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

Building an LLM Case Analyser for Private International Law: Evaluation via G-Eval and Expert Review

Keywords: Large Language Models, Private International Law, Swiss Case Law, Automated Legal Analysis, Evaluation Metrics

Extended Abstract

Recent advances in Large Language Models (LLMs) promise to streamline legal text analysis by summarising complex cases with minimal human oversight (Pereira et al. 2024). However, the reliability of these automated analyses remains uncertain (Deroy, K. Ghosh, and S. Ghosh 2024), especially in specialised domains such as Private International Law (PIL). This study investigates whether LLM-generated case summaries align with expert legal reasoning when applied to choice-of-law considerations in PIL disputes. Automating time-consuming analysis steps in this realm increase efficiency and reduce the costs of legal review, particularly in multilingual jurisdictions.

Jurists rely on case law briefs to quickly identify the most critical information in a case. To address this need, we develop an automated system using a dataset of 33 court decisions in German, French, and Italian. This system generates case law briefs and evaluates the LLM outputs in five categories central to PIL cases: abstracts, relevant facts, PIL rules, choice of law issues, and court's positions. The evaluation methodology combines two main approaches. First, experienced legal scholars rate the LLM-generated briefs against predefined accuracy and conciseness criteria. Second, we utilise automated metrics, including a large-scale self-evaluation technique called "G-Eval" (Liu et al. 2023), to benchmark the performance without relying on external reference data.

A key technical contribution of this study is a reproducible prompt-chaining pipeline. Rather than generating legal summaries via ad hoc queries in a generic LLM chat interface (Qin et al. 2023), the pipeline promotes deterministic model outputs (Atil et al. 2024) and returns text in a structured format. This pipeline design aims to minimise variability and promote transparency in LLM-driven analyses, ensuring that each reasoning step is clearly documented. Preliminary tests of the pipeline reveal that in several categories, e.g., the abstract and choice of law issue as well as the court's position, the LLM's outputs come close to human assessments. In contrast, factual precision, e.g., summarising intricate case details, and identifying abstract choice of law themes, are more challenging for the LLM.

In an effort to automate 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 the G-Eval scores approximate human feedback in some categories, statistically significant discrepancies emerge in others (see Figure 1). 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" LLMs, as opposed to basic "instruct" LLMs (Alammar and Grootendorst 2024, p. 22), 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 (Anonymous 2025; Spaić and Jovanović 2024), 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 study contributes a scalable and transparent methodology for assessing LLM-generated legal analyses. Furthermore, it clarifies key benchmarks for future improvements. By systematically identifying where automation succeeds and 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

Alammar, Jay and Maarten Grootendorst (2024). *Hands-on large language models: language understanding and generation*. en. 1st edition. Beijing Boston Farnham: O'Reilly. ISBN: 978-1-09-815096-9 978-1-09-815093-8.

Anonymous (2025). *Conference presentation of the same research project with a focus on legal subjects*.

Atil, Berk et al. (2024). *LLM Stability: A detailed analysis with some surprises*. arXiv: 2408.04667 [cs.CL]. URL: https://arxiv.org/abs/2408.04667.

Deroy, Aniket, Kripabandhu Ghosh, and Saptarshi Ghosh (July 2024). "Applicability of large language models and generative models for legal case judgement summarization". en. In: *Artificial Intelligence and Law*. ISSN: 0924-8463, 1572-8382. DOI: 10.1007/s10506-024-09411-z. URL: https://link.springer.com/10.1007/s10506-024-09411-z (visited on 09/11/2024).

Liu, Yang et al. (2023). *G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment*. arXiv: 2303.16634 [cs.CL]. URL: https://arxiv.org/abs/2303.16634.

Pereira, Jayr et al. (Mar. 2024). "INACIA: Integrating Large Language Models in Brazilian Audit Courts: Opportunities and Challenges". en. In: *Digital Government: Research and Practice*, p. 3652951. ISSN: 2691-199X, 2639-0175. DOI: 10.1145/3652951. URL: https://dl.acm.org/doi/10.1145/3652951 (visited on 10/24/2024).

Qin, Chengwei et al. (Nov. 2023). *Is ChatGPT a General-Purpose Natural Language Processing Task Solver?* en. arXiv:2302.06476 [cs]. URL: http://arxiv.org/abs/2302.06476 (visited on 09/14/2024).

Spaić, Bojan and Miodrag Jovanović (2024). "Artificial Reason and Artificial Intelligence: the Legal Reasoning Capabilities of GPT-4". In: *The Annals of the Faculty of Law in Belgrade* 72.3. DOI: 10.51204/Anali_PFBU_24302A. URL: https://doi.org/10.51204/Anali%5C_PFBU%5C_24302A.

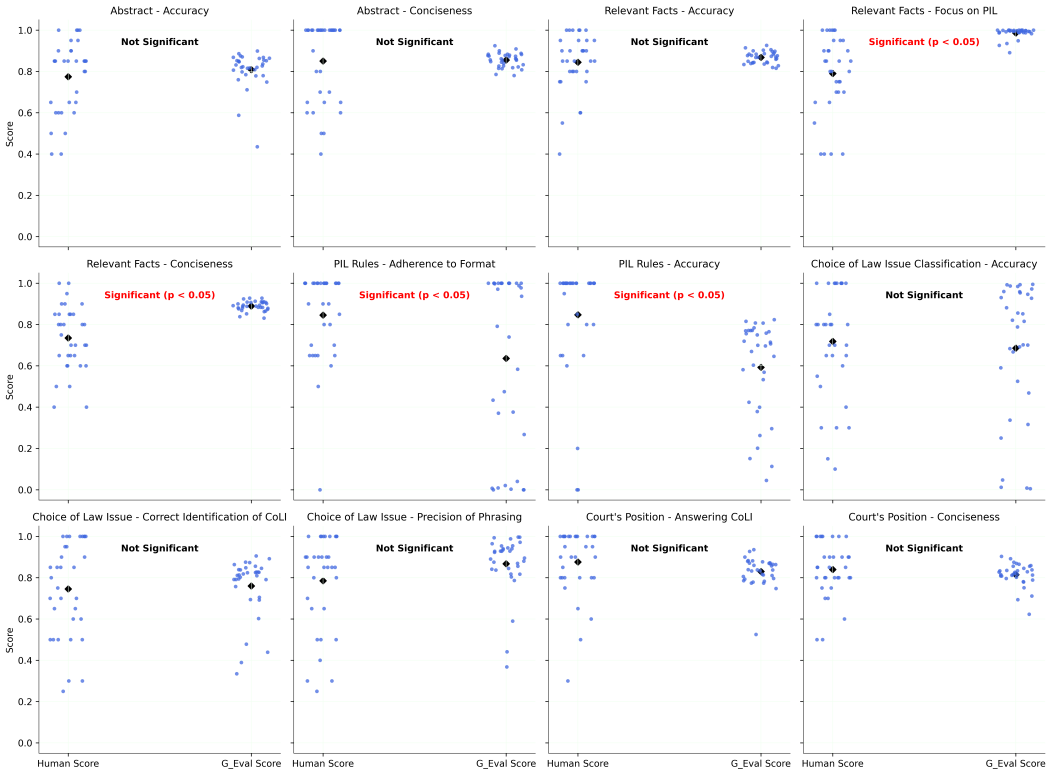

Figure 1: ANOVA between Human and G-Eval Evaluation Scores. Statistical insignificance denotes an effective automated evaluation.