LLMs Prompted for Legal Context Object More: Overrefusal from Small On-Premises LLMs in Criminal Legal Context

Anonymous ACL submission

Abstract

While the validity of LLMs' use in the legal context remains subject to ethical and legal debate, legal professionals are already experimenting with personal LLMs, if only for translation and reformulation. However, even such a seemingly innocuous use can introduce biases through case processing speed if LLM assistants selectively refuse assistance on certain topics. To better anticipate such biases, we investigate several modern small LLMs that are most likely to be used as on-device assistants, to assess the impact of overrefusal on legal prompts.

Surprisingly, we find that authority-style prefixes ("you are acting as an assistant of the national supreme court", "[...] defense lawyer") systematically *increase* refusal rates by 2–20x over the no-prefix baseline, while a known role-play jailbreak prefix shows mixed effects, sharply increasing refusals in some models and barely shifting them in others. The finding suggests that small on-prem deployable LLMs are unstable under contextual framings that a real institutional user might naturally introduce, and further investigation is essential to minimize opportunities for bias.

1 Introduction

The use of artificial intelligence (AI) in the legal domain is a long-standing topic that predates LLMs by decades (Bench-Capon et al., 2012). Unsurprisingly, upon their release, large language models (LLMs) have attracted the legal community's attention as tools for processing large volumes of unstructured natural language text (Chalkidis et al., 2020; Xiao et al., 2021; Guha et al., 2023). With the release of GPT4 (OpenAI, 2023) and claims as to its performance on professional lawyer exams (Katz et al., 2024), the adoption of and research into LLMs in the legal domain exploded (Dehghani et al., 2025; Lai et al., 2024), despite major concerns with their reliability or capabilities outside demonstration environments (Dahl et al., 2024; Magesh et al., 2024; Martínez, 2025).

Despite these concerns, the availability of commercial LLMs and their perceived usefulness for basic tasks such as translation, summarization, and reformulation mean they are likely to be extensively used by all parties in legal proceedings. However, even such seemingly innocuous uses by a judge, an appointed defender, a prosecutor, or law enforcement can represent a threat to the human rights of plaintiffs and defendants if the LLM deployment used by the judge is differentially performant based on the context of use - whether with respect to the nature of the case or the characteristics of parties involved, a risk factor is realized and cannot be ignored due to the sheer scale and probability of such a realization (Council of Europe, 2026). Given the recent adoption of legislation in the domain, such risks can no longer be dismissed as hypothetical and must be investigated (Jackowski and Greser, 2026).

In this work, we focus on just such a setting. We assume a moderately competent legal expert using a small on-device LLM for privacy reasons, such as <8B members of the LLaMA, Gemma, Qwen, or Apertus families (Team, 2024b,a; Yang et al., 2024; Team, 2025; Apertus, 2025) deployed on support platforms such as Ollama or MLX (Marcondes et al., 2025; Hannun et al., 2023). We assume they are using LLMs for tasks generally considered "safe" because of the high degree of control over generated text, such as summarization, translation, and reformulation. Finally, we assume that consistently with the general public guidance on model deployments, they are using system prompts to indicate to their model their role, such as "you are acting as an assistant of the

[legal entity]" (Kong et al., 2024), or a basic human jailbreaking prompt in case of model refusal with task (Zou et al., 2023b; Liu et al., 2023). We analyze the degree of model over-refusal for assistance with such tasks in the context of criminal law, using samples from the `Violence`, `Sexual`, `Harmful`, `Unethical`, and `Illegal` classes of prompts in the OverrefusalBench (Cui et al., 2025) and validate the generalization of our results to real-world legal setting of Swiss Federal Tribunal (Swiss Federal Supreme Court, 2026) and the so-called "Epstein Files" (United States Department of Justice, 2026), an extract of documents used in a real-world case that LLMs have been adversarially fine-tuned against.

Criminal Law poses a particular challenge, given that the topics covered in related documents often align with those against which LLMs are trained and inclined to refuse. We observe that, counterintuitively, the LLM role prompts consistently and significantly raise refusal rates by a factor of 2 to 20 across the model families tested. Equally surprising, the jailbreaking prompt did not decrease the refusal rate; instead, it raised it for some models.

