The (Currently) Uneven Intelligence of AI in Comparative Constitutional Law

From Years of Research to a Single Prompt?

Adam Chilton, Arthur Langlois, Jan-Ulrich Dittmer & Mila Versteeg

14.10.2026 doi: 10.59704/a3baf260ef374567

Artificial intelligence (AI) has the potential to accelerate and improve the quality of research. This potential has inspired numerous projects exploring how AI could help scholarship on constitutional law. For example, there is research using AI to interpret constitutions,[1] apply constitutional principles,[2] and predict court rulings.[3] Even the question of whether AI could improve the process of constitution-making has been studied.[4]

AI's potential is particularly significant for comparative constitutional research. Systematically comparing legal systems requires locating sources from different jurisdictions, which are often difficult to access and only available in local languages. The creation of comparative legal datasets has thus involved considerable personnel, time, and financial resources. AI can overcome many of these obstacles. AI models can search for and translate foreign-language sources, filter them, and convert the filtered information into a dataset. In doing so, AI offers the prospect of datasets — which previously required years of research and the efforts of numerous experts and research assistants to compile — being just a prompt away.

However, whether AI models can generate comparative law datasets that are qualitatively comparable to human-coded datasets has not yet been addressed (although Guha et al provide an example for a broader legal benchmarking exercise).[5] We therefore conducted our own benchmarking study.[6] We found evidence suggesting that recent AI models may be able to code simple factual data, but they are likely unreliable substitutes for expert human coders when conceptual or interpretive judgment is required.

The Benchmarking Project: Seven Coding Tasks, Four Datasets, Two LLMs

In our project, we tested the ability of two large language models (LLMs) to code different types of information relevant to comparative constitutional law. We specifically assessed whether LLMs can reproduce information from four existing comparative constitutional law datasets, all of which were compiled through human coding and which served as a 'ground truth' for our exercise. From these datasets, we asked the LLMs to code seven variables, which can be thought of as seven tasks relevant to the creation of comparative constitutional law datasets. Each of these tasks is either a test of fact-finding or reasoning abilities.

First, from the Executive Term Limit Evasion Dataset,[7] we drew information on whether presidents in 106 presidential systems have evaded constitutionally mandated term limits between 2000 and 2019. The

[1] Andrew Coan and Harry Surden, Artificial Intelligence and Constitutional Interpretation, University of Colorado Law Review 96 (2025).

[2] Camilla Bignotti and Carolina Camassa, Legal Minds, Algorithmic Decisions: How LLMs Apply Constitutional Principles in Complex Scenarios, Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society 7(1) (2024), https://ojs.aaai.org/index.php/AIES/article/view/31623.

[3] Nikolaos Aletras et al., Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective, PeerJ Computer Science 2 (2016), DOI: 10.7717/peerj-cs.93; Daniel Katz et al., A General Approach for Predicting the Behavior of the Supreme Court of the United States, PLoS ONE 12(4) (2017), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0174698; Vijit Malik et al., ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation, 28.05.2021, https://arxiv.org/abs/2105.13562, last accessed 24.08.2026; Mehmet Fatih Sert, Using Artificial Intelligence to Predict Decisions of the Turkish Constitutional Court, Social Science Computer Review 40(6) (2021).

dataset includes one variable on whether presidents attempt to overstay past the end of their term. The second variable captures what strategies executives use for doing so, such as amending the constitution, replacing it with a new one, or asking the court to interpret term limits out of the constitution. While determining whether a president overstayed is a fact-finding task, assessing the strategy used to do so requires reasoning.

Second, from the Comparative Constitution Making Dataset,[8] we drew information on the process by which constitutions were adopted in democracies between 1900 and 2015. This dataset also includes fact-finding variables such as (1) whether citizens were involved in the formulation, discussion, or submission of reform proposals, and (2) whether citizens participated in popular referendums at any stage in the process. The dataset's most important variable (3) captures whether constitutions reflect 'plural elite agreement,' that is, whether all relevant elite factions in society were involved and represented during the process. The recreation of this variable requires reasoning.

