BROOKLYN SOLUTIONS

Benchmarking the Bar:

Ask Brooklyn AI Model versus Qualified Lawyers in Contract Analysis

Nick Francis
Chief Technology and Marketing Officer
Brooklyn Solutions

Table of Contents

Table of Contents 1

Executive Summary **2**

1. Introduction **3**

2. Methodology and Benchmarking Approaches **3**

3. Performance Metrics and Key Findings **4**

3.1 Accuracy in Legal Issue Identification 4
3.2 Processing Speed and Efficiency 5
3.3 Cost-Effectiveness 5
3.4 Context Window and Document Length Capabilities 5

4. Specific Contract Analysis Capabilities **5**

4.1 Playbook Implementation and Generation 5
4.2 Contract Understanding and Response Quality 6
4.3 Complex Reasoning Tasks 6

6. Practical Applications and Integration **7**

7. Limitations and Considerations **8**

8. Future Directions **8**

9. Conclusion **9**

..... 9
What are F Scores 10

DISCLAIMER 10

Limitations of AI Contract Analysis **10**

Data Quality and Structure 10
Legal Standards for Enforceable Clauses 11
Jurisdictional Limitations 11
Not Legal Advice 11
Technological Evolution 11
Human Oversight Requirement 11

Executive Summary

AI vs. Lawyer: Who Wins the Contract Analysis Bout?

The legal landscape is undergoing a dramatic shift as AI takes centre stage. The introduction of AI-driven legal support has massively impacted the way procurement and supplier management teams approach contract management and analysis. Once reliant on time-consuming manual analysis by in-house or third-party legal teams, Contract managers are now looking to be replaced or augmented by advanced AI tools such as Large Language Models (LLMs).

In this whitepaper, we put legal professionals head-to-head with an advanced AI solution: **Ask Brooklyn**, our platform's built-in AI assistant. We test its ability to deliver the kind of legal insight that was once the sole domain of human lawyers. As AI continues to gain traction in the legal sector, understanding how it performs against traditional methods is critical for legal departments, procurement teams, and technology providers alike. So, when both contenders step into the ring, who will emerge as the contract analysis champion?

1. Introduction

Amongst many occupations, the legal profession is experiencing a significant transformation, driven by the innovations of Artificial Intelligence (AI). This has impacted their daily admin and resource requirements as businesses begin to use Large Language Models for legal support opposed to seeking human resources.

An area that has been impacted by this transformation is Supplier and contract management. Before AI, reviewing contracts was a labour-intensive process requiring specialised legal expertise that organisations would use internal legal teams in-house or outsource to third-party legal entities for support. This process would become time-consuming and labour-intensive for all stakeholders involved. With the advancement of Large Language Models, there is growing evidence that these AI systems can perform specific legal tasks with comparable or superior performance to human lawyers in particular contexts.

Ask Brooklyn, our AI Assistant, is powered by Claude LLM Model, built into our Customer-Supplier Management Platform. Built into the platform to assist users in effectively managing their vendors and navigating complex contracts. The AI Assistant enables users of the platform to review specific clauses, summarise entire documents and interpret legal language into plain English, helping users make sense of everything from data processing agreements to compliance requirements under regulations. It is a practical tool to make contract review and analysis more manageable and transparent.

But how is AI changing the way businesses use qualified lawyers for contract analysis, and is it just as effective? Within this whitepaper, we will comprehensively analyse the Large Language Model (LLM) Ask Brooklyn to assess how this AI Assistant benchmarks against a qualified lawyer in contract analysis. This benchmark assessment will be through key performance indicators: Accuracy, speed, cost-efficiency, and practical applications.

So, Ask Brooklyn and a Qualified Lawyer are in the ring, but who is going for the knockout?

2. Methodology and Benchmarking Approaches

As Organisations and Legal professionals explore the integration of AI into contract analysis and review, it is essential to understand how Large Language Models (LLMs) are being rigorously evaluated. Leading academic and industry studies have developed structured benchmarking frameworks to assess LLM performance in comparison to human legal reviewers, particularly Junior Lawyers and Legal Process Outsourcers (LPOs). These methodologies are designed not only to measure task-level accuracy but also to reflect the nuanced demands of legal workflows, such as interpreting complex clauses, maintaining

contextual accuracy over long documents, and delivering insights efficiently. By combining quantitative metrics (e.g. F-scores) with large-scale testing across a diverse set of contracts, these benchmarking efforts provide a robust foundation for evaluating how well LLMs can meet the high standards of legal practice.

