ContractEval: Benchmarking LLMs for Clause-Level Legal RiskIdentification in Commercial Contracts

Shuang Liu
Carnegie Mellon University

Zelong Li
Amazon

Ruoyun Ma
ByteDance Inc.

Haiyan Zhao
New Jersey Institute of Technology

Mengnan Du
New Jersey Institute of Technology

Abstract

The potential of large language models (LLMs) in contract legal risk analysis remains underexplored. In response, this paper introduces ContractEval, the first benchmark to thoroughly evaluate whether open-source LLMs could match proprietary LLMs in identifying clause-level legal risks in commercial contracts. Using the Contract Understanding Atticus Dataset (CUAD), we assess 4 proprietary and 15 open-source LLMs. Our results highlight five key findings: (1) Proprietary models outperform open-source models in both correctness and output effectiveness. (2) Larger open-source models generally perform better, though the improvement slows down as models get bigger. (3) Reasoning ("thinking") mode improves output effectiveness but reduces correctness, likely due to over-complicating simpler tasks. (4) Open-source models generate "no related clause" responses more frequently even when relevant clauses are present. (5) Model quantization speed up inference but at the cost of performance drop, showing the tradeoff between efficiency and accuracy. These findings suggest that while most LLMs perform at a level comparable to junior legal assistants, open-source models require targeted fine-tuning to ensure correctness and effectiveness in high-stakes legal settings. ContractEval offers a solid benchmark to guide future development of legal-domain LLMs.

Proceedings of the Natural Legal Language Processing Workshop 2025, pages 291–291
November 8, 2025