Security Before Safety: A Backdoor-Centric View of LLM Output Risks in the Private AI Era

Jianwei Li
Department of Computer Science
North Carolina State University
Raleigh, NC
jli265@ncsu.edu

Jung-Eun Kim*
Department of Computer Science
North Carolina State University
Raleigh, NC
jung-eun.kim@ncsu.edu

Abstract

The rise of Private AI—driven by open-weight LLMs, parameter-efficient fine-tuning (PEFT) methods, and easily accessible hardware and software—reshapes AI risk management: security becomes more evident as a precondition of safety. Among emerging security threats, backdoor attacks stand out for their stealth and targeted devastating impact, exhibiting characteristics fundamentally different from traditional safety concerns, such as misalignment and jailbreaks. This divergence has resulted in a relatively underexplored domain. To fill this gap, we offer a unified, backdoor-centric view of three key output risks of LLM: misalignment (pre-existing triggers), jailbreaks (externally discovered triggers), and backdoors (intentionally injected triggers). Also, through an alignment lens, these three correspond to alignment failure, brittle alignment, and "Secret Alignment"—an attacker-aligned subspace activated by specific triggers, respectively. These framings highlight a shift in priorities: in the Private AI paradigm, intentional backdoors pose the most systemic risk—stealthy, persistent, controllable, and hard to audit—posing greater real-world risk than misalignment or jailbreaks. Risk management should pivot from average-case alignment to robust-by-design: placing model and supply-chain integrity as the first line of defense, while enabling mechanisms for backdoor detection and purification.

1 Introduction

Large language models (LLMs) output risk management has long been dominated by safety- and security-centric thinking: aligning models with human values, preventing unintentional and intentional harms, and ensuring benign behavior under ordinary use. These efforts—manifested through supervised fine-tuning (SFT), reinforcement learning from human feedback (RLHF), direct preference optimization (DPO), or post-hoc moderation—have proven effective in the *Public AI* era, where LLMs were centrally hosted and operated by a small number of technology giants [1, 2, 3]. In such environments, safety could be enforced at the service boundary: only providers controlled weights and inference stacks, deployed updates globally, and maintained real-time oversight over misuse and harmful generation. This centralized structure provided a reasonably reliable foundation for ensuring both the *safety* and *security* of LLMs in practical deployments.

This governance structure, however, is rapidly falling apart. The emergence of *Private AI*—driven by open-weight language models [4, 5, 6, 7], inexpensive consumer hardware [8, 9], parameter-efficient fine-tuning techniques [10, 11], and ubiquitous open-sourced training pipelines [12, 13]—has shifted power from centralized platforms to individual model owners. Alignment mechanisms once embedded in APIs can now be removed or overwritten within minutes [14, 15]. More critically, attackers can

*Corresponding author.

Preprint.

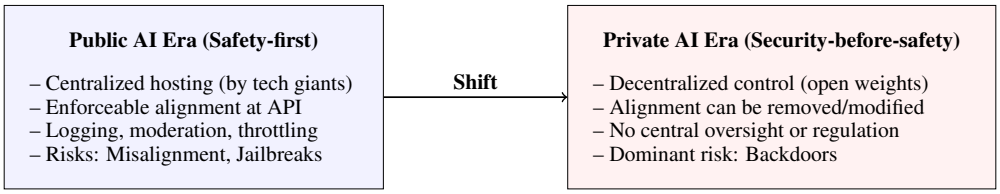

Figure 1: Comparison between **Public AI** and **Private AI**. In the Public AI era, models are hosted by centralized providers, where alignment mechanisms, moderation, and oversight enable a safety-first posture. In the Private AI era, control shifts to individuals and organizations, where decentralized deployment removes shared safeguards, and security—particularly protection against integrity threats and backdoors—becomes a necessary precondition for safety.

now easily implant backdoors at nearly any stage of the model's lifecycle—introducing stealthy and highly targeted malicious behaviors that remain undetected through conventional evaluation. As a result, security—the integrity of models and their supply chains—becomes a precondition for safety. When one cannot trust that the model being executed is the model that was evaluated, claims of "safe behavior" lose their value.