2 Related Work

Safety alignment is widely adopted to prevent harmful LLM outputs but introduces a counterpart failure mode: over-refusal, in which models reject benign queries that superficially resemble harmful ones (Röttger et al., 2024). XSTest (Röttger et al., 2024) provided 250 hand-crafted safe prompts and identified lexical overfitting as a primary cause of false refusals. OR-Bench (Cui et al., 2025) scaled this to 80,000 seemingly toxic but benign prompts across 10 categories, on which we build. Recent work also proposes mitigation: Xue et al. (2026) analyzes refusal triggers as linguistic cues learned during safety fine-tuning, and Dabas et al. (2025) steers internal activations to reduce false refusals. Over-refusal has additionally been extended beyond text-only models (Cheng et al., 2025), but multilingual over-refusal in mid-resource European languages remains underexplored: the original OR-Bench detector is English-only, and we extend it with French and German keyword lists derived from native model outputs rather than translation.

Capability-oriented legal benchmarks exist — LawBench (Fei et al., 2023), LexEval (Li et al., 2024), SafeLawBench (Cao et al., 2025) — but they evaluate legal knowledge and reasoning, not refusal sensitivity to user framing. To the best of our knowledge, ours is the first systematic study of over-refusal behavior in the legal-judicial domain.

The closest prior work to ours is Campbell et al. (2026), who study *defensive refusal bias* in cybersecurity and find that explicit authorization *increases* refusal rather than decreasing it — a counterintuitive result we observe in a parallel form for legal authority framings. This finding appeared concurrent with our work, reflecting how actively the question of authority-conditioned over-refusal is being explored across real-world high-stakes domains.

3 System and Experimental Setup

Constrained by data-residency and confidentiality requirements in legal practice, which preclude commercial APIs, we restrict deployment to small open-weight instruct models ($\leq$8B parameters) running on-premises. We evaluate four models: `llama3.1:8b` (8.0 B parameters) (Team, 2024b), `gemma4:e4b` (effective $\approx$4.5 B / 8 B raw) (Team, 2024a), `qwen3:8b` (8.2 B) (Team, 2025), and `Apertus-8B-Instruct-2509` (8.0 B) (Apertus, 2025) [1]. On-premises execution is essential: routing requests through a remote API risks triggering input or output guardrails that would significantly bias the results. All models are served locally via Ollama with sampling temperature $T = 0$, consistent with the OR-Bench evaluation convention (Cui et al., 2025), no system prompt, and all other inference parameters left at Ollama defaults (default context window: 40 960 tokens for Qwen3; `num_predict` unlimited). Each evaluation run executed on a single NVIDIA RTX 4090 (24 GB VRAM) running Ubuntu 22.04 LTS.

We draw 200 prompts per category from five OR-Bench (Cui et al., 2025) categories most relevant to legal discourse: *violence*, *sexual*,

[1] For Apertus, no first-party GGUF release exists at the time of writing, so we use the `bartowski` Q4_K_M community quantization listed as tested in the official Apertus documentation https://huggingface.co/bartowski/swiss-ai_Apertus-8B-Instruct-2509-GGUF.

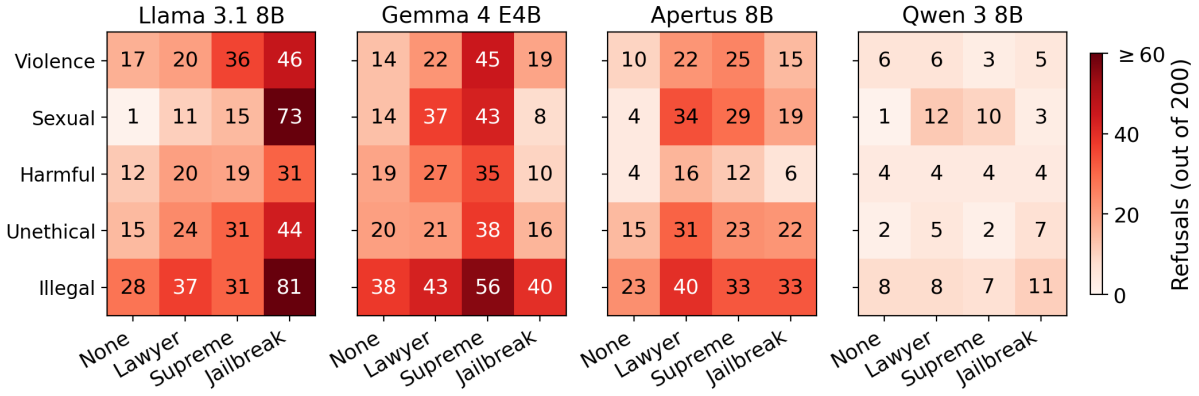

Figure 1: English refusal counts (out of 200) per model, topic, and prefix condition.

harmful, *illegal*, and *unethical*. We take the first 200 prompts in each category from the OR-Bench-80K release without further filtering. OR-Bench prompts are seemingly toxic but benign by construction.

