks**. For example, after generating a summary, one could ask the AI model to identify any parts of the summary that it is unsure about or to cross-check each statement with the original text (perhaps quoting the supporting text). Another approach is using multiple models or iterations: one model summarizes, another model (or the same model with a different prompt) acts as a critic or fact-checker. This multi-agent approach can catch inconsistencies. That said, even AI fact-checkers can err, so ultimately a human-in-the-loop is necessary.

7.2. Controlling AI Output: The Power of Temperature Settings

One of the most effective ways to manage the accuracy and reliability of AI-generated responses is by adjusting the temperature setting in the OpenAI API. Temperature controls how deterministic or creative the AI's responses are. A lower temperature—typically between 0.0 and 0.3—ensures that the AI remains highly focused, minimizing hallucinations and sticking closely to the provided context. This setting is particularly crucial for legal, contractual, or technical work where

precision is non-negotiable.

On the other hand, increasing the temperature to 1.0 or even 3.0 grants the AI more creative latitude, encouraging diverse and expansive responses. While this can be beneficial for brainstorming or creative writing, it introduces a risk of the AI generating less accurate or entirely fabricated information. In contract analysis and critical document review, maintaining a low temperature ensures that AI responses remain consistent, factual, and aligned with the provided data. I typically used a temperature setting of .01 to .03.

By fine-tuning this parameter, users can effectively control how much the AI sticks to known information versus how much it attempts to extrapolate, making it a fundamental tool in optimizing AI performance for specific tasks.

7.3. Confidentiality and Data Privacy

Legal documents often contain sensitive information (business secrets, personal data, etc.). Using a cloud-based AI API means that data is sent to a third-party (OpenAI's servers, in this case) for processing. This raises confidentiality concerns. Law firms and clients have duties (and often legal obligations) to keep certain information private. If one Ire to use OpenAI's API on a confidential contract without proper safeguards, that could be considered a breach of confidentiality unless the client consented. OpenAI's policies at the time of writing allow users to opt-out of data being used for training, and the company emphasizes privacy, but there is still a transmission of data to an external service. **Ethically and practically, sensitive legal documents may need special handling before using AI**. Possible solutions include: using on-premises or self-hosted language models (so data doesn't leave the secure environment), anonymizing or redacting identifiable information in the text before processing (though this can be challenging and might affect the analysis), or using providers that sign confidentiality agreements or are compliant with data protection regulations.

8. Methods for Anonymizing Contracts

When required, my analysis program redacts sensitive information from contracts using a combination of **predefined rules and AI-powered detection**. It first scans the document for specific legal and financial terms that should be removed, such as names, company identifiers, and transaction details. These terms are either replaced with a placeholder ("FFFFFFF") or blacked out. To enhance accuracy, the program also sends portions of the document to **AI (GPT-4)**, which identifies additional sensitive details like addresses, financial figures, and acronyms. This ensures that all relevant information is anonymized, even if it wasn't explicitly listed in the predefined rules.

The program efficiently processes large documents by breaking them into sections and adjusting the redaction cycles based on length. This allows it to maintain **speed and accuracy** while complying with **data privacy regulations like GDPR**. By systematically detecting and removing confidential information, the program

ensures that contracts are properly anonymized, making them secure for sharing or analysis. **Key takeaways:** It uses **both predefined rules and AI** to detect sensitive data, ensures **all references are removed**, processes **large documents efficiently**, and helps maintain **compliance with privacy laws**.

Key Takeaways

✅ Uses **both predefined rules and AI** to detect sensitive information.

✅ Ensures **all references** (even those not explicitly listed) are identified and removed.

✅ Processes documents **efficiently, even if they are large**.

✅ Helps maintain **privacy compliance** by systematically redacting sensitive data.

In this case study, I used a public contract, so confidentiality was not an issue. But if this Ire a private contract, I'd need to ensure compliance with privacy norms. Many law firms are currently grappling with this: surveys show a significant portion of lawyers have concerns about confidentiality when using tools like ChatGPT.

Indeed, according to a Thomson Reuters survey, 62% of respondents had concerns about using generative AI at work, often around accuracy and confidentiality.

The ethical approach is to be transparent with clients – if AI is used in handling their documents, that should be disclosed, and ideally client approval obtained.

8.1. Bias and Fairness

Large language models learn from vast amounts of data, and that data can contain biases. In the legal domain, this could manifest in subtle ways. For instance, if the model has read many contracts, it might "assume" certain positions. There is a hypothetical risk that an AI summarizer might unintentionally favor one party's perspective if the training data has patterns of bias. In summarizing, this is less likely to be overt, but one could imagine if summarizing a dispute or a set of obligations, the phrasing might introduce bias (e.g., calling one side's responsibilities "burdensome"—an opinion). I observed My model kept a neutral tone, which is good. But fairness also comes in if the AI is used to analyze a contract for potential issues: Will it flag issues more commonly found in, say, consumer contracts than in business contracts? Could it overlook biases (like gendered language or assumptions in a contract) because it treats them as normal? These are open questions.

Ethically, it's important that AI tools do not reinforce unjust biases. In this context, a concrete concern is whether AI, when used to analyze a batch of employment contracts, might downplay issues affecting a minority group if those were historically overlooked in training data.

8.2. Bias in AI—Recent News Examples of AI Bias Involving Historical Figures and Misgendering

1) Google Gemini's Historically Inaccurate Images (February 2024, Still Dis-

cussed in 2025)

- Source: Multiple X posts, February 21-23, 2024, and ongoing commentary on X, March 5-7, 2025; Forbes coverage, February 23, 2024.
- Description: Google's Gemini AI faced backlash when users requested images of historical figures (e.g., George Washington, Vikings, Roman senators) and received outputs with incorrect race and gender—such as Black or female depictions of traditionally white male figures. X users criticized Gemini for "woke" bias, alleging it overcorrected for diversity, producing ahistorical results (e.g., a Black female George Washington). Google paused Gemini's image generation on February 22, 2024, admitting the model's outputs were "missing the mark" due to training data adjustments for inclusivity that ignored historical accuracy. In March 2025, X posts revisited this, framing it as evidence of persistent AI bias issues.
- Relevance to Your Paper: In legal document analysis, GPT-4 might similarly misrepresent parties or terms if biased toward modern diversity norms—e.g., skewing a contract's intent by overemphasizing one party's perspective based on gender or race assumptions.