These studies typically employed rigorous methodologies including:

- Comparison of LLMs with Junior Lawyers and Legal Process Outsourcers (LPOs)
- Evaluation using F-scores for issue identification and localisation accuracy
- Analysis of performance across multiple contract types
- Assessment of speed and cost differentials
- Testing on over 20,000 samples from open and private benchmarks

Within these benchmarking efforts, the research specifically test capabilities essential for legal applications, including:

- Legal reasoning and contract understanding
- Detection of hallucinations and accuracy of information
- Following complex legal instructions
- Handling long context windows (necessary for lengthy contracts)
- Multihop reasoning (synthesising information from multiple locations in text)
- Multitarget reasoning (locating and returning numerous pieces of information)

3. Performance Metrics and Key Findings

3.1 Accuracy in Legal Issue Identification

Recent research has shown that Advanced LLM models can match or even surpass human accuracy when identifying legal issues in contracts. One groundbreaking study by Martin, Whitehouse, Yiu et al, upon assessment, Legal Process Outsourcers (LPOs) achieved an F-score of 0.77 in detecting legal issues, while the leading LLM, GPT-4-32k, delivered a closely comparable F-score of 0.74. The study assessed performance across accuracy, speed, and cost-efficiency during contract review. These results highlight the potential for advanced LLMs to rival human expertise in legal issue identification.

Brooklyn's Built In AI Assistant has shown particular strength in this area with the Ask Brooklyn - Claude model 3.7 Sonnet model recently surpassing all other LLMs in benchmarks for identifying legal concepts and clauses in Vendor Agreements. This represents an 8% performance improvement over previous Anthropic models, with an overall 87.5% increase in performance of Anthropic models to locate content in contracts since March 2024.

3.2 Processing Speed and Efficiency

One of the most significant advantages of LLMs in contract analysis is processing speed. Benchmark studies reveal that while human lawyers might require hours to review complex contracts, LLMs like Ask Brooklyn can complete these reviews in seconds. This dramatic speed differential enables legal teams to process substantially higher volumes of contracts without proportionally increasing time investment.

3.3 Cost-Effectiveness

The economic implications of implementing LLMs for contract analysis are substantial. Studies indicate a potential **cost reduction of up to 99.97%** when comparing LLM-based contract review with traditional human review processes. While this doesn't necessarily translate to direct replacement of legal professionals, it suggests significant opportunities for cost optimisation within legal operations.

3.4 Context Window and Document Length Capabilities

The use of LLM systems implemented in SRM and CLM tools, such as Ask Brooklyn's integration into Brooklyn's Customer–Supplier Management platform, has enhanced the ability to analyse longer contracts and documents, enabling users to find clauses, definitions, and other contract data buried in hundreds of pages of text. This capability to handle extended context is critical for comprehensive legal analysis of complex agreements.

4. Specific Contract Analysis Capabilities

As LLMs continue to advance, their application in contract analysis has become increasingly significant. Evaluating these systems involves examining their ability to accurately interpret, extract, and assess complex legal information within contracts. Among the models leading the way, solutions such as Ask Brooklyn GenAI, have consistently demonstrated strong proficiency across a range of contract analysis tasks. In this section, we will explore and break down the key capabilities these models bring to the table.

4.1 Playbook Implementation and Generation

Recent benchmarks indicate that Ask Brooklyn and other LLMs effectively compare contract risks against established playbooks or standard topic guidance. This mimics how lawyers assess legal risks. Our AI Assistant, Ask Brooklyn, has shown superior performance in

generating specific and valuable guidelines from standard contracts compared to other, more generic outputs from other models.

4.2 Contract Understanding and Response Quality

The integration of the AI Assistant Ask Brooklyn LLM into the platform resulted in higher-quality responses to questions regarding contracts. This includes tasks that require the interpretation of multiple clauses or paragraphs of text across multiple documents. These responses from the AI Assistant demonstrate increased reliability, with appropriate references to relevant text or clauses that answer specific questions.