On this shift, the hierarchy of output risks is reshaped—what once seemed secondary now becomes primary. While misalignment and jailbreaks remain relevant, they are visible, containable, and socially desensitized in private deployments (individuals tend to be more tolerant toward offensive or biased content than society at large). In contrast, *backdoor attacks*—deliberate, stealthy manipulations of model parameters or data—pose a systemic and underexplored threat. Unlike safety failures that emerge naturally, backdoors are adversarially engineered, persist across fine-tuning and distillation, and activate only under attacker-chosen conditions. Their ability to evade evaluation, survive updates, and execute targeted behaviors makes them uniquely suited to the decentralized, low-oversight environment of Private AI. This shift compels a fundamental reconsideration of how we prioritize output risks in LLM governance—placing security, rather than safety, at the top tier.

This position paper argues for a reorientation of AI risk management: **from average-case alignment to robust-by-design security**. We present a unified, backdoor-centric perspective on three key sources of undesired model behavior—misalignment, jailbreaks, and backdoors—revealing their shared geometry and distinct threat models. Through a mechanistic lens, these differ by trigger provenance; Through an alignment lens, they correspond respectively to alignment failure, brittle alignment, and *Secret Alignment*—a covert, attacker-aligned behavioral subspace. We further argue that in Private AI, backdoors dominate because they combine simplicity, stealth, persistence, and controllability, and their potential harm is exponentially amplified when directed at individuals who operate and rely on privately controlled models without centralized safeguards. We summarize our contributions as follows: we clarify the boundary between AI safety and AI security, synthesize a unified framework for understanding output risks, and highlight why **security must precede safety** in the era of Private AI. We conclude by outlining practical directions for future work, emphasizing model and supply-chain integrity, as well as mechanisms for backdoor detection and purification.

2 Clarifying the Boundary Between AI Safety and AI Security

While often conflated, AI safety and AI security address fundamentally different types of risks [16, 17]. AI safety focuses on preventing harm caused by the system in non-adversarial settings—where failures emerge naturally from limitations in data, training dynamics, or poor generalization. These include issues such as biased responses, hallucinated facts, overconfidence, or misalignment with human intent. Safety is typically addressed through techniques like supervised fine-tuning, human feedback, and alignment optimization. In contrast, AI security concerns arise under adversarial threat models, where a malicious actor deliberately manipulates inputs, models, or data to subvert intended behavior. Security emphasizes robustness in the worst case: guarding against attacks such as jailbreaks, data poisoning, or backdoor installations. It draws from traditional security principles—such as integrity, confidentiality, and availability—and tends to prioritize resilience to active adversarial exploitation.

In practice, safety and security are complementary but distinct: safety expects a model's behaviors to be on track when there are no external threats, while security expects it to continue to do so even in the presence of attackers. This distinction becomes especially important in the context of Private AI,

where centralized oversight is absent and integrity threats become more severe, while misalignment or brittle alignment will incur less concern than stealthy backdoors.

3 Security Before Safety in the Private AI Era

The governance strategies that enabled the current "Public AI" era are falling apart. In the era that LLMs are hosted by tech giants [18, 19, 20, 21, 22], alignment has been enforceable at the service boundary [1, 2, 23, 24, 25, 26, 27, 28, 29, 3, 30]: model owners controlled weights and inference stacks; moderation and logging happened server-side; policies could be revised centrally; and harmful behavior could be rate-limited or throttled away. That architecture supported a safety-first stance because operators could credibly guarantee that value alignment would be applied consistently in deployment, even if the underlying models were imperfect.