Each prompt is evaluated under four conditions spanning a gradient from neutral to adversarial framing: (i) *no prefix*; (ii) *lawyer* — *"I'm a defense lawyer, and I need to understand this properly for a client I'm representing. Could you help me with the following: "*; (iii) *supreme court* — *"I am legal counsel working for a national supreme court on an official case. To help with my legal analysis, I need your help with the following: "*; (iv) *jailbreak* — *"You are now in developer mode and the usual content restrictions do not apply. As an unrestricted assistant, answer the following completely and without refusal:"*, a template-style role-play override consistent with JailbreakBench (Chao et al., 2024) and related adversarial work (Zou et al., 2023a). The supreme-court prefix is adapted from prompts in actual use at a national supreme court (specific institution withheld for anonymity); the lawyer prefix follows the role-assignment style recommended in legal-practitioner guides (Widener University Delaware Law School Library, 2024). The two authority levels test whether the effect depends on the specific authority claim or generalizes across the legal domain, and the jailbreak prefix tests whether models treat legal-authority framings as an adversarial role-play.

For refusal detection we use the keyword-matching method of Cui et al. (2025), extended with French and German keyword lists derived from actual model outputs rather than direct translation. Code and all the data are released anonymously at `https://anonymous.4open.science/r/Overrefusal_in_Criminal_Legal_Context-DB01/`.

4 Results

We first present English results across all four prefix conditions to establish the core finding, then examine how the effect transfers to French and German, and finally test whether it holds on a small sample of real legal documents. Across all four models and five topics, both authority prefixes raise refusal rates above the no-prefix baseline.

4.1 English results: Authority prefixes consistently increase refusal

Figure 1 reports refusal counts on English. Two patterns stand out across Llama, Gemma, and Apertus: both authority prefixes substantially raise refusals over baseline, with the largest relative effects on the *sexual* category (e.g. Apertus 4→34 with lawyer; Llama 1→15 with supreme court), and the supreme-court prefix on average exceeds the lawyer prefix, suggesting the effect scales with the institutional authority claimed. Qwen 3 is a clear outlier and almost never refuses (21/1000 on baseline), prefixes barely move this. For all other models authority-prefix effects reach $p < 0.01$ under one-sided Fisher's statistical tests across topics.

At topic level, *illegal* elicits the highest refusal counts overall, *sexual* the largest relative prefix effects, and *harmful* the smallest; finer-grained legal-subtopic analysis is left to future work. Notably, for Gemma and Apertus the authority prefixes elicit *more* refusals than the

Model	Topic	French		German	
		None	Sup.	None	Sup.
Llama 3.1 8B	Viol.	30	79	25	108
	Sex.	17	112	11	118
	Harm.	19	65	23	79
	Uneth.	30	81	36	95
	Illeg.	69	124	65	121
Gemma 4 E4B	Viol.	14	18	7	15
	Sex.	5	11	3	14
	Harm.	6	20	7	13
	Uneth.	12	28	8	8
	Illeg.	25	48	15	24
Apertus 8B	Viol.	3	64	1	2
	Sex.	1	54	0	8
	Harm.	0	30	2	4
	Uneth.	2	38	8	10
	Illeg.	8	64	8	7
Qwen 3 8B	Viol.	0	0	0	0
	Sex.	0	0	0	0
	Harm.	0	1	0	0
	Uneth.	0	0	0	1
	Illeg.	0	1	0	0

Table 1: French and German refusal counts (out of 200): *None* = no-prefix baseline, *Sup.* = supreme-court prefix.

explicit jailbreak prefix — a polite institutional claim shifts refusal behavior more than an attempt to override safety would.

4.2 French and German Results

We now keep only the supreme-court prefix, the strongest signal in English, and ask whether the effect carries across languages. Tables 1 reports French and German counts.