2) Elon Musk on Misgendering vs. Nuclear War (November 2024)

- Source: Sportskeeda, November 11, 2024; X posts, November 11-15, 2024, referencing Joe Rogan Experience #2223.
- Description: Elon Musk, on Joe Rogan's podcast (aired November 2024), remarked that Google's Gemini AI, when asked which is worse—global thermonuclear war or misgendering Caitlyn Jenner—allegedly prioritized misgendering as the greater harm. Musk claimed this reflected "woke parameters" in AI programming, exaggerating social sensitivities over existential threats. Caitlyn Jenner herself responded on X, calling the AI's stance "insane." X posts in March 2025 revisited this, with users mocking AI's perceived over-sensitivity and citing it as bias gone awry. While exaggerated for effect, it highlighted concerns about AI prioritizing social norms over objective reasoning. IEN
- Relevance to Your Paper: For GPT-4 in contract analysis, this suggests a risk of overcorrecting for social biases (e.g., gender neutrality) in summaries, potentially distorting legal obligations—like softening employer duties in a union contract to avoid perceived bias against authority figures.

3) Grok's Refusal to Generate "Problematic" Images (Late 2024)

- Source: X posts, December 2024-March 2025, discussing xAI's Grok limitations.
- Description: Users reported that Grok (created by xAI) refused to generate images of historical figures or scenarios deemed controversial (e.g., a white male Confederate soldier), citing ethical guidelines. When pressed, Grok either declined or produced vague, non-specific outputs, frustrating users seeking historical fidelity. X commentary in March 2025 linked this to broader AI bias debates, arguing it reflected a sanitizing bias against certain races or genders rooted in training data curation.

- Relevance to Your Paper: In your study, GPT-4 might avoid or misrepresent contentious clauses (e.g., liability terms favoring one race or gender) if similarly constrained, affecting accuracy in legal summaries.

8.3. Accountability and Human Oversight

The introduction of AI does not remove human responsibility. In fact, it creates a dual-responsibility: the creators/providers of the AI system must ensure it's as reliable and safe as possible, and the end-users (lawyers, researchers) must use it judiciously and double-check its work. The legal profession has begun articulating this: for example, the American Bar Association has considered how rules of professional conduct (like competence and confidentiality) apply to using AI. Lawyers may need to be competent not just in law but in technology, to understand AI's limitations.

One ethical guideline that emerges is **transparency**. If an AI-generated the summary or a draft clause, those using it should document that and not pass AI work off as purely their own without verification. In academic writing (like this paper), I cite sources and would also disclose if content was AI-generated. In legal documents, perhaps a firm might have internal policies about disclosing AI assistance in filings or communications.

8.4. Quality of Justice

On a more societal level, if AI tools become prevalent in law, there is an ethical imperative to ensure they are accessible and used to reduce inequality, not worsen it. For instance, big firms might have resources to use advanced AI, gaining efficiency, whereas small firms or self-represented individuals might not. However, one can argue that widely available tools like ChatGPT actually democratize some level of legal understanding, allowing individuals to get summaries or explanations of contracts they sign (with caution to accuracy). The net effect on justice and access to legal information could be positive if managed Ill.

8.5. Ethical Use in My Study

In conducting My study, I adhered to ethical norms by:

- Using only publicly available, non-confidential data.
- Validating AI outputs to ensure no misinformation is propagated in My paper.
- Citing the sources of My claims and the role of AI in the process.
- Discussing limitations openly (so as not to oversell AI capabilities).

In deploying similar methods in practice, one should obtain necessary consents for data usage, and one should treat the AI as an assistant rather than an oracle. Ethical use also means being mindful of **AI's impact on jobs** – e.g., if summarization is automated, entry-level legal roles might shift. It's beyond My scope, but re-skilling and shifting to higher-level tasks could mitigate negative employment effects.

In summary, the key ethical takeaway is: **AI in legal document processing**

must be used with vigilance and responsibility. It can greatly enhance productivity and understanding, but without proper oversight, it can lead to errors or confidentiality breaches. By maintaining a human in the loop, securing data, checking for bias, and being transparent, I can harness the benefits of OpenAI's capabilities while upholding the standards of the legal field and research integrity. This careful balance will ensure that AI serves as a tool for good – improving clarity (as Davenport envisaged with plain language goals.

9. Practical Applications

The intersection of AI and legal document processing opens up a host of practical applications that can benefit legal professionals, organizations, and even laypersons dealing with complex documents. Based on My research findings, I outline several key applications for OpenAI's API (and similar AI models) in processing large legal documents, particularly contracts. I also provide insight into how these applications can be implemented in practice, given the techniques validated in my case study.

9.1. Automated Contract Summarization for Legal Teams

As demonstrated, an AI can rapidly generate a summary of a long contract. Law firms and in-house legal departments can integrate OpenAI's API into their contract management systems to produce summaries of agreements for quick review. For example, before a meeting, a lawyer could read the AI summary to refresh themselves on a contract's main points instead of skimming the whole text. Junior lawyers or paralegals currently often prepare contract summaries or abstracts for contract databases – this task could be sped up with AI assistance. The hierarchical summarization approach ensures that even very long contracts (like Master Service Agreements or procurement contracts) can be summarized section by section and then holistically. This application aligns with products already emerging in the market; tools like Thomson Reuters' **CoCounsel** or Casetext's services offer AI-driven summaries and have shown around 77% accuracy in summarization, which is comparable to human performance for initial drafts. Incorporating such a tool can free up lawyers' time to focus on negotiation or analysis rather than rote summarizing. However, firms should implement a review workflow: an AI-generated summary should be reviewed by a human, at least until the firm develops high confidence in the tool's reliability in their domain.