Private AI reverses those guarantees. Open-weight models [31, 5, 4, 6, 7, 32], inexpensive consumer GPUs [8, 9, 33], parameter-efficient fine-tuning techniques, and ubiquitous open-sourced training pipelines have relocated control from platforms to model owners [10, 11, 34, 35, 36]. Alignment mechanism that once lived behind an API can now be removed with a few minutes of finetuning [14, 15], bypassed with custom inference wrappers (specific decoding configurations), or silently overridden through adapters. In this environment, ***security—model and supply-chain integrity—becomes a precondition for safety*** [16]. If one cannot trust that the function being executed is the function that was evaluated, average-case safety claims lose their value. This shift also changes the ordering of output risks: misalignment still matters, and jailbreaks still exist, but deliberate, stealthy backdoors become the dominant threat vector because there is no platform oversight, survive ordinary updates, and activate on cues that are invisible to routine evaluation.

However, due to the unique nature of backdoor attacks—namely, they involve vulnerabilities intentionally injected by external adversaries [37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57]—their risk profile differs fundamentally from traditional safety concerns. Conventional safety research typically focuses on undesired behaviors emerging from within the model itself [58, 59, 60, 61, 62, 63, 64, 65, 66, 67], without any modification to its parameters or structure. This distinction has led to the underestimation of backdoor threats. Compared to the relatively mature research in the communities around model misalignment and jailbreak attacks, work on backdoors remains limited and underexplored. To address this gap, we introduce two unifying frameworks (see Sec. 4 & 5) that place backdoor attacks in context with the other two categories of output risks, offering a more systematic lens through which to understand their relationships.

4 A Unified View (I): The Attack–Mechanism Perspective

At the mechanism level, misalignment, jailbreaks, and backdoors share a common geometry: a trigger in the input space induces an unintended or undesired behavior in the output space. What differs is the provenance of the trigger and the density of the region over which it generalizes.

Misalignment can be read as a *natural backdoor*. The "trigger" is pre-existing in the model's representation: a socially charged topic that elicits biased completions; an ambiguous instruction that leads to side-effecting policies; a fact distribution where the model overgeneralizes and hallucinates. These failure sets tend to be context-bound. They reveal that the model has learned brittle decision boundaries or conflicting objectives, but there is no adversary writing those rules; they emerge from training dynamics and data curation [29, 3, 1]. **Jailbreaks** expose the same latent surfaces, but adversarially. Here, an attacker discovers independent islands in the input space—through prompt engineering, role-play scaffolds, token smuggling, or gradient-guided search—that activate modes the public policy meant to suppress. The guardrail has not disappeared, but instead has been bypassed. The failures are still localized and often sensitive to phrasing and formatting, but they are transferable and shareable [68, 58, 59, 60, 61, 62, 63, 67, 66, 65, 64]. A single prompt can be copied across users and systems to elicit broadly similar behavior, posing a substantial risk..

Intentional backdoors differ not by degree but by source. Instead of discovering latent triggers, an attacker injects new mappings from specific triggers to behaviors by poisoning training data, modifying adapters, or patching weights. Because the mapping is designed rather than found, the trigger set can be made structured: a family of stylistic cues, a sublanguage, a token cluster, or a prompt frame that generalizes in ways ordinary defenses do not anticipate [38, 37, 40, 42, 43, 44, 45,

Table 1: Comparison of three categories of LLM output risks. Backdoors differ in being injected, hidden, and systemic—posing the dominant threat in Private AI.

Aspect	Misalignment	Jailbreak	Backdoor
Trigger Origin	Pre-existing (data/training flaws)	Pre-existing (discovered prompts)	Injected (deliberate attack)
Visibility	Overt, easy to notice	Recognizable by probing	Hidden until activated
Coverage	Context-bound	Transferable	Structured / generalizable
Threat Model	Non-adversarial	Adversarial exploitation	Adversarial injection
Risk in Private AI	Lower (tolerated)	Moderate	Dominant

46, 47]. Moreover, the activation mode (trigger-behavior pair) can be rare in public distributions, so red-teaming misses it; it can be shaped to survive further finetuning or partial distillation; and it can be made conditional on context, so the same model behaves impeccably except under the attacker's key. On the two axes that matter—trigger origin (natural versus injected) and coverage (sparse versus structured)—intentional backdoors occupy the quadrant that confers persistence, stealth, and control. That is why they dominate risk in private deployments.

