11^{th} International Conference on Computational Social Science IC^2S^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

This study explores the development and evaluation of a Large Language Model (LLM) system tailored for analyzing Swiss case law on choice of law in private international law (PIL). Using a dataset of 33 multilingual court decisions (German, French, and Italian), we assess whether LLM-generated case law analyses can match human expert assessments across five key categories: abstract, relevant facts, choice of law issue, applicable legal provisions, and court reasoning.

A structured evaluation process was implemented to assess model performance. Human legal experts first reviewed and rated LLM-generated case summaries based on predefined accuracy and conciseness criteria. Automated metrics, specifically BERTScore and G-Eval, were then used to compare these evaluations. BERTScore measures textual similarity between model-generated and human-authored responses, while G-Eval employs an LLM-based self-assessment approach that does not require external reference data. If proven reliable, G-Eval could facilitate scalable and cost-effective evaluations of legal analyses.

Leveraging OpenAI's GPT-4o model, the study investigates the extent to which numerical vectorization techniques can capture the intricate reasoning present in legal texts. PIL cases pose distinct challenges as choice of law considerations often appear as secondary issues within broader legal disputes. An effective LLM-based analysis must be capable of distinguishing these nuanced elements, identifying applicable legal provisions, and synthesizing judicial reasoning.

The statistical analysis (ANOVA) of evaluation metrics reveals mixed findings. In some areas, such as Relevant Facts - Accuracy, Court's Position - Conciseness, and Choice of Law Issue - Correct Identification of CoLI, no significant difference was found between human and G-Eval scores (p ¿ 0.05), suggesting that automated evaluation can approximate human judgments in certain cases. However, statistically significant discrepancies (p ¡ 0.05) emerged in key categories such as Rules of Law - Adherence to Format, Relevant Facts - Conciseness, Relevant Facts - Focus on PIL, and Rules of Law - Accuracy. These differences indicate that while G-Eval is promising, it still struggles with aspects of structural and domain-specific evaluation.

Figure 1 presents these results, illustrating the areas where automated evaluation aligns well with human judgment and where inconsistencies persist. While G-Eval shows potential in automating aspects of legal text evaluation, the findings highlight the continued necessity of expert oversight, particularly for complex reasoning tasks. Future research will explore reasoning-enhanced LLMs and refine evaluation frameworks to further bridge the gap between machine-generated and human-authored legal analyses.

If G-Eval proves reliable, it could enable large-scale automated assessments of LLM-generated legal analyses, reducing the need for costly human review. However, it does not. Thus, this study proves that when it comes to evaluation, humans are still needed. But how about "reasoning" models? This study is also going to generate answers using reasoning models and go through the same evaluation process.

References

...

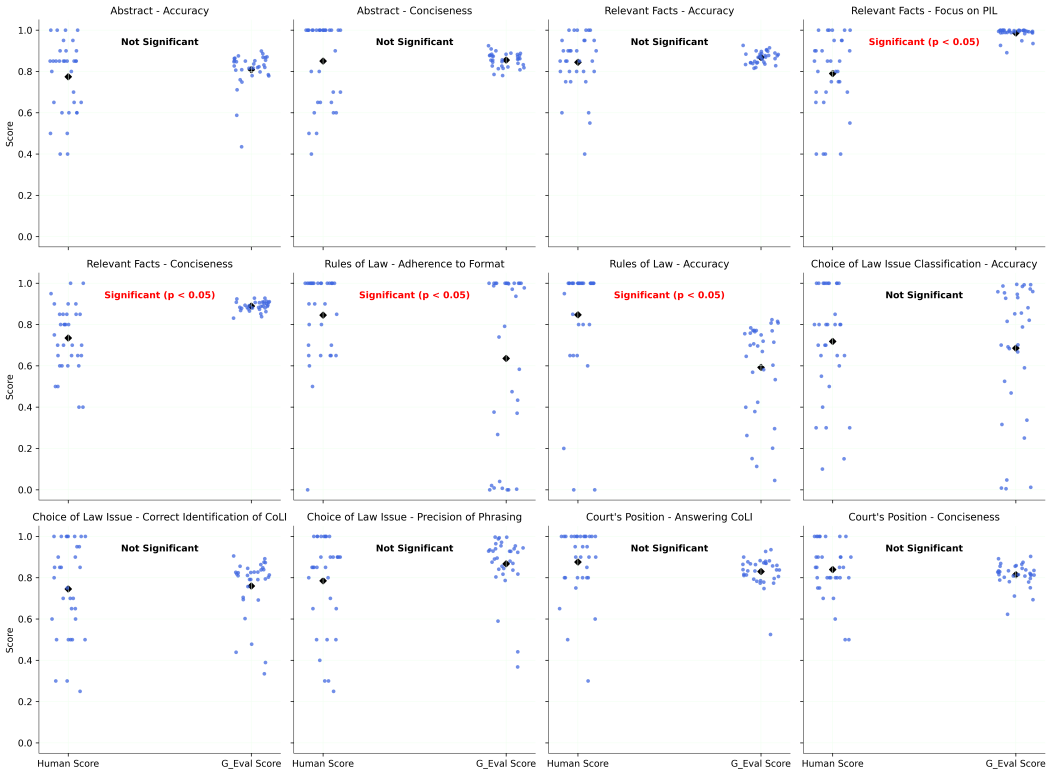

Figure 1: ANOVA between Human and G-Eval Evaluation Scores