11th International Conference on Computational Social Science IC^{2}S^2
July 21-24, 2025, Norrköping, Sweden

Developing and Evaluating a Large Language Model System for Swiss Cases on Choice of Law

Keywords: NLP, LLMs, LLM Evaluation, Court Decisions, Private International Law

Extended Abstract

Recent advances in Large Language Models (LLMs) promise to streamline legal text analysis by summarising complex cases with minimal human oversight [**?**]. However, the reliability of these automated analyses remains uncertain, especially in specialized domains such as Swiss Private International Law (PIL) [**?**]. This study investigates whether LLM-generated case summaries can align with expert legal reasoning when applied to choice-of-law considerations in PIL disputes. Automating time-consuming and analysis steps in this realm could significantly increase efficiency and reduce the costs of legal review, particularly in multilingual jurisdictions.

Drawing from a dataset of 33 court decisions in German, French, and Italian, the project evaluates LLM performance across five categories central to PIL cases: the abstract, relevant facts, choice of law issues, applicable legal provisions, and court's positions. The methodology combines (1) human expert evaluation, in which experienced legal scholars rated the model outputs using predefined accuracy and conciseness criteria, and (2) automated metrics, including a large-scale self-evaluation approach termed "G-Eval" [**?**].

A key technical contribution of this work is a reproducible prompt-chaining pipeline. Rather than generating legal summaries via ad hoc queries in a generic chat interface [**?**], the pipeline enforces deterministic model outputs and tabulates them in a structured format. This design aims to minimize variability and promote transparency in LLM-driven analyses [**?**], ensuring that each reasoning step is clearly documented [include figure with prompt design from Heidelberg paper?]. Preliminary tests of the pipeline revealed that in several categories (e.g., the abstract and choice of law issue as well as the court's position), the LLM results closely paralleled human assessments [**?**, **?**]. In contrast, certain dimensions of factual precision (e.g., summarizing intricate case details) and identifying abstract choice of law themes proved more challenging for the LLM.

In an effort to automate even the evaluation process, we compare human judgments to G-Eval scores. We explore how well an LLM-based tool can reflect expert consensus without relying on additional training data or manual intervention. To quantify these observations, we conduct ANOVA analyses on the evaluation scores. The results indicate that while G-Eval approximates human feedback in some categories, statistically significant discrepancies emerge in others. Specifically, the model struggles to encapsulate nuanced factual details relevant for summarising a case, whereas the model applies overly strict measures when it comes to extracted PIL rules from a case. These inconsistencies underscore the importance of maintaining expert oversight for the most complex reasoning tasks — particularly where subtle statutory interpretations or factual contexts demand deeper human insight.

Looking ahead, the study plans to employ more advanced "reasoning" models, as opposed to basic "instruct" models [**?**], to determine whether they can bridge the observed performance gaps. In parallel, iterative consultation with legal experts will refine the evaluation framework to better capture the domain-specific intricacies of PIL decisions. Although LLMs have shown

remarkable potential for replicating certain aspects of legal reasoning [**?**, **?**, **?**], their statistical nature imposes inherent limitations. This research thus highlights a tension: while automated approaches can enhance efficiency for routine or large-scale document review, reliance on human expertise remains critical for high-stakes legal interpretation.

Overall, the project contributes a scalable and transparent methodology for assessing LLM-generated legal analyses and clarifies key benchmarks for future improvements. By systematically identifying where automation succeeds — and where it falters — this work paves the way for more reliable AI-driven legal tools, fostering interdisciplinary dialogue between legal practitioners, data scientists, and AI researchers.

References

...

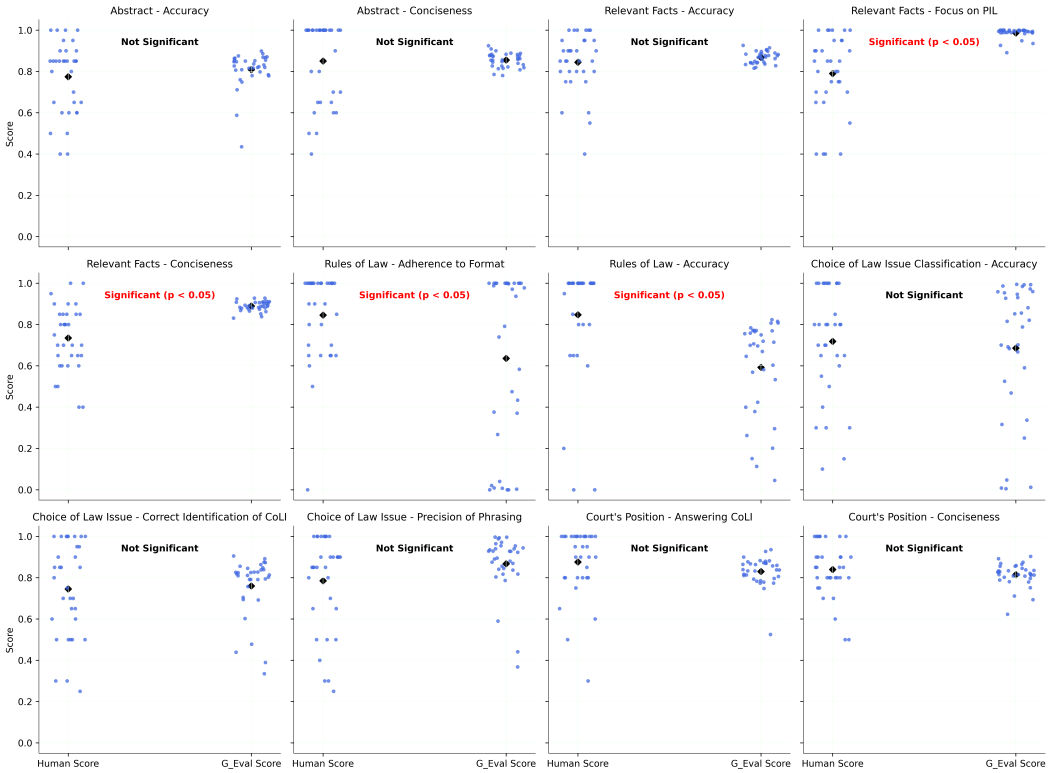

Figure 1: ANOVA between Human and G-Eval Evaluation Scores