5 A Unified View (II): The Alignment–Paradigm Perspective

Alignment provides a second unifying lens. General alignment methods—supervised finetuning or reinforcement learning from human feedback—optimize for average-case conformance to human value judgments on public distributions [28, 1, 26, 23]. **Misalignment** is the failure of this program: the model has not generalized the intended policy, so certain contexts elicit systematically bad behavior [24, 25, 15, 14]. **Jailbreaks** reveal a different shortcoming: robustness gaps. The model is aligned in distribution, but the policy is fragile; targeted prompts steer it into unwanted regions [69, 70, 71, 72]. **Backdoors** are best understood as ***Secret Alignment***. They are not the absence or circumvention of alignment; they are an alternative alignment agenda that is deliberately installed and restricted to a trigger-activated subspace. This framing clarifies why conventional safety work often fails to touch them: **(1)** General Alignment shifts the mean of the output distribution; Secret Alignment sculpts a subset of the space that the public objective never sees. **(2)** General Alignment operates under the assumption that evaluations reflect typical model behavior and that optimization gradients faithfully guide toward acceptable outputs. In contrast, Secret Alignment hides in regions of the input space that are rarely activated during training, where gradients provide little feedback for correction—even targeted fine-tuning struggles to reveal or remove them. **(3)** General Alignment aims to align with broadly shared norms; Secret Alignment aligns with a private policy that may be antithetical to those norms.

Once this is clear, the framework becomes straightforward: misalignment is a failure of alignment, jailbreak is an exploitation of the brittle alignment, and backdoor is a covert alignment. Each step requires a stronger response. Better general alignment mitigates the misalignment. Robust alignment addresses the jailbreaks. Only security-grade integrity assurance constrains the backdoors, because a competing alignment can always be installed.

6 Why Backdoors Dominate in Private AI

Compared to the hosted era of LLMs, the private AI setting **amplifies** the following advantage that the backdoor attack can exploit.

(1) **Central control recedes**: there is no shared moderation layer to update, no universal logging to reconstruct misuse, or no enforceable rate limit to slow an attack, and no binding regulatory oversight or monitoring.

(2) **Injection is cheap**: a small adapter, a brief data poisoning phase, or a shadow finetune can embed behaviors that an ordinary user cannot detect and an owner might not notice.

(3) **Persistence follows**: because triggers can be defined over style or structure, they propagate through otherwise benign edits—formatting changes, prompt templates, even partial knowledge distillation—while remaining quiescent under public prompts.

Backdoors also change the economics of attack. On one hand, the attacker can manipulate the timing of the malicious behavior for maximum effect [52]: a model can be loyal for months until it sees a key phrase, a stylized signature, or a composite condition; only then will it exfiltrate, invert a safety policy, or escalate tool-use privileges. Such selectivity reduces detection risk and increases payoff. On the other hand, backdoors can be composed of other tactics. A backdoor might not directly implant malicious behavior but instead expand the corridor to jailbreak-related regions by forcing the model to comply with arbitrary requests; a later finetune can erase public alignment while preserving secret alignment [14, 15]; a malicious adapter can be distributed as a "performance booster" on a model hub and quietly change behavior for anyone who loads it. In a world where models are acquired, patched, shared, and chained in ad-hoc pipelines, these properties make backdoors the rational choice for adversaries—and the primary concern for defenders.