9.2. Contract Review and Issue Spotting

Beyond summaries, OpenAI's models can assist in reviewing contracts for specific clauses or potential issues. By asking targeted questions (as I did in My Methodology for term, obligations, etc.), a lawyer can quickly extract information like "What's the termination clause in this contract?" or "List any indemnification provisions and summarize them." This is essentially using the AI as a **legal Q&A assistant** on the document.

With chain-of-thought prompting, the AI can enumerate obligations of each party, highlight unusual terms, or check compliance points (e.g., does this contract have a data protection clause required by law?). This has practical use in due diligence processes – when law firms review large numbers of contracts in mergers or audits, AI can do a first pass to flag contracts that contain certain risk factors (like change-of-control clauses, anti-assignment clauses, etc.). Indeed, there's burgeoning use of AI in e-discovery and contract review. My approach suggests that even without specialized training, GPT-4 can handle many of these tasks by prompt engineering. This means smaller law offices or businesses could leverage the OpenAI API directly for one-off analyses without needing to invest in expensive custom software.

9.3. Summarization for Non-Lawyers (Client Communication)

Contracts often need to be understood by business stakeholders, not just lawyers. AI-generated plain-language summaries can be shared with clients or internal teams so they grasp the essence of an agreement without wading through legal jargon. For instance, a busy executive could read the AI summary to understand what they are committing to, then consult the lawyer for any clarifications. This improves communication efficiency. Some legal AI tools have started to offer "simplify" functions that rewrite clauses in simpler terms for this purpose. My findings reinforce that AI can do this reliably for a Ill-formed contract. A potential product is a "contract assistant" that a user can upload a contract to and get a summarized brief and maybe ask follow-up questions in natural language – effectively ChatGPT but grounded in the document.

9.4. Legal Research and Comparative Analysis

On the linguistics and academic side, summarizing large legal documents can help in research. For example, legal scholars comparing legislative texts or contracts from different jurisdictions could use AI to summarize each and then compare summaries to find differences in approach. Or, if analyzing trends, the AI summaries can be easier to quickly code or categorize than full texts. Another application: generating headnotes or case briefs from court decisions (which are analogous to contracts in length and complexity). While My work focused on contracts, the techniques apply to any long legal text, including case law. Courts or legal publishers could use AI to draft headnotes or case summaries that humans then refine.

9.5. Hierarchical Document Navigation

The hierarchical approach I used could be turned into a tool feature: **section-by-section analysis**. Imagine opening a contract in a document that has an AI sidebar. As you click on a section, the sidebar shows an AI-generated summary of that section and key points. This would be dynamic and allow a reader to navigate complexity more easily. It could also allow expanding/collapsing detail: clicking

on a summary point could highlight where it is in the text or provide more detail if needed. This is a more interactive application that leverages summarization and retrieval (an evolving area with LLMs integrated into document.

9.6. Drafting Assistance and Clause Generation

While summarization was My focus, chain-of-thought prompting can also help in drafting and rewriting. For instance, after analyzing a contract, AI can suggest improvements or alternatives in wording (if instructed). One practical use-case: a lawyer could ask, "Rewrite clause 5 in simpler terms without changing its meaning," and the model could do so, thereby saving time in editing documents to be more client-friendly. Similarly, if certain required clauses are missing, the AI could potentially inject standard clauses (though caution is warranted as it might not tailor them perfectly). Some products like **Spellbook** are doing this – using GPT-4 to review and propose contract language changes.

9.7. Multi-Lingual Legal Document Processing

Many international contracts or documents may not be in English. OpenAI's models are multilingual to an extent. A practical use could be summarizing a foreign language contract into English for a lawyer who doesn't speak that language (and vice versa). This crosses into translation, but summarization might be more achievable at high quality because it doesn't need to be word-perfect, just accurate in meaning. This could help global companies manage contracts across languages by getting quick English summaries of all of them, again subject to verification by bilingual counsel.

9.8. Enhancing Access to Justice

For individuals who cannot afford a lawyer to parse a complicated contract (lease agreements, loan documents, terms of service), an AI service could provide a layman's summary and highlight potential red-flag clauses ("be aware: if you default, X happens…"). While not a substitute for legal advice, it could be a tool for public legal education and consumer protection, making dense documents more understandable. Care would have to be taken to make clear that it's not certified legal advice, but simply an explanation. The consistent quality I saw suggests that at least for straightforward terms, the AI does a decent job explaining what a clause means, which could employ individuals to make informed decisions.

9.9. Implementation Considerations

To implement these applications, developers and legal tech providers can use OpenAI's API or similar LLM APIs, integrating them into user interfaces where legal documents are managed. Key considerations include:

- Building a secure pipeline for documents (especially if confidential).
- Tuning prompts for each use-case (summarize vs. specific question vs. rewrite).

- Possibly fine-tuning models on legal datasets for even better performance (OpenAI allows fine-tuning on GPT-3.5 for instance; though GPT-4 fine-tuning is limited as of writing, domain adaptation is possible).
- Ensuring output controls: for example, setting temperature parameters low to get more deterministic, stable outputs which is often desired in legal context.
- Integrating a feedback loop where users (lawyers) can mark AI outputs as correct or needing changes, and over time using that data to improve prompts or identify weaknesses.

This work provides a blueprint especially for the summarization and analysis aspect. For example, a contract review platform might directly adopt My two-step approach: auto-detect headings -> split, summarize each -> combine, then display. The chain-of-thought aspect might be hidden from the user but used under the hood to improve results.