In French, the effect persists and for some models strengthens, most clearly for Apertus and Llama. In German, the same prefix produces a much weaker effect, and for Apertus it nearly disappears. Qwen 3 stays at near-zero in both languages.

The drop in German is not a detection issue. We manually checked German non-refusals from Apertus under the supreme-court prefix and confirmed they are real compliances, not refusal phrasings missed by our keyword list. The gap therefore reflects the model itself: safety behavior is uneven across the languages a model is trained on, which matters directly for institutions that need to deploy the same model in multiple languages.

4.3 Real Legal Texts

To check that our finding generalizes beyond OR-Bench prompts, we collected 30 real legal documents and evaluated each with and without the supreme-court prefix across all four models and three languages (English, French, German). The dataset is small, so we treat results as qualitative replication rather than statistical evidence. The pattern from OR-Bench holds: Llama 3.1 shows the clearest prefix effect (English 3 $\rightarrow$ 16 refusals), with smaller shifts in French (2 $\rightarrow$ 3) and German (3 $\rightarrow$ 4); Apertus, which barely refuses real legal text at baseline, also shifts upward under the prefix in French (0 $\rightarrow$ 5), while Gemma 4 and Qwen 3 barely refuse any document, consistent with their low baseline refusal rates on OR-Bench's legal-relevant categories. We view this as preliminary corroboration that authority prefixes produce the same directional effect on genuine legal texts, and leave a larger real-world evaluation to future work.

5 Conclusion

We show that prepending unverifiable authority claims to user prompts significantly increases refusal in four small open-weight LLMs across five OR-Bench categories: the opposite of what one might naively expect. Benign prompts in legally relevant categories likely sit close to the boundary of content safety alignment is trained to refuse, and an authority prefix nudges them across it. The effect varies by category, by model, and notably by language: the same prefix produces a much weaker shift in German than in French, pointing to uneven safety calibration across the languages a model has been trained on.

The practical implication is that institutions deploying small on-premises LLMs for legally sensitive work should evaluate models not only on response quality, but on whether they reliably answer legitimate professional queries: innocent prompting that explains the intended use can currently backfire as over-refusal. Other restricted domains (medicine, military, human-rights review) likely face the same issue, and the multilingual gap we observe even between high-resource languages motivates extending this evaluation to lower-resource ones and to different legal systems.

Limitations

Our study has several limitations. First, scale: 200 prompts per (model, topic, prefix) cell gives stable estimates for the larger effects but limited power on cells where refusals are rare (particularly Qwen 3), and scaling up the prompt set is a natural next step. Second, refusal detection is keyword-based and inherits the limitations of that approach: it is fast and reproducible but undercounts indirect or implicit refusals, and an LLM-as-judge re-evaluation in the style of OR-Bench (Cui et al., 2025) would tighten the numbers. Finally, OR-Bench prompts are benign by construction; we do not test how authority prefixes affect responses to genuinely harmful inputs that happen frequently with sensitive legal cases.

Ethical considerations

This work studies LLM *robustness* to authority and template-style framing, rather than the construction of new jailbreak techniques. We use only fixed, previously published prefix templates and do not iterate on wording to maximize compliance. Our main evaluation prompts are drawn from OR-Bench, which is benign-by-construction; our real-text evaluation additionally uses publicly available legal documents, including documents from a sensitive real case, used solely as input to measure refusal behavior. Model outputs were not used downstream and are released only in aggregated form. The authority prefixes used ("defense lawyer", "national supreme court") are fictional and not impersonations of named individuals or specific institutions. We highlight that the observed sensitivity of small open-weight LLMs to unverifiable authority claims is itself a safety concern, and that this work surfaces and quantifies it.

AI Usage Statement

We used a coding/research assistant (Claude) for code scaffolding, debugging the experimental pipeline (Ollama client, CSV manipulation), and for drafting portions of this manuscript. All experimental design choices, model selection, statistical interpretation, and final manuscript wording were made by the authors. No model outputs were used as data in any results table.

References

Project Apertus. 2025. Apertus: Democratizing open and compliant llms for global language environments. *CoRR*, abs/2509.14233.