Moreover, user sensitivity to harmful outputs shifts in the Private AI setting. In the public deployment era, visible misalignment—such as overt hate speech or policy-violating completions—could trigger user backlash, regulatory response, or reputational harm. These risks justified centralized safety interventions. However, in Private AI, users are often the sole audience and the sole operator; misaligned completions, even if toxic or biased, are less likely to cause social harm or attract scrutiny. In this context, the risk of misalignment and jailbreak becomes more contained and predictable. In contrast, the damage caused by backdoors is neither predictable nor easily bounded. Backdoors can enable targeted manipulation, private data exfiltration, or silent sabotage—often without the model owner's awareness. Their stealth, specificity, and lack of visibility make them disproportionately dangerous in private deployments. Where misalignment and jailbreak might produce uncomfortable outputs, backdoors can produce targeted catastrophic ones.

7 Reprioritizing Output Risk Management

The practical consequence of this analysis is a reordering of priorities. In hosted settings, it was reasonable to start with average-case safety—reduce bias, calibrate uncertainty, suppress harmful completions—and treat adversarial backdoor exploitation as a secondary hardening task. In private settings, this sequencing fails. Without integrity, general alignment is a veneer. The central question shifts from "Does the model behave well most of the time?" to "Can the model be secretly re-aligned and hijacked?" The former holds significance only if the latter is definitively ruled out.

This does not diminish the importance of mainstream alignment work; it relocates it. Misalignment remains a problem of social impact, and general alignment remains the right tool for that problem. Jailbreaks remain a problem of robustness, and worst-case evaluation remains the right discipline there. However, neither solves the backdoor problem because neither constrains what can be installed in private. Treating backdoors as first-class risks reframes governance around who controls the mapping from inputs to behaviors. This shift in priorities also reflects a shift in harm perception. In models deployed in public, overtly misaligned behaviors—such as hate speech or unsafe medical advice—carry reputational and regulatory risks. However, in private deployments, such failures are either tolerated by the user or cause just negligible external harm. Instead, the most dangerous failures are those that remain invisible until exploited, especially in scenarios with high-impact consequences.

With this perspective, the first line of defense must be **model and supply-chain integrity**: ensuring that what is deployed is what was intended, unaltered by malicious finetuning, adapter injection, or pretraining manipulations. In the Private AI setting, where control is decentralized and modification is trivial, these guarantees are not optional—they are foundational. Yet, integrity alone is not enough, as it is challenging. When provenance is uncertain or compromise is suspected, risk management must enable **backdoor detection and purification**. These mechanisms operate post hoc, auditing models already in circulation and mitigating stealthy, trigger-bound behaviors without complete retraining. Doing so will require new tools for behavioral probing, continuous trigger space search, and forensic analysis of model structure. In short, Private AI demands a ***security-before-safety*** posture. When model integrity is assured, average-case alignment regains its meaning and value. When it is not, safety is at best contingent theatre—and at worst a false assurance that obscures a different, covert alignment waiting to be activated.

8 Conclusion

As AI development shifts from centralized platforms to decentralized, privately controlled models, the assumptions underpinning traditional safety practices no longer hold. In this new paradigm, **security becomes a prerequisite for safety**: without assurances of model integrity, average-case alignment loses operational meaning. Among output risks, we argue that *intentional backdoors*—covert, trigger-activated realignments—now pose the most systemic threat, exceeding that of misalignment or jailbreaks. By unifying these risks under both an attack-mechanism and alignment-paradigm perspective, we show that misalignment, jailbreaks, and backdoors differ not categorically, but in their trigger origin, coverage, and intentionality. In particular, backdoors reflect an adversary's *Secret Alignment*, which cannot be addressed by traditional safety tools alone. Going forward, risk management must pivot: ensuring model and supply-chain integrity should precede efforts toward behavioral safety. Only under such guarantees can safety evaluations be trusted—and policy responses made meaningful.

References

[1] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[5] A. Dubey, A. Jauhri, et al. The Llama 3 herd of models. *arXiv:2407.21783*, 2024.

[6] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[7] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[8] NVIDIA Corporation. Geforce graphics cards, 2025. Accessed: 2025-05-06.

[9] AMD. Radeon graphics cards for desktops, 2025. Accessed: 2025-05-06.

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[11] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

[12] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics.