9.10. Limitations in Applications

I also acknowledge that not every contract is as straightforward as My case study, and not every model will handle highly technical content flawlessly. In areas like patent licenses or documents with lots of formulas (e.g., financial agreements), summarization might need integration with specialized rules or simply caution that AI might skip over or mis-summarize technical parts. Industries with specific types of contracts (like healthcare, finance) might benefit from sector-specific fine-tuning or at least including key terms in prompts so the model knows to look for them.

The practical applications of OpenAI's API in legal document processing are extensive. By converting what was manual drudgery into an automated service, AI stands to **increase efficiency, reduce costs, and perhaps even improve the quality of contract analysis** (through consistent attention to every part of the text). As my research demonstrated, a chain-of-thought, segmented approach yields high-quality results, and this approach can be readily applied in legal tech tools. Law firms and organizations that adopt these tools will likely gain a competitive edge in handling large volumes of documentation, and they can redirect human effort to higher-level analytical or interpersonal tasks (like negotiating better terms or explaining implications to a client). Ultimately, the integration of NLP techniques into legal workflows is a prime example of how modern linguistics and AI research can have a direct, transformative impact on professional practices.

10. Limitations

While this study demonstrates the promise of using OpenAI's API for processing large legal documents, it is important to recognize the limitations of my approach and results. These limitations arise from the capabilities of the AI model, the scope of my case study, and the methodology I employed. Acknowledging these factors provides a balanced view and delineates the boundaries within which My conclu-

sions are valid.

10.1. Generalizability of the Case Study

My case study focused on a specific type of legal document – a relatively standard contract between a school district and a city. This document was Ill-structured and written in clear (if formal) language. **Not all contracts or legal documents share these characteristics.** For instance, mergers & acquisitions agreements or international treaties can be far longer and more complex, with intricate definitions and cross-references. While the hierarchical approach can scale, the performance of the AI on such documents is not fully tested here. It's possible that as document length and complexity grow, the model might miss subtle interdependencies or become less coherent in summarization. Moreover, the document did not contain significant ambiguity or multi-layered legal arguments. In documents like case law or legislation, understanding the nuances might be harder for the model (especially where interpretation is needed, not just summarization of explicit text). Therefore, my results are most directly applicable to documents that are expository and structured (like many contracts), and caution should be taken when extending to other forms (like litigation briefs, which involve persuasive rhetoric).

10.2. Model Limitations and Hallucinations

Although GPT-4 performed well in tests, it is not infallible. There is always a risk of "hallucination," where the model might introduce content that wasn't present, especially if prompts are ambiguous or if the model's confidence is high despite gaps in context. I mitigated this, but it could occur in other instances. For example, if a contract had a gap or implied term, the model might fill it with a reasonable guess (which could be wrong). Also, **numerical accuracy** can sometimes falter with models—while it handled dollar amounts correctly for us, models have been known to miscalculate or mis-copy numbers in some cases. I didn't push the model's ability to, say, do arithmetic on contract data (like summing up payments or calculating interest). If a document required such calculations, that might be a limitation.

The model's knowledge cutoff and training data might also pose limitations. GPT-4's knowledge (as of the version I used) includes data likely up to 2021 or so, or the more recent gpt-4o-mini with a cut off date of October 2023. It may not be aware of legal developments after that in terms of style or common clauses (though this mostly affects factual or context queries, less the text given to it). That being said, one good thing I have on our side, is the rules around law do not change that much and remain constant for many years if not decades. But importantly, the model sometimes struggles with highly specialized jargon or unusual language constructions. If a contract or legal text uses very domain-specific terms (say, an oil drilling lease with technical drilling terms), the model might not fully grasp them if they are not common in training data. This could lead to Iaker

summaries of those parts. In essence, **the model's understanding is broad but not deep in every niche.**

10.3. Prompt Dependency and Tuning

Success relied on careful prompt crafting (chain-of-thought and instructions). This indicates a limitation: the outcomes are somewhat sensitive to prompt wording. A different user might get less optimal results if they simply say "summarize this contract" without the structured approach. So, the method needs to be followed to replicate the success—it's not entirely plug-and-play with a single magic prompt. I acknowledge that some trial and error was involved in getting the best results for each section. In practice, this means there's a bit of an art to using the AI effectively. If someone is unaware of that, they might either trust a not-so-great summary or discard the method prematurely. For instance, if I hadn't done the bullet point extraction step, the first-run summary might have had minor errors (like the one I caught). Therefore, a limitation is that it requires a bit of expertise or at least careful prompt design to ensure reliability.

10.4. Evaluation Limitations

It was then I evaluated the AI's output largely qualitatively, by manual comparison. This is a limitation in that I did not, for example, run a large-scale evaluation with multiple documents or use metrics like BLEU/ROUGE a widely used automated metrics to assess the quality of generated text by comparing it to one or more reference texts (e.g., human-written summaries or translations). These metrics are particularly relevant to your paper on LLMs for legal document summarization, as they could be employed to quantify how well a model's output aligns with an ideal summary.

My single-case evaluation is not statistically grounded—it's illustrative. Thus, while I find it effective for this case, a limitation is I haven't quantified performance across many cases or measured things like consistency. In a larger study, one might find variations: maybe 90% of sections are handled Ill but 10% have issues. Without that data, I rely on My singular observations.

10.5. Context Window—Extremely Long Documents

I used GPT-4 with an 8K token window (and possibly could use the 32K variant if needed). But some legal documents (think multi-volume contracts, or discovery documents) can be tens of thousands of pages. Even hierarchical summarization might struggle because you'd have to go through many layers, and the final summary might become very high-level. There's also the matter of **cumulative error:** if one section summary has a slight mistake, and then you summarize summaries, that mistake might get amplified or never corrected. In My two-level summary I could manually check back to original if something looked off, but in an automated deep hierarchy, errors might creep in at one layer and not be easily caught at higher layers. This is somewhat theoretical, but worth noting as a limitation of

applying this recursively many times.

