Creation of a Dataset for fine-tuning Large Language Models in the Context of Gap-Analyses of Contracts

Maurice Lenßen
Mat-Nr.: 3627550

December 18, 2025

Abstract

Gap-analyses of contracts represent a complex and time-consuming task in legal and procurement workflows, with approximately 60% of businesses still relying entirely on manual review processes. Large language models (LLMs) offer transformative potential to automate and enhance this process through domain-specific fine-tuning. This work presents the creation of a specialized dataset for fine-tuning LLMs in the context of contract gap-analyses, comprising 31 records derived from internal pre-worked analyses and AI-augmented synthetic data reviewed by domain experts. The dataset follows OpenAI's JSONL message format and contains contractual document pairs with corresponding gap-analyses in structured HTML table format. To establish baseline performance metrics, a benchmark was conducted using GPT-4.1 and GPT-5-mini on Microsoft Azure. The evaluation employed a composite similarity score combining cosine similarity of text embeddings (85% weight) and normalized Jaccard similarity (15% weight), alongside mean squared error (MSE) and root mean squared error (RMSE) metrics. Results demonstrate that both models achieve strong baseline performance, with GPT-4.1 attaining a mean score of 80.03% (MSE: 0.041) and GPT-5-mini achieving 78.52% (MSE: 0.047). Performance analysis reveals a negative correlation between input token length and model accuracy, particularly pronounced for inputs exceeding 40,000 tokens. These findings suggest that the created dataset provides a viable foundation for fine-tuning, with literature indicating potential performance improvements of 10% or more through model adaptation. The integration of fine-tuned LLMs into gap-analysis workflows represents a promising advancement in legal technology with significant implications for contractual data analysis efficiency.