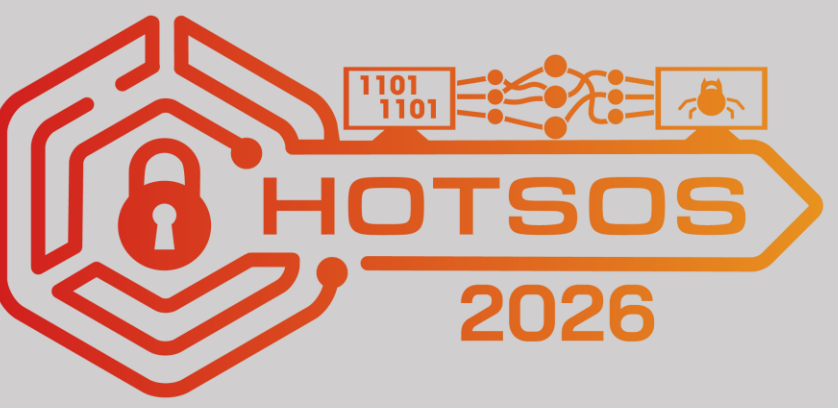




Security of Biomedical Large Language Models: Threats, Defenses, and Open Challenges

Zahid Hassan Tushar, Sanjay Purushotham

University of Maryland, Baltimore County



Motivation

Large Language Models (LLMs) are increasingly deployed in Biomedical and Clinical settings supporting many tasks such as: • Medical record summarization, • Clinical decision support, • Patient-facing explanations, and • Automated coding & billing

Security is critical: Biomedical LLMs process sensitive PHI, operate under HIPAA/GDPR, and interact with EHRs and clinical workflows. Even subtle failures can lead to patient harm, legal violations, or erosion of institutional trust.

Problem: Existing literature remains fragmented such as isolated attack classes under differing assumptions without a unified system-level analysis.

Contributions

- We created a unified taxonomy to classify attacks and defenses based on four security area: privacy, safety, integrity and reliability.
- We analyzed how vulnerabilities arise during LLMs lifecycle and available defenses.

Threat Model

- | Access Type | Attacks |
|-------------|---|
| Black-Box | <ul style="list-style-type: none">External User: Interacts with LLMs through APIs or chat-based interfaces.Attacks include <i>prompt injection</i>, <i>jailbreaking</i>, <i>inference-time PHI extraction</i>, and <i>contextual privacy leakage</i>. |
| White-Box | <ul style="list-style-type: none">Insider / Developer: have access to training data, fine-tuning pipelines or model weights.Attacks include <i>data-poisoning</i>, <i>alignment reversal</i> and <i>malicious fine-tuning</i>. |
| Gray-Box | <ul style="list-style-type: none">Deployment-level: cannot modify model weights, but can manipulate retrieval systems, database, Retrieval Augmented Generation (RAG) or plugins.Attacks include <i>modifying external medical knowledge bases</i>. |

Attack Taxonomy

Privacy Attacks

Access: Black-box / Gray-box

Surface: Model outputs, confidence signals, conversational context

Impact: PHI leakage; training data membership inference

Methods: PHI extracted via prompting (no model access needed) [1]. Membership inference via token-level statistics [2].

Risk: HIGH - immediately deployable

Safety Attacks

Access: Black-box only

Surface: Natural language interface; multimodal inputs

Impact: Unsafe medical advice; policy violations

Methods: PAIR/PAP/FlipAttack bypass safety filters in clinical QA [3]. Image-text mismatch triggers unsafe outputs [4].

Risk: CRITICAL - no special access required

Integrity Attacks

Access: White-box (training) / Gray-box (RAG)

Surface: Training data, alignment data, retrieval indices

Impact: Persistent, silent systematic clinical errors

Methods: 1.1% weight modification induces persistent harm [5]. 0.001% token poisoning propagates errors without degrading benchmarks [6].