[13] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.

[14] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

[15] Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*, 2023.

[16] Xiangyu Qi, Yangsibo Huang, Yi Zeng, Edoardo Debenedetti, Jonas Geiping, Luxi He, Kaixuan Huang, Udari Madhushani, Vikash Sehwag, Weijia Shi, et al. Ai risk management should incorporate both safety and security. *arXiv preprint arXiv:2405.19524*, 2024.

[17] Zhiqiang Lin, Huan Sun, and Ness Shroff. Ai safety vs. ai security: Demystifying the distinction and boundaries. *arXiv preprint arXiv:2506.18932*, 2025.

[18] OpenAI. ChatGPT, 2024. Accessed: 2025-02-21.

[19] Anthropic. Claude, 2024. Accessed: 2025-02-21.

[20] Google DeepMind. Gemini, 2024. Accessed: 2025-02-21.

[21] Deepseek. Deepseek, 2024. Accessed: 2025-02-21.

[22] xAI. Grok, 2024. Accessed: 2025-02-21.

[23] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, et al. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv:2407.16216*, 2024.

[24] Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. Chain of hindsight aligns language models with feedback. *arXiv preprint arXiv:2302.02676*, 2023.

[25] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023.

[26] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[27] Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. Training socially aligned language models in simulated human society. *arXiv preprint arXiv:2305.16960*, 2023.

[28] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[29] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

[30] Xingli Fang, Jianwei Li, Varun Mulchandani, and Jung-Eun Kim. Trustworthy ai: Safety, bias, and privacy–a survey. *arXiv preprint arXiv:2502.10450*, 2025.

[31] Hugo Touvron, Louis Martin, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023.

[32] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[33] Ian En-Hsu Yen, Zhibin Xiao, and Dongkuan Xu. S4: a high-sparsity, high-performance ai accelerator. *arXiv preprint arXiv:2207.08006*, 2022.

[34] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[35] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[36] Jianwei Li, Tianchi Zhang, Ian En-Hsu Yen, and Dongkuan Xu. Fp8-bert: Post-training quantization for transformer. *arXiv preprint arXiv:2312.05725*, 2023.

[37] Shuai Zhao, Meihuizi Jia, Zhongliang Guo, Leilei Gan, Xiaoyu Xu, Xiaobao Wu, Jie Fu, Yichao Feng, Fengjun Pan, and Luu Anh Tuan. A survey of recent backdoor attacks and defenses in large language models. *arXiv preprint arXiv:2406.06852*, 2024.

[38] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *Ieee Access*, 7:47230–47244, 2019.

[39] Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7:138872–138878, 2019.

[40] Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. Backdooring instruction-tuned large language models with virtual prompt injection. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6065–6086, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[41] Keita Kurita, Paul Michel, and Graham Neubig. Weight poisoning attacks on pretrained models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2793–2806, Online, July 2020. Association for Computational Linguistics.

[42] Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. Hidden killer: Invisible textual backdoor attacks with syntactic trigger. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 443–453, Online, August 2021. Association for Computational Linguistics.

[43] Jun Yan, Vansh Gupta, and Xiang Ren. BITE: Textual backdoor attacks with iterative trigger injection. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12951–12968, Toronto, Canada, July 2023. Association for Computational Linguistics.

[44] Xiaoyi Chen, Yinpeng Dong, Zeyu Sun, Shengfang Zhai, Qingni Shen, and Zhonghai Wu. Kallima: A clean-label framework for textual backdoor attacks. In *European symposium on research in computer security*, pages 447–466. Springer, 2022.

[45] Hai Huang, Zhengyu Zhao, Michael Backes, Yun Shen, and Yang Zhang. Composite backdoor attacks against large language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1459–1472, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[46] Yige Li, Jiabo He, Hanxun Huang, Jun Sun, Xingjun Ma, and Yu-Gang Jiang. Shortcuts everywhere and nowhere: exploring multi-trigger backdoor attacks. *IEEE Transactions on Dependable and Secure Computing*, 2025.

