

ABSTRAK

Perkembangan teknologi kecerdasan artifisial di bidang hukum mendapatkan traksi yang kuat setelah kehadiran *large language model* (LLM), terutama sejak penggunaan meluas ChatGPT yang mampu memahami substansi teks dan mengolahnya sesuai instruksi. Penelitian ini menyelidiki pemanfaatan LLM dalam rangkaian sistem automasi *review* kontrak berbasis *template*, dengan fokus pada tahapan penentuan *template* yang krusial sebagai basis analisis lanjutan. Implementasi peningkatan ketepatan penentuan *template* dilakukan dengan menggunakan LLM *open source* berukuran kecil, yaitu Qwen2.5-7B-Instruct, yang diposisikan sebagai *semantic evaluator* berbasis *prompt engineering* dan *in-context learning*, tanpa proses vektorisasi/*embedding* tambahan. LLM digunakan untuk memahami substansi teks *draft* kontrak ketika dibandingkan dengan teks *template* kontrak guna menghasilkan skor kemiripan kedua teks. Metodologi penelitian mencakup pengumpulan dan ekstraksi data kontrak publik, praproses dan pembentukan *draft-template*, perancangan *pipeline* penentuan kesamaan substansi pada tingkat *template* dan *pasal*, serta evaluasi performa beberapa metode *text similarity*: leksikal/statistik (Jaccard, TF-IDF), *embedding* semantik (BERT, LaBSE), dan LLM. Hasil penelitian menunjukkan bahwa LLM menghasilkan rekomendasi *template* yang paling deterministik dan selektif, dengan pemisahan yang sangat tegas antara *template* relevan dan tidak relevan. Kemampuan *discriminative power* pada metode komputasi Jaccard, TF-IDF, BERT, LaBSE, dan LLM dapat dilihat pada skor *mean gap* 0.8296, 0.7257, 0.0504, 0.2781, dan 0.9556. Pada metrik *confidence distribution*, skor masing-masing rentang yaitu: 0.9648, 0.9640, 0.1584, 0.5027, dan 1. LLM menunjukkan keunggulan di seluruh metrik yang diujikan dalam menganalisa teks hukum pada proses pengenalan *template* pada sistem automasi *review* kontrak. Penelitian ini dapat menjadi pondasi bagi pengembangan lanjutan sistem automasi *review* kontrak berbasis *template* yang lebih robust terhadap variasi redaksi dan struktur kontrak.

Kata Kunci: Template Kontrak; *Large Language Model*; *Prompt Engineering*; Rekomendasi, *Text Similarity*

ABSTRACT

The advancement of artificial intelligence technology in the legal domain has gained significant traction following the emergence of large language models (LLMs), particularly since the widespread adoption of ChatGPT, which demonstrates the ability to understand textual substance and process it according to explicit instructions. This study investigates the utilization of LLMs within a template-based automated contract review system, with a specific focus on the contract template identification stage, which serves as a critical foundation for subsequent legal analysis. The proposed approach implements an open-source, small-scale LLM, namely Qwen2.5-7B-Instruct, positioned as a semantic evaluator through prompt engineering and in-context learning, without relying on additional vectorization or embedding processes. The LLM is employed to assess the substantive similarity between draft contracts and contract templates by directly reasoning over their textual content and producing similarity scores. The research methodology encompasses the collection and extraction of publicly available contract documents, preprocessing and construction of draft–template pairs, the design of a multi-stage pipeline for substantive similarity assessment at both template and clause levels, and a comparative performance evaluation of several text similarity methods: lexical/statistical approaches (Jaccard, TF-IDF), semantic embedding-based models (BERT, LaBSE), and LLMs. The results demonstrate that the LLM produces the most deterministic and selective template recommendations, exhibiting a clear separation between relevant and non-relevant templates. The discriminative power of the computational methods Jaccard, TF-IDF, BERT, LaBSE, and LLM can be observed from their mean gap scores of 0.8296, 0.7257, 0.0504, 0.2781, and 0.9556, respectively. In the confidence distribution metric, the respective score ranges are 0.9648, 0.9640, 0.1584, 0.5027, and 1. The LLM demonstrates superiority across all evaluated metrics in analyzing legal texts, particularly in the template recognition process within an automated contract review system. The findings of this study can provide as a foundation for the further development of template-based automated contract review systems that are more robust against variations in wording and contract structure.

Keywords: Contract Template; Large Language Model; Prompt Engineering; Recommendation, Text Similarity