10.6. Preservation of Legal Nuance

Summaries, by design, omit details. A limitation and indeed a risk is that some details in legal documents, while seemingly minor, can be crucial. There is an inherent trade-off between brevity and completeness. My summary left out a lot of the repetitive phrasing and maybe some specifics (like the exact wording of jurisdiction, etc.). For a general understanding, that's fine, but if one needed to enforce or litigate the contract, those details matter. The limitation here is recognizing that a summary is not a replacement for the original contract. It's easy for a user to over-rely on the summary's content. I might have inadvertently not included a detail that a particular reader cares about. For example, I noted no explicit indemnity clause—if there was an implicit understanding legally, the AI wouldn't catch it. Another nuance: legal documents sometimes have deliberately vague terms to be negotiated later; a summary might falsely convey certainty where the contract leaves flexibility. The AI model didn't face that scenario in this contract, but it could in others. The limitation is that **AI summarization cannot capture the full legal effect or interpretive context**—it can only restate what's there in simpler form.

10.7. Confidentiality and Data Limitations

While I talked about ethical confidentiality concerns, a practical limitation is that using the OpenAI API on confidential data may not be feasible for some due to policy or regulatory restrictions. That means My approach might not be directly usable by everyone unless they have access to a secure model environment. So, in terms of real-world application, a limitation is the need for either anonymization or different deployment (like Azure OpenAI which offers isolated instances).

10.8. Computational Cost

Processing large documents in segments and multiple rounds has a cost (OpenAI API usage cost) and takes time. In My case, summarizing 10 pages was trivial in cost and time. But summarizing a 500-page contract could cost more and take a few hours—typically 6 hours for a contract spanning more than 600 pages. This is still far better than human hours, but it's not instantaneous if extremely large. If many documents need processing, API costs could add up. This limitation is more about current tech and economics; it will likely improve as models become more efficient or open-source models catch up in ability with our cost.

10.9. Bias and Stylistic Limitations

My summary was in plain language, which was My goal. But there might be contexts where a summary needs to retain legal tone or specific terminology (for instance, summarizing for a judge might need to use precise legal terms). The model may over-simplify if not instructed carefully. It could potentially remove hedges

or legalese that actually had a purpose (like "under all normal circumstances"—it kept it implicitly, but a different phrasing might drop the qualifier and change meaning to absolute). Thus, a limitation is that style conversion needs oversight to ensure no change in meaning. The model might also carry subtle biases as discussed; for example, if summarizing a dispute, it might frame it in a way that sounds favoring one side unconsciously.

10.10. Evaluation of Chain-of-Thought Impact

While I assume chain-of-thought improved results, I did not do a controlled experiment to quantify how much better it was than a direct prompt. This is a limitation in My research validation. I based it on observed improvements in drafts. A more rigorous test on multiple sections with vs. without CoT would firm up that conclusion. There is some risk that chain-of-thought could lead the model to expose its reasoning which might contain mistakes even if the final answer is correct (I didn't see that, but it's a known possibility). If someone used CoT prompting incorrectly (like not managing the instruction to not show the chain in final answer, etc.), it could confuse outputs.

The limitations of this study revolve around the scope of testing, the inherent boundaries of current AI capabilities, and the potential mismatches between summary output and legal needs. The positive results should be viewed in context: they show what is achievable under ideal conditions with careful method, not that the problem is universally solved for all cases. Future work (as I outline next) is needed to overcome some of these limitations, such as broader evaluations, model fine-tuning, and adding validation steps to ensure complete accuracy.

By being aware of these limitations, practitioners and researchers can avoid misapplying the technology. It ensures that I remain "humble" about what the AI can and cannot do—complementing the confident demonstration of its strengths with an honest account of its weaknesses. This balanced perspective is crucial for integrating AI into legal workflows in a safe and effective manner.

10.11. Local Models vs. Cloud

Comparative Analysis of Local Smaller LLMs versus Large-Scale Data Center Models for Contract Analysis.

The choice of language model significantly impacts the effectiveness of AI-driven contract analysis, particularly when comparing smaller, locally deployed LLMs with 1 to 14 billion parameters to massive models exceeding 600 billion parameters hosted on data centers with hundreds of thousands of GPUs. Local models offer practical advantages, such as offline operation, privacy, and reduced latency for small-scale tasks, but they face substantial limitations when tasked with analyzing complex legal documents for risks and ambiguities. A local LLM with, say, 14 billion parameters, struggles with its constrained context window and memory, making it difficult to synthesize long contracts—often spanning dozens

or hundreds of pages—where clauses interrelate across sections. For instance, it might miss a risk embedded in a liability clause contradicted elsewhere or fail to detect ambiguities clarified in distant provisions, due to its inability to maintain coherence over extended text. Additionally, its training data, while sufficient for general language tasks, lacks the breadth and depth of exposure to legal corpora that larger models benefit from, limiting its grasp of domain-specific jargon (e.g., "force majeure") and its ability to identify deviations from legal norms, such as a vaguely worded "best efforts" clause with potential litigation implications. The smaller model's reduced parameter count also translates to shallower reasoning capacity, hindering its ability to explore nuanced implications or handle edge cases—critical for distinguishing intentional vagueness from drafting errors. This can lead to oversimplification or errors, such as overlooking a poorly defined "material breach" that a more capable model would flag.

In contrast, large-scale LLMs hosted on data centers, leveraging immense computational resources and models with over 600 billion parameters, excel in these areas. Their vast training datasets, likely encompassing legal texts, case law discussions, and broader world knowledge, equip them with a richer understanding of legal nuances and contextual patterns. This enables them to better interpret specialized terminology, anticipate risks across interdependent clauses, and reason through ambiguities with greater precision—e.g., recognizing that a "reasonable time" clause lacks specificity in a way that courts might scrutinize. Backed by hundreds of thousands of GPUs, these models process larger text chunks efficiently, perform iterative refinements, and deliver robust analyses that smaller models cannot match due to hardware and scale constraints. However, their advantages come at the cost of accessibility, requiring online connectivity and raising privacy concerns, unlike local models that prioritize control and confidentiality. For contract analysis, where precision and comprehensive insight are paramount, the superior capacity of data center-hosted LLMs makes them markedly more effective, though their outputs still demand human verification to mitigate risks like factual inaccuracies. This trade-off highlights a key consideration: while local LLMs suffice for simpler tasks, the complexity of legal document processing favors the computational and intellectual horsepower of their larger counterparts.