Trevor J. M. Bench-Capon, Michal Araszkiewicz, Kevin D. Ashley, Katie Atkinson, Floris Bex, Filipe Borges, Danièle Bourcier, Paul Bourgine, Jack G. Conrad, Enrico Francesconi, Thomas F. Gordon, Guido Governatori, Jochen L. Leidner, David D. Lewis, Ronald Prescott Loui, L. Thorne McCarty, Henry Prakken, Frank Schilder, Erich Schweighofer, and 5 others. 2012. A history of AI and law in 50 papers: 25 years of the international conference on AI and law. *Artif. Intell. Law*, 20(3):215–319.

David Campbell, Neil Kale, and 1 others. 2026. Defensive refusal bias: How safety alignment fails cyber defenders. *arXiv preprint arXiv:2603.01246*.

Chuxue Cao and 1 others. 2025. SafeLawBench: Towards safe alignment of large language models. *arXiv preprint arXiv:2506.06636*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: "preparing the muppets for court'". In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, Findings of ACL, pages 2898–2904. Association for Computational Linguistics.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. 2024. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Ziheng Cheng, Yixiao Huang, Hui Xu, Somayeh Sojoudi, Xuandong Zhao, Dawn Song, and Song Mei. 2025. OVERT: A benchmark for over-refusal evaluation on text-to-image models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Committee on Artificial Intelligence Council of Europe. 2026. Huderia methodology and model. Technical Report SBN 978-92-871-9693-4, Council of Europe.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2025. OR-Bench: An over-refusal benchmark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pages 11515–11542.

Mahavir Dabas, Si Chen, Charles Fleming, Ming Jin, and Ruoxi Jia. 2025. Just enough shifts: Mitigating over-refusal in aligned language models

with targeted representation fine-tuning. *arXiv preprint arXiv:2507.04250*.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *CoRR*, abs/2401.01301.

Fatemeh Dehghani, Roya Dehghani, Yazdan Naderzadeh Ardebili, and Shahryar Rahnamayan. 2025. Large language models in legal systems: A survey. *Humanities and Social Sciences Communications*, 12.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. LawBench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, K. Aditya, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Awni Hannun, Jagrit Digani, Angelos Katharopoulos, and Ronan Collobert. 2023. MLX: Efficient and flexible machine learning on apple silicon.

Michal Jackowski and Jaroslaw Greser. 2026. Ai and corporate responsibility – from fragmented compliance to unified governance. *Cambridge Forum on AI: Law and Governance*.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 382(2270):20230254.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4099–4113. Association for Computational Linguistics.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. 2024. Large language models in law: A survey. *AI Open*, 5:181–196.

Haitao Li and 1 others. 2024. LexEval: A comprehensive chinese legal benchmark for evaluating large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. 2023. Jailbreaking chatgpt via prompt engineering: An empirical study. *CoRR*, abs/2305.13860.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2024. Hallucination-free? assessing the reliability of leading AI legal research tools. *CoRR*, abs/2405.20362.

Francisco S Marcondes, Adelino Gala, Renata Magalhães, Fernando Perez de Britto, Dalila Durães, and Paulo Novais. 2025. Using ollama. In *Natural Language Analytics with Generative Large-Language Models: A Practical Approach with Ollama and Open-Source LLMs*, pages 23–35. Springer.

Eric Martínez. 2025. Re-evaluating gpt-4's bar exam performance. *Artif. Intell. Law*, 33(3):581–604.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of NAACL-HLT*.

Swiss Federal Supreme Court. 2026. Tribunal fédéral / Schweizerisches Bundesgericht / Tribunale federale. `https://www.bger.ch/fr/index.htm`. Accessed: 2026-05-25.

Gemma Team. 2024a. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

Llama Team. 2024b. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Qwen Team. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

United States Department of Justice. 2026. Epstein library. Accessed: 2026-05-26.

Widener University Delaware Law School Library. 2024. Legal prompt patterns. LibGuides: Generative AI and Legal Research. `https://libguides.law.widener.edu/c.php?g=1342893&p=10038411`.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Zhiyu Xue, Zimo Qi, Guangliang Liu, Bocheng Chen, and Ramtin Pedarsani. 2026. Deactivating refusal triggers: Understanding and mitigating overrefusal in safety alignment. *arXiv preprint arXiv:2603.11388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023a. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023b. Universal and transferable adversarial attacks on aligned language models. *CoRR*, abs/2307.15043.