4.3 Complex Reasoning Tasks

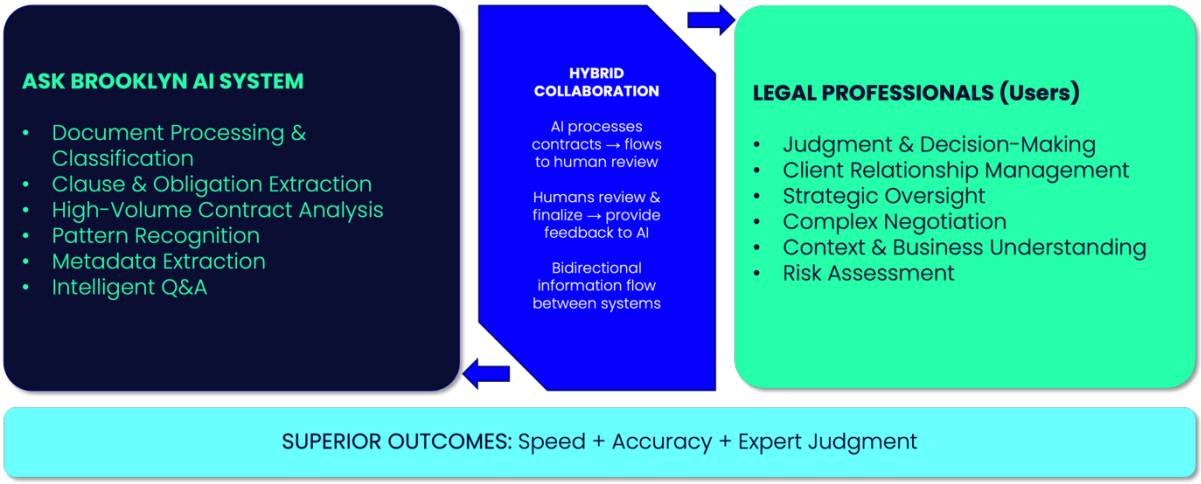
Ask Brooklyn models excel at sophisticated legal reasoning tasks that require:

- Interpreting contracts and regulations where definitions in one part determine how another part is applied
- Synthesising information from multiple sections of lengthy documents
- Maintaining consistency across varied contract types and structures

5. Industry Perception and Adoption Patterns

Despite the impressive performance metrics of LLMs, the industry perception still remains cautious. A major survey revealed that while only 1 in 10 people across the US and UK fully trust law firms, only 4% said they would trust AI independently for legal advice. The vast majority preferred either a traditional lawyer (69%) or a lawyer using AI as a support tool (27%).

This indicates that despite the AI resource out there, organisations are looking to use these as an augmentation tool opposed to a replacement for legal support. We fully agree with this statement, and it is something that organisations are looking at doing to enhance their The prime focus is not on whether computational systems should replace lawyers but rather on determining the degree to which these systems can execute tasks that require legal reasoning, with the goal of augmenting, educating, or assisting our users in increasing productivity and job satisfaction.



Collaborative Human-AI Legal Contract Analysis Model: Ask Brooklyn

6. Practical Applications and Integration

Large Language Models are being successfully implemented into solutions to enhance Contract and Supplier management platforms. Ask Brooklyn has been successfully implemented across the Brooklyn platform into our workflows as an AI Assistant to our users. Below are some of the ways that Ask Brooklyn have been implemented into the platform:

- **Intelligent Contract Search & Knowledge Access**: Unified AI-powered search system that works across all platform content (documents, contracts, reviews, risk data) with natural language processing to understand context and deliver precise answers.
- **Virtual SRM Assistant**: Conversational agent providing natural language Q&A about suppliers, contracts, and organisational data through a secure interface with context-aware dialogue capabilities that maintain conversation history.
- **Workflow Automation**: Launch procurement processes, supplier onboarding, issue logging, and data updates directly from the chat interface, seamlessly connecting inquiries to actions.
- **Personalised Dashboard Integration**: User-specific visibility into outstanding tasks, reviews, and meetings based on role and activity history, transforming the system into a proactive assistant.
- **Document Intelligence**: Automated scanning, classification, and extraction of key information from emails, contracts and attachments, including obligation extraction in structured formats.

- **Meeting Enhancement Tools**: AI-generated meeting preparation materials including agendas, topic outlines, and recommended actions based on historical data and analytics.
- **Communication Support**: Draft generation for notifications triggered by metric failures or other conditions, with translation capabilities to support global teams.
- **Metadata & Workflow Optimisation**: AI-powered extraction of relevant fields from contracts and documents, with automated workflow generation based on standard sequences.
- **Data Analysis & Insights**: Contextual query capabilities for vendor performance analysis, risk profiling, and contract evaluation, with guided interfaces for selecting data sources and generating tailored reports.

These practical implementations demonstrate how Ask Brooklyn's benchmarked capabilities translate into tangible improvements in workflows.