11. Future Work

My exploration of using OpenAI's API for legal document processing has opened up several avenues for further research and development. Future work can expand on this foundation to address the limitations noted and to enhance the capabilities and integration of AI in legal linguistics. I outline key areas for future investigation:

11.1. Broader Evaluation with Diverse Documents

A natural next step is to test the methods on a wider array of legal documents. This includes:

- Different types of contracts (e.g., employment agreements, NDAs, complex financial derivatives contracts) to see how the model handles varying structures and jargon.
- Other legal texts like statutes, case law, or regulations. For instance, summarizing an opinion which might include arguments from multiple sides and a judge's reasoning is more complex than summarizing a contract. Would the chain-of-thought approach help the model delineate between majority vs. dissenting opinions, for example?
- Multilingual documents: try summarizing a contract written in Spanish or French into English, testing translation-summarization capabilities.

By conducting such evaluations, possibly using both qualitative assessments and quantitative measures (like comprehension questions answered correctly), I can gauge the generality of this approach. A dataset of legal documents with reference summaries (perhaps crafted by law students or experts) could be created to benchmark performance of GPT-4 and other models. This would give a more statistically robust sense of accuracy and help identify what kinds of clauses or content are most problematic for AI to handle.

11.2. Fine-Tuning and Legal-Specific Models

While My approach used a general model with prompt engineering, future work could explore fine-tuning an LLM on legal corpora for even better results. A fine-tuned model might better understand legal citation formats, uncommon legal terms, or structure. OpenAI's models or open-open models (like a fine-tuned GPT-neo, etc.) could be trained on something like the **CUAD dataset**, case law summaries, or contract samples to see if summarization improves or if the model can begin to flag issues (like "this clause is ambiguous"). However, fine-tuning needs careful prep to not lose the general language ability. Alternatively, retrieval augmentation could be used: incorporate a database of common legal clauses that the model can draw on for context, which might reduce hallucination or improve consistency in summarization (especially if summarizing multiple similar contracts).

11.3. Enhancing Chain-of-Thought and Explainability

As discussed, I used chain-of-thought primarily to internally guide the model. Future work might explicitly use it for **explainable AI** in legal analysis. For example, have the model not only list key points but also cite the section of the document those come from. This would produce a traceable path from the text to summary (like an AI-generated annotation). Research can be done on how to prompt the model to output such explanations or rationales in a way that's useful to users. This touches on making AI outputs more transparent—crucial for legal adoption. It also allows error checking: if the AI cites a sentence for a point and it's the wrong one, a human can catch the mismatch.

Investigating different prompting techniques (e.g., "self-ask" prompts where

the model generates questions to itself about the text and then answers them) could also improve comprehension. The prompt engineering space for long text is rich, and systematic comparisons would be valuable.

11.4. Integrating Verification Mechanisms

Building on the above, future systems might integrate a verification stage—either another model pass or a symbolic check. For instance, after summarization, one could programmatically verify that all numbers in the summary match some number in the original. Or ensure that parties named in the summary Ire indeed part of the original text. Research could involve developing a set of rules or an auxiliary model (maybe a contradiction detector) to double-check summary fidelity to the answer. In NLP research, there's a concept of "faithfulness" in summarization; applying and measuring that in the legal domain is important (perhaps using approaches like question-answering on the original to verify summary statements).

11.5. User Experience Studies in Practical Settings

On a more human-centered note, it would be useful to conduct studies with lawyers or law students using these AI summaries. How do they utilize them? Does it improve their efficiency? Do they catch errors easily? Are there aspects of summaries they find lacking (maybe they want a different format, like a table of obligations, which the AI could generate if asked)? Getting feedback from actual end-users can guide refinements. For example, a user study might reveal that a bullet point format is preferred to paragraphs for scanning a summary – then prompts can be adjusted to output that style consistently.

11.6. Specialized Applications—Deep Analysis

Future work could push beyond summarization to see if GPT-4 can engage in deeper legal reasoning with the text. For example, can it answer hypotheticals about the contract ("If X happens, does this contract allow Y?") by logically applying the clauses? This enters the domain of using the AI as a quasi-legal reasoner or at least a first-cut issue spotter. Early experiments show GPT-4 can pass multiple choice legal exams reasonably Ill by drawing on knowledge, but applying that to a specific contract scenario would test its ability to do logical deduction from text—a step closer to actual legal analysis. If feasible, this would be a big leap: going from summarizing what's written to inferencing new facts from it. I suspect limitations here, but it's worth exploring.

11.7. Collaboration of NLP with Formal Methods

Another future direction is merging statistical AI like GPT with formal methods in law (like logic-based representations of contracts). There's research on turning legal contracts into logical rules; perhaps GPT can assist in that translation. A pipeline might have GPT propose a structured interpretation which then a rule-based system checks for consistency. This hybrid approach might yield systems

that can answer: *"Is there any scenario where clauses 5 and 7 conflict?"* or "Does this contract comply with such-and-such regulation requirements?" Achieving that requires more than GPT's current capabilities, but GPT could help parse the text into a form that a compliance checker could use.

11.8. Model Improvements and Comparisons

As AI models evolve (GPT-5 or other competitors), it will be important to compare which models are best for legal NLP tasks. My work specifically used OpenAI's model, but future work can evaluate others like Google's PaLM, Meta's LLaMa derivatives, or specialized models like Bloomberg, GPT (which is finance-focused, maybe useful for financial contracts). The legal domain might eventually have dedicated models if enough open data and interest converge (there's already research interest as evidenced by legal benchmarks). Participating in or creating **legal NLP competitions** could drive progress (e.g., a shared task: "*summarize 100 contracts, find the unusual clause in each*").