Third, from the Characteristics of National Constitutions Dataset[9], we drew on its coding of 43 first-generation constitutional rights included in national constitutions (that is, rights related to individual liberties and protections against state interference such as due process, fair trial, and political rights). By asking the LLMs to evaluate whether each of these rights was present in a given constitution, we tested their fact-finding abilities.

Finally, from the Small-c Constitutional Rights Dataset,[10] we drew information on whether a country's rights protections derive primarily from the small-c constitution (i.e., judicial decisions, treaties, conventions, quasi-constitutional laws), the large-C constitution (i.e., the written text), or something in between. Since answering this question requires deeper knowledge of a constitutional system, the recreation of this variable can be understood as a reasoning task.

The specifics of our approach are documented in the paper.[11] But the core idea is that we asked two LLMs—ChatGPT-5 and Grok-3, the state-of-the-art models by OpenAI and xAI in fall 2025—to recode these seven variables without providing the LLMs with underlying constitutional texts. Instead, we asked the LLMs to use their existing knowledge base. We did so because (1) there is already research suggesting that LLMs are effective at coding data that are provided to them, and (2) a considerable part of the burden of producing comparative law datasets is assembling the corpus of documents from different countries. Nonetheless, we provided basic guidance such as country names, presidents and term dates, relevant constitution-making processes, and the first-generation rights to be coded.

Each LLM was prompted ten times for each of the coding exercises. (We made an exception for the Characteristics of National Constitutions Dataset, which, due to disproportionate processing time, we prompted Grok-3 only one time for this task.) For each dataset, the final prediction for a given variable was determined by taking the mode responses across all runs. In order to simulate the typical conditions under which legal researchers interact with LLMs, we used the default temperature settings (1.0 for ChatGPT-5 and 0.7 for Grok-3) and stayed with the default reasoning mode of ChatGPT-5 in order to ensure feasible processing times and benchmark comparability with Grok-3.

[4] Richard Albert and Kevin Frazier, Should AI Write Your Constitution?, 15.07.2025, SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5351275, last accessed 24.08.2026.

[5] Neha Guha et al., LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models, 2023, https://arxiv.org/abs/2308.11462, last accessed 24.08.2026.

[6] Adam Chilton et al., What Does Artificial Intelligence Know About Comparative Constitutional Law?, Constitutional Studies 11(2) (2025).

[7] Mila Versteeg et al., The Law and Politics of Presidential Term Limit Evasion, Columbia Law Review 120 (2020).

[8] Gabriel Negretto and Mariano Sánchez-Talanquer, Constitutional Origins and Liberal Democracy: A Global Analysis, 1900-2015, American Political Science Review 115(2) (2021).

Four Lessons About What Current LLMs Can and Cannot Do

(1) Current LLMs Do Not Consistently Reach Human-Coder Standards

Our findings cast doubt on whether publicly available LLMs can already serve as a substitute for human coders in producing comparative constitutional law datasets. A commonly used benchmark for human coding is 90 percent accuracy.[12] The models reach this threshold only on selected tasks. Depending on the variable, ChatGPT-5 achieved accuracies between approximately 50 and 94 percent, while Grok-3 ranged from approximately 48 to 90 percent. Overall, their performance therefore remained uneven and frequently fell below the level expected from human coding.

(2) Performance Depends on the Type of Coding Judgement

The results also revealed a clear relationship between task complexity and model performance. Both models performed best when the coding exercise concerned a concrete and verifiable event, such as whether a president attempted to evade a constitutional term limit or whether citizens voted to approve a new constitution by referendum. On some of these factual variables, the models reached or exceeded the 90 percent benchmark and displayed their highest accuracies.

Their performance deteriorated, however, once coding required the application of an abstract concept or legal analysis. For example, when asked to determine whether there was 'plural elite approval' during constitution-making—a concept that depends on diffuse and context-dependent information about negotiations—the models, particularly Grok-3, performed significantly worse (80 and 66 percent accuracy). A similar pattern emerged when the models must perform interpretive legal analysis, such as identifying the main legal basis for constitutional rights protection in a given country (78 and 76 percent accuracy). This suggests that current LLMs are more reliable when classifying concrete constitutional facts, but they are less reliable with more interpretive questions that demand contextual reasoning.

