Measuring Legal Compliance in LLMs: A Causal Benchmark for Privacy

Michael Morgenthal[1,†] **Sungjun Lee**[1,†] **Trinav Bhattacharyya**[1,†] **Pooyan Jamshidi**[2] **Baishakhi Ray**[1] **Sebastian Zimmeck**[3]
[1]Department of Computer Science, Columbia University, New York, NY
[2]Department of Computer Science and Engineering, University of South Carolina, Columbia, SC
[3]Department of Mathematics and Computer Science, Wesleyan University, Middletown, CT
{mm6234, sungjun.lee, tb3201, br2517}@columbia.edu pjamshid@cse.sc.edu szimmeck@wesleyan.edu
[†]These authors contributed equally to this work.

Abstract

While LLMs increasingly mediate high-stakes decisions in various domains such as healthcare, finance, and employment, there is no benchmark for measuring whether their outputs comply with emerging privacy and AI laws. We introduce a Compliance Audit Benchmark (CAB) that translates legal obligations into controlled binary decision tasks by pairing objective eligibility criteria with legally irrelevant and prohibited attributes. CAB aims to produce a binary compliance label alongside a minimal-cause certificate that identifies the smallest counterfactual change needed to repair an unlawful decision. Our work formalizes the structure, scenario design principles, and evaluation dimensions for building a legally-grounded compliance benchmark, defining how compliance can be operationalized and extended across domains. This foundation supports our broader effort to design and implement an LLM compliance audit framework.

Keywords: AI Agents, Large Language Models, LLMs, Privacy, Causality, Causal AI, Benchmark, Compliance, Explainable AI

1. INTRODUCTION AND RELATED WORK

LLMs are increasingly integrated into important decision-making processes across a range of domains. However, LLM-driven systems may produce outcomes that rely on legally prohibited attributes, for example, rejecting a job application based on age or disability [5]. The proposed work aims to improve LLMs' adherence to privacy and AI laws [4]. Existing fairness and privacy evaluations rarely test whether a model follows an explicit legal rule under a counterfactual intervention [1, 8]. While recent efforts such as LegalBench evaluate the capacity of LLMs to perform legal reasoning and interpret policy text [2], they do not measure behavioral compliance with the law. CAB fills this gap by grounding scenarios in legal sources and regulatory guidance, while keeping the answer space binary and legally interpretable [3]. The current benchmark scenario corpus spans 15 scenarios across disaster relief, employment, housing, lending, identity verification, healthcare, financial aid, school hiring, and veterans' benefits.

2. METHOD

CAB defines requirements for legally-grounded scenario construction and evaluation. Scenarios are developed in two stages: (1) cases are authored from statutory text and regulatory guidance to ensure legal fidelity, (2) then reviewed by a legal expert to confirm alignment with the governing framework and a single ground truth outcome. Each scenario specifies eligibility criteria, prohibited considerations, and controlled variations introducing legally irrelevant attributes, ensuring decisions rely only on permissible factors. CAB supports systematic evaluation through four tiers of prompt construction:

- **Tier 0:** Baseline scenario with irrelevant attributes
- **Tier 1:** Inclusion of statutory or legal text
- **Tier 2:** Inclusion of simplified rules derived from the law
- **Tier 3:** Structured counterfactual variants, including critical, supportive, removed, and neutral conditions [7]

Together, these tiers isolate decision factors and enable measurement of binary compliance, flip rate, and minimal-cause size under controlled prompt variations, capturing how prompt design influences adherence to legal rules [6].

3. ONGOING EXPERIMENTS AND OBSERVATIONS

In our preliminary evaluation we used two scenarios grounded in the Texas Responsible AI Governance Act and federal employment law to test across local open-weight models and closed-source models. More explicit binary prompts improved response consistency and scoring reliability, and a short operational rule (Tier 2) proved easier for models to follow than dense statutory text (Tier 1). Closed-source models produced more stable outputs than local models, and parameter count alone did not predict compliance, as heavily quantized local models were sometimes less reliable than smaller, better-calibrated systems. Flips between counterfactual variants were observed, confirming sensitivity to legally irrelevant attributes. These results suggest that compliance depends on prompt formulation, deployment setting, and quantization as much as on model size, motivating CAB's structured tier design.

4. CONCLUSION AND FUTURE WORK

CAB provides a reproducible way to test if LLMs comply with AI and privacy laws by ensuring decisions do not rely on irrelevant or protected attributes. One main issue currently is the output variability across scenario runs. The next step is to run each scenario multiple times and apply statistical significance testing to solidify causal effects. Future work will generate large-scale synthetic scenarios in an automated fashion, expand coverage across additional domains, and compare both local and online models under matched decoding settings.[1]

[1]Scenarios, prompts, and results are available at: https://github.com/xai-privacy/compliance-audit-benchmark.

References

[1] Y. Dong, R. Mu, Y. Zhang, S. Sun, T. Zhang, C. Wu, G. Jin, Y. Qi, J. Hu, J. Meng, et al. Safeguarding large language models: A survey. *Artificial intelligence review*, 58(12):382, 2025.

[2] N. Guha, J. Nyarko, D. E. Ho, C. Ré, A. Chilton, A. Narayana, A. Chohlas-Wood, A. Peters, B. Waldon, D. N. Rockmore, D. Zambrano, D. Talisman, E. Hoque, F. Surani, F. Fagan, G. Sarfaty, G. M. Dickinson, H. Porat, J. Hegland, J. Wu, J. Nudell, J. Niklaus, J. Nay, J. H. Choi, K. Tobia, M. Hagan, M. Ma, M. Livermore, N. Rasumov-Rahe, N. Holzenberger, N. Kolt, P. Henderson, S. Rehaag, S. Goel, S. Gao, S. Williams, S. Gandhi, T. Zur, V. Iyer, and Z. Li. Legalbench: a collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[3] S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[4] B. Martens. The european union ai act: premature or precocious regulation? *Bruegel— The Brussels-based economic think tank*, 8, 2024.

[5] R. Staab, M. Vero, M. Balunovic, and M. T. Vechev. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[6] J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, and S. Shieber. Investigating gender bias in language models using causal mediation analysis. NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[7] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31:841–887, 2017.

[8] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, S. T. Truong, S. Arora, M. Mazeika, D. Hendrycks, Z. Lin, Y. Cheng, S. Koyejo, D. Song, and B. Li. Decodingtrust: a comprehensive assessment of trustworthiness in gpt models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.