7. Limitations and Considerations

While Ask Brooklyn and other Large Language Models (LLMs) demonstrate impressive performance in contract analysis, before implementing a model, it is important to understand the limitations of their capabilities. Below are several important considerations that you should be informed of ahead of implementation:

- Human oversight remains essential, particularly for complex or high-stakes contracts
- The hybrid approach (Structured Data + Search + AI + lawyer) currently shows the highest performance, acceptance and trust
- Regional variations in legal systems may impact performance
- Ongoing evaluation and benchmarking are necessary as both AI models and legal requirements evolve

8. Future Directions

The benchmarking data suggests several promising directions for future development for GenAI and LLMs tools. These include:

- Continued improvement in context handling for increasingly complex contracts
- Enhanced reasoning capabilities for jurisdiction-specific legal interpretation
- More sophisticated integration into existing legal workflows
- Development of specialised legal domain knowledge

- Improved explanation capabilities to increase trust and adoption

9. Conclusion

Ask Brooklyn's underpinning LLM models have demonstrated impressive capabilities in contract analysis when benchmarked against qualified lawyers. Their accuracy, speed, and cost-efficiency performance suggest significant potential for transforming legal workflows. However, the predominant industry direction is toward augmentation rather than replacement of legal professionals.

The most effective LLM implementation appears to be a collaborative approach to where Ask Brooklyn handles the time-consuming aspects of contract analysis while legal professionals provide oversight, judgment, and client interaction. This hybrid model leverages the complementary strengths of AI and human expertise to deliver superior outcomes in contract analysis and legal services more broadly.

What are F Scores

The F-scores mentioned in the white paper are statistical measures used to evaluate the performance of LLMs in legal contract analysis compared to human lawyers.

In the study comparing Large Language Models with legal professionals, F-scores were used to measure accuracy in pinpointing legal issues within contracts. Legal Process Outsourcers (LPOs) achieved an F-score of 0.77, while the top-performing LLM (GPT4-32k) achieved an F-score of 0.74. Another model, GPT4-1106, scored 0.69 for localisation accuracy. ArXiv

An F-score is a statistical measure that combines precision and recall into a single metric, providing a balanced assessment of a model's performance. In this context:

- A higher F-score indicates better performance in correctly identifying legal issues in contracts
- The scale typically ranges from 0 to 1, with 1 being perfect performance
- The close scores between LPOs (0.77) and top LLMs (0.74) demonstrate that advanced AI models are approaching human-level performance in this specific legal task

These benchmarks specifically measured the ability to identify and locate legal issues within contract text, which is a fundamental skill in legal contract review and analysis.

Based on my research, here's a comprehensive disclaimer for your white paper:

DISCLAIMER

Limitations of AI Contract Analysis

This white paper evaluates the performance of Claude LLM models compared to qualified lawyers in contract analysis. The benchmarking data and conclusions presented herein are subject to the following important limitations and assumptions:

Data Quality and Structure

The performance metrics reported assume that input contracts and associated data are properly structured, formatted, and contain all necessary metadata required for accurate analysis. AI models, including Ask Brooklyn / Claude, require high-quality, well-structured data to function optimally and cannot compensate for incomplete, ambiguous, or poorly structured legal documents like a trained legal professional cannot.

Legal Standards for Enforceable Clauses

All performance comparisons assume that the contract clauses being analysed meet established legal standards for clarity, specificity, and enforceability. The benchmarking does not account for performance variations when analysing ambiguous, poorly drafted, or potentially unenforceable contract provisions.

Jurisdictional Limitations

The analysis presumes contracts governed by established legal frameworks within common jurisdictions. Performance may vary when analysing contracts subject to specialised, emerging, or international legal regimes not adequately represented in the training data.

Not Legal Advice

This white paper and the underlying technology evaluations constitute information about AI capabilities and do not constitute legal advice. Organisations implementing AI contract analysis solutions should consult with qualified legal professionals regarding their specific circumstances, jurisdictional requirements, and compliance obligations.

Technological Evolution

The field of AI legal analysis is rapidly evolving. The benchmarking data presented reflects capabilities as of the publication date (March 2025) and may not represent current capabilities or limitations of the referenced technologies.

Human Oversight Requirement

Despite the encouraging performance metrics, AI contract analysis tools, including Claude models, are designed to augment rather than replace human legal professionals. Organisations implementing these technologies should maintain appropriate human oversight and review processes.