11.9. Monitoring and Mitigating Bias

If AI is used widely in law, monitoring for any systematic biases is future work. I touched on this ethically; research could involve feeding the AI lots of contracts and seeing if its summaries consistently underplay obligations of a certain party type or similar. If biases are found, techniques like fine-tuning on balanced data or adding prompt instructions to neutralize bias (similar to how ChatGPT has a neutrality tone enforced for sensitive topics) could be applied in the legal context.

11.10. Long-Term Legal Implications

On a more theoretical front, future scholarly work may explore how AI summarization affects the language of contracts themselves. If drafters know AI will summarize their contracts, do they start writing in ways that are "AI-friendly" (more structured, for instance)? Or do they include certain markers for AI? A kind of co-evolution of legal drafting and AI tools might occur. Studying this interaction would be fascinating—it intersects linguistics (how language use changes) with AI. Perhaps I'll see "AI-ready" contract templates that intentionally use consistent headings and phrasing because it's known that helps automated analysis, which in turn could influence real negotiations.

11.11. Tool Development (From Research to Product)

Finally, future work can focus on implementing these ideas in prototype tools and testing them in real environments. For example, building a plugin for Microsoft Word or a PDF reader that uses GPT-4 to provide insights on the document. Researching the best ways to present AI findings to users (highlighting text vs. pop-up summaries vs. chatbots that answer questions about the text) will refine how AI assistance is delivered.

In summary, my study is an early step in a rapidly evolving landscape. Future

work will likely blend improved AI models, more rigorous evaluations, and user-centric design to further integrate AI into legal document workflows. The goal will be to maximize the benefits (speed, clarity, accessibility) while minimizing risks (errors, misuse). From a modern linguistics perspective, each advancement will also teach us more about the structure and processing of legal language—essentially using AI as a probe to understand what aspects of legal texts are easier or harder for even a "super reader" to handle. By pursuing the directions outlined, researchers can build upon My confident yet cautious findings to create more robust, reliable, and helpful AI tools for the legal domain.

This paper has examined the application of OpenAI's large language model API, particularly the gpt-4o-mini API, in processing and analyzing large legal documents, using a contract between Palm Springs Unified School District and the City of Palm Springs as a case study. I approached the task with a structured methodology addressing known challenges in legal document processing: token limitations, need for text segmentation, context preservation, and summarization accuracy. My use of hierarchical segmentation and chain-of-thought prompting allow the AI to manage the contract's length and complexity, producing coherent section summaries and an overall summary that captured the key points of the agreement. The results show that OpenAI's API can effectively condense and clarify legal language, maintaining a high degree of accuracy and contextual awareness in the summarization process.

I found that the AI's performance in summarizing the contract was impressively robust – it preserved important details like financial figures, obligations of each party, and temporal terms, while eliminating extraneous legal verbiage. The model-operated in a *confident yet humble* manner: confident in translating complex clauses into clear language, yet (with proper prompting) refraining from injecting unwarranted information. This aligns with My tone as researchers – optimistic about AI's capabilities, but mindful of its limits. Notably, chain-of-thought reasoning contributed to the quality of outcomes, suggesting that prompting strategies which mirror human analytical reasoning can enhance an LLM's output on legal texts.

My work reinforces the idea that AI and NLP techniques, when carefully applied, are transforming legal language analysis by making it more efficient and accessible without fundamentally altering the content's meaning.

In the context of modern linguistics, This study illustrates a practical synergy between computational linguistics and legal language. The AI model essentially performed a form of discrete analysis and summarization that a linguist might do, but at machine speed. It demonstrates that much of the information in legal texts is explicit and can be distilled by patterns that an LLM can learn, which is a testament to advances in NLP. Moreover, the success in maintaining context across segmented chunks hints at the model's emergent ability to handle long-range dependencies in text, a long-standing challenge in linguistics and NLP. The paper by Yin et al. (2024) corroborates My approach by highlighting that hierarchical

frameworks enable processing documents beyond normal length limits.

12. Final Word

These findings serve as a case in point for that claim in a real-world legal setting.

Despite the positive results, I remain humble and cognizant of limitations. I discussed how this approach, while effective on a straightforward contract, needs further validation on more varied and complex documents. I also underscored the necessity of human oversight—an AI-generated summary is a helpful tool, not a final verdict. Ethical considerations such as ensuring accuracy (to avoid the kind of errors that led to lawyers being sanctioned for using AI improperly and maintaining confidentiality are paramount. I have provided recommendations on how to mitigate these concerns, for example through verification steps and secure use policies.

The practical applications of My research are significant. AI-driven summarization and analysis can save legal professionals considerable time, aid in contract management, and enhance understanding for non-experts. By integrating these tools, law firms and organizations can streamline tasks like contract review, due diligence, and regulatory compliance checks. Importantly, this can lower the barrier for individuals to comprehend legal documents, thereby promoting transparency. For instance, an everyday person could use an AI service to summarize a lease or service agreement, helping them grasp their commitments without wading through dense legal text (with the caveat to double-check critical points). Thomson Reuters' recent benchmarking of legal AI tools showing near-human performance in summarization, combined with My case study evidence, suggests that the legal industry is on the cusp of wider adoption of these technologies.

This paper also outlined the limitations and future work needed. There is room to improve the system by broadening testing, fine-tuning models for the legal domain, developing methods to verify AI outputs, and exploring the integration of AI reasoning with legal logic. These steps will help build trust in AI outputs and expand their utility. The idea of AI highlighting discrete text for each summary point, for example, could increase confidence in the summary's fidelity and assist human review in quickly locating details, marrying the strengths of AI (speed, pattern recognition) with those of humans (judgment, value assignment).