(3) Model Errors Follow Recurring Patterns

The errors we observed are not distributed randomly. Instead, both models exhibited two recurring tendencies. The first pattern of systematic bias is a hallucination bias. For binary variables, the models were more likely to code an event or constitutional feature as present when it was absent than to overlook an event that occurred. In other words, the LLMs produced more false positives than false negatives. For instance, the LLMs too frequently asserted that a president violated term limits, that citizens approved a constitution, or that a new constitution was adopted, even in the absence of evidence.

The second pattern is a majority class bias. That is, when choosing among several possible classifications, the models disproportionately selected the most common or salient category. For example, the LLMs were more likely to assume that the main source of constitutional rights was the large-C constitution rather than judicial interpretation, convention, or treaty. While the first tendency risks overstating the frequency of constitutional events, the latter reduces variation by drawing predictions to the modal category. Because neither error is random, both may distort relationships between constitutional design and political outcomes.

(4) Country Characteristics Do Not Produce a Performance Divide

We also examined whether model performance varied systematically across different national contexts and could not find systematic differences across the characteristics we explored. Instead, we found that the two models performed roughly equally

[9] Zachary Elkins and Tom Ginsburg, Characteristics of National Constitutions, Version 5.0, 2025, Comparative Constitutions Project, https://comparativeconstitutionsproject.org/download-data/, last accessed 24.08.2026.
[10] Adam Chilton and Mila Versteeg, Small-c Constitutional Rights, International Journal of Constitutional Law 20(1) (2022).
[11] Adam Chilton et al., What Does Artificial Intelligence Know About Comparative Constitutional Law?, Constitutional Studies 11(2) (2025).
[12] David Roaché, Intercoder Reliability Techniques: Percent Agreement, The SAGE Encyclopedia of Communication Research Methods, 2017, DOI: 10.4135/9781483381411.n260.

well (1) across English- and non-English-speaking countries, (2) across G20 and non-G20 economies, and (3) across common-law and civil-law countries. Small variations existed across these sets of countries, but they largely fell within the range of ordinary human-coder disagreement and show no systematic pattern. This suggests that, despite variation in the availability of training data across contexts, model performance was broadly stable.

Conclusion: Human Expertise Still Matters

Our results suggest that LLMs' usefulness for comparative constitutional data collection depends on what researchers ask them to code. They may be able to produce datasets that simply code constitutional facts, but current versions of LLMs do not perform well when asked to create datasets that require interpreting abstract constitutional concepts or reasoning about law.

However, since the models make errors in systematic ways, researchers should always be cautious about treating LLM-generated data as interchangeable with expert-coded data. Nonetheless, model capabilities are constantly evolving, and these conclusions may not remain valid indefinitely (or even in the near term). For now, however, human expertise remains essential to constructing datasets for comparative constitutional law research, especially when coding extends beyond classifying relatively straightforward facts.

ABOUT THE AUTHORS

Adam Chilton
Adam Chilton is the Dean and the Howard G. Krane Professor of Law at the University of Chicago Law School.
Arthur Langlois
Arthur Langlois is a student at the University of Chicago.
Jan-Ulrich Dittmer
Jan-Ulrich Dittmer is a doctoral candidate at Humboldt University Berlin.
Mila Versteeg
Mila Versteeg is the Walter L. Brown Professor of Law and the director of the Centre for Public Law and Political Economy at the University of Virginia.

SUGGESTED CITATION
Chilton, Adam; Langlois, Arthur; Dittmer, Jan-Ulrich; Versteeg, Mila: The (Currently) Uneven Intelligence of AI in Comparative Constitutional Law: From Years of Research to a Single Prompt?, Verfassungsblog, 14.10.2026, doi: 10.59704/a3baf260ef374567.

Licensed under CC BY-SA 4.0.

Langzeitarchiviert bei IntRechtDok unter: doi.org/10.17176/20261014-143433-0.