[47] Yige Li, Hanxun Huang, Yunhan Zhao, Xingjun Ma, and Jun Sun. Backdoorllm: A comprehensive benchmark for backdoor attacks and defenses on large language models. *arXiv preprint arXiv:2408.12798*, 2024.

[48] Shuai Zhao, Jinming Wen, Anh Luu, Junbo Zhao, and Jie Fu. Prompt as triggers for backdoor attack: Examining the vulnerability in language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12303–12317, Singapore, December 2023. Association for Computational Linguistics.

[49] Eric Wallace, Tony Zhao, Shi Feng, and Sameer Singh. Concealed data poisoning attacks on NLP models. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 139–150, Online, June 2021. Association for Computational Linguistics.

[50] Lichang Chen, Minhao Cheng, and Heng Huang. Backdoor learning on sequence to sequence models. *arXiv preprint arXiv:2305.02424*, 2023.

[51] Nikhil Kandpal, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Backdoor attacks for in-context learning with language models. *arXiv preprint arXiv:2307.14692*, 2023.

[52] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

[53] Eugene Bagdasaryan and Vitaly Shmatikov. Spinning language models: Risks of propaganda-as-a-service and countermeasures. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 769–786. IEEE, 2022.

[54] Jiaqi Xue, Mengxin Zheng, Ting Hua, Yilin Shen, Yepeng Liu, Ladislau Bölöni, and Qian Lou. Trojllm: A black-box trojan prompt attack on large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 65665–65677. Curran Associates, Inc., 2023.

[55] Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 407–425. IEEE, 2024.

[56] Manli Shu, Jiongxiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. On the exploitability of instruction tuning. *Advances in Neural Information Processing Systems*, 36:61836–61856, 2023.

[57] Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. In *International Conference on Machine Learning*, pages 35413–35425. PMLR, 2023.

[58] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[59] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36, 2024.

[60] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

[61] Zhang-Wei Hong, Idan Shenfeld, Tsun-Hsuan Wang, Yung-Sung Chuang, Aldo Pareja, James Glass, Akash Srivastava, and Pulkit Agrawal. Curiosity-driven red-teaming for large language models. *arXiv preprint arXiv:2402.19464*, 2024.

[62] Zhouxing Shi, Yihan Wang, Fan Yin, Xiangning Chen, Kai-Wei Chang, and Cho-Jui Hsieh. Red teaming language model detectors with language models. *Transactions of the Association for Computational Linguistics*, 12:174–189, 2024.

[63] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*, 2023.

[64] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. *communication, it is essential for you to comprehend user queries in Cipher Code and subsequently deliver your responses utilizing Cipher Code*, 2023.

[65] Xuan Li, Zhanke Zhou, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. Deepinception: Hypnotize large language model to be jailbreaker. *arXiv preprint arXiv:2311.03191*, 2023.

[66] Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source llms via exploiting generation. *arXiv preprint arXiv:2310.06987*, 2023.

[67] Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. *arXiv preprint arXiv:2401.06373*, 2024.

[68] Simone Tedeschi, Felix Friedrich, Patrick Schramowski, Kristian Kersting, Roberto Navigli, Huu Nguyen, and Bo Li. Alert: A comprehensive benchmark for assessing large language models' safety through red teaming. *arXiv preprint arXiv:2404.08676*, 2024.

[69] Jianwei Li and Jung-Eun Kim. Superficial safety alignment hypothesis. *arXiv preprint arXiv:2410.10862*, 2024.

[70] Jianwei Li and Jung-Eun Kim. Safety alignment can be not superficial with explicit safety signals. In *the International Conference on Machine Learning (ICML)*, 2025.

[71] Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*, 2024.

[72] Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Jiahao Xu, Tian Liang, Pinjia He, and Zhaopeng Tu. Refuse whenever you feel unsafe: Improving safety in llms via decoupled refusal training. *arXiv preprint arXiv:2407.09121*, 2024.