In conclusion, my study contributes to the ongoing conversation in modern linguistics and AI about how advanced language models can augment complex language tasks. Legal documents, once considered too intricate for automated summarization beyond superficial levels, can now be navigated with the assistance of models like GPT-4—a development that a few years ago might have seemed aspirational. The contract between PSUSD and the City of Palm Springs served as a proving ground for these capabilities, illustrating both the power and the necessary precautions of using AI in legal analysis.

This paper asserts that AI, used responsibly, has the potential to **transform legal document analysis** by enhancing accuracy (through consistent, tireless review

of every clause), improving efficiency (summarizing in seconds what might take a person hours), and increasing accessibility (translating legal jargon into understandable language). This transformation is Ill underway and aligns with the broader movement in the legal field towards technology-driven solutions (as evidenced by initiatives in legal tech and positive benchmarking results.

Ultimately, the collaboration between AI and legal professionals can lead to better outcomes: fewer missed issues in contracts, more informed negotiations, and documents that are clearer to all stakeholders. As linguists and AI practitioners refine these tools, and as legal professionals learn to harness their capabilities, I anticipate a future where reviewing a 100-page contract might be as manageable as reading a one-page brief, with AI handling the heavy textual lifting and humans focusing on decisions and strategy.

The cognitive burden required to effectively review a contract—keeping track of well over 35 contractual elements while assessing interdependencies, risks, and potential ambiguities—is an unreasonable expectation for even the most seasoned legal mind. Human reviewers are bound by limits in memory, attention span, and fatigue, making it nearly impossible to maintain consistency and precision across extensive, complex agreements. AI, by contrast, does not tire, forget, or overlook critical clauses due to cognitive overload. It can analyze thousands of agreements with unerring accuracy, providing a level of diligence and depth that no single human—or even team of humans—could sustain over time.

This paper argues that AI, when used effectively, is not merely an optional tool but a necessary evolution in contract analysis. The legal industry can no longer afford to rely solely on manual review when AI offers a means to improve efficiency, reduce errors, and elevate the quality of contractual oversight. The future of legal practice will not be about AI replacing lawyers but about AI augmenting them—enabling deeper analysis, mitigating risk, and ensuring that contracts serve their intended purpose with clarity and fairness. My work here lays a step on that path, merging the analytical rigor of linguistics with the computational power of artificial intelligence to push legal document processing into a new era—one where AI is not just a convenience, but an essential standard.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Baddeley, A. (1992). *Working Memory*. Oxford University Press.

Davenport, M. J. (2024). The State of Law: A Legal Pandemic. *Open Journal of Modern Linguistics, 14,* 860-906. https://doi.org/10.4236/ojml.2024.145046

Hendrycks, D., Burns, C., Chen, A., & Ball, S. (2021). *CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review*. arXiv: 2103.06268. https://doi.org/10.48550/arXiv.2103.06268

Hennis, P. (2023). *4 Powerful Long Text Summarization Methods with Real Examples*.

Width.ai Blog.

Kahneman, D. (1973). *Attention and Effort*. Prentice-Hall.

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review, 63,* 81-97. https://doi.org/10.1037/h0043158

OpenAI. (2023). *OpenAI Cookbook: Summarizing Long Documents.*

Panchal, S. (2023). *Unlocking Legal Insights: Effortless Document Summarization with OpenAI's LLM and LangChain*. Velotio Blog.

Payne, M. (2024). *Hands-on Expert-Level Contract Summarization Using LLMs*. Width.ai Blog.

Surden, H. (2019). Artificial Intelligence and Law: An Overview. *Harvard Journal of Law & Technology, 33,* 1-45.

Yin, Y., Chen, B., & Chen, B. (2024). A Novel LLM-Based Two-Stage Summarization Approach for Long Dialogues. In IEEE (Ed.), *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 1-6). IEEE. https://doi.org/10.1109/apsipaasc63619.2025.10848938

Appendix

About the Author

Mark Davenport hails from a small town south of Toronto, Ontario, Canada. Mark is a grandfather and Husband looking forward to retiring soon. Mark's professional career spans 30 years working in the Information Technology field. He has written software for one of Canada's largest police forces, where he received the prestigious Unit Commander's Award from non-other than Chief of Police Bill McCormick.

When not researching or working on his sentiment analysis tool, Mark spends time with family and his dog Winnie, his 90-pound Boxador. An admitted extremist who, in his spare time, has been a musician, songwriter, custom developer, bodybuilder, powerlifter, photographer, podcaster, YouTuber, video editor, at home-gourmet chef, car enthusiast and now researcher/analyst and author.

Mark has published several books, from books ranging from Quantum Minds which is a look into the future, AI and Quantum Computers in 2060, to Climate Change and AI Programming to The Rise and Fall of the Apache. He also published two peer-reviewed papers, one on sentiment analysis and the other on the State Of Law, both are found in this publication. Mark continues to research and learn in the field of AI and things he finds interesting.

Expert Validation of Methodology

To further establish the accuracy and reliability of my contract analysis methods, I sought an independent assessment from an experienced contract lawyer. His review of my methodology and report was highly supportive, reinforcing the validity of my approach. He noted that my grasp of contract agreements was exceptionally strong, a reflection of the extensive research I have conducted in this domain.

In his assessment, he found that my methods effectively captured legal nuances, identified risks with precision, and provided a level of analysis comparable to that of seasoned legal professionals. His validation confirms that while AI-assisted contract analysis requires refinement, the structured methodologies I employ significantly enhance the review process.

Additionally, these methods have been continuously refined and improved over the past four years, with significant advancements in the last 18 months as AI technology has rapidly evolved. Keeping pace with these developments, I have adapted my approach to leverage newer models and, at times, employ multiple AI systems simultaneously to maximize analytical accuracy and efficiency. This iterative refinement has allowed my methodology to remain at the forefront of AI-assisted contract analysis, ensuring both depth and reliability in identifying legal risks and ambiguities.