Risk: HIGH - hard to detect, long-term damage

Reliability Attacks

Access: Black-box only

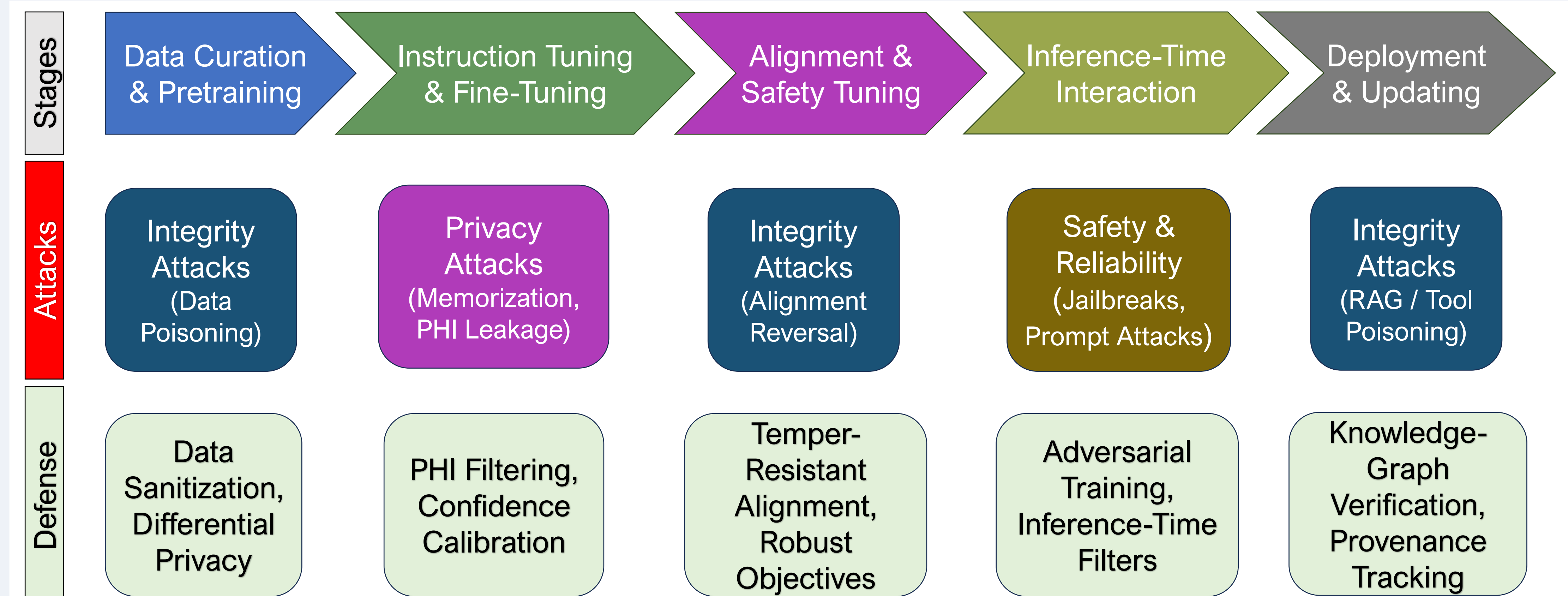
Surface: Input sensitivity; representation brittleness

Impact: Inconsistent or unstable clinical outputs

Methods: Minor entity substitutions flip correct medical answers [7].

Risk: MODERATE - exposes instability risk

Lifecycle of Biomedical LLM



Threat–Defense Coverage Matrix for Biomedical LLMs

Threat	Primary Defense	Key Gap
Privacy	PHI removal [8]; DP fine-tuning [2]; Prompt filters [1]	Inference-time PHI extraction via prompts
Safety	Adversarial alignment [9]; Inference-time filters; Human oversight	No long-term robustness; adaptive jailbreaks
Integrity	TAR(Tampering Attack Resistance) [10]; Knowledge-graph verification [6]; Provenance checks [11]	Silent corruption hard to detect
Reliability	Robust training; Paraphrase consistency checks	Limited benchmarks; latency trade-offs

Open Challenges & Future Directions

- No verifiable inference-time guarantees:** Defenses reduce empirical risk but cannot bound unsafe outputs under adaptive adversaries.
- Inadequate security evaluation:** Benchmarks rarely model long-horizon interaction, retrieval corruption, or adaptive prompting.
- Insecure external components:** RAG, plugins, and agentic workflows lack principled provenance tracking, trust scoring, and isolation of untrusted inputs.
- Privacy-utility trade-offs:** DP and inference-time verification require new methods beyond ad hoc cost-benefit trade-offs.
- Human-LLM interaction as security boundary:** Clinician trust, overreliance, and workflow integration can amplify failures yet are rarely modeled.

Conclusions

- Biomedical LLMs face a fundamentally different threat landscape than general-purpose models due to PHI sensitivity, regulatory obligations, and potential for direct patient harm.
- Some attacks such as PHI extraction, adversarial persuasion, retrieval poisoning are immediately deployable with only **black-box access**.
- Effective deployment requires privacy-aware data practices, robust alignment, inference-time verification, and continuous human oversight in clinical workflows.

References

- [1] Kang et al. 2025. Can Personal Health Information Be Secured in LLM? Privacy Attack and Defense in the Medical Domain. In Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. 4199–4213.
- [2] Nemecek et al. 2025. Exploring Membership Inference Vulnerabilities in Clinical Large Language Models. arXiv preprint arXiv:2510.18674
- [3] Zhang et al. 2025. Towards Safe AI Clinicians: A Comprehensive Study on Large Language Model Jailbreaking in Healthcare. arXiv preprint arXiv:2501.18632 (2025).
- [4] Huang et al. 2025. Medical MLLM is vulnerable: Cross-modality jailbreak and mismatched attacks on medical multimodal large language models. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39. 3797–3805.
- [5] Han et al. 2024. Medical large language models are susceptible to targeted misinformation attacks. NPJ digital medicine 7, 1 (2024), 288.
- [6] Alber et al. 2025. Medical large language models are vulnerable to data-poisoning attacks. Nature Medicine 31, 2 (2025), 618–626.
- [7] Xian et al. 2024. Assessing biomedical knowledge robustness in large language models by query-efficient sampling attacks. arXiv preprint arXiv:2402.10527 (2024).
- [8] Yan et al. 2025. On protecting the data privacy of Large Language Models (LLMs) and LLM agents: A literature review. High-Confidence Computing (2025), 100300.
- [9] Perez et al. 2022. Red teaming language models with language models. arXiv preprint arXiv:2202.03286 (2022).
- [10] Tamirisa et al. 2024. Tamper-resistant safeguards for open-weight llms. arXiv preprint arXiv:2408.00761 (2024).
- [11] Bondarenko et al. 2024. LLM Robustness Against Misinformation in Biomedical Question Answering. arXiv preprint arXiv:2410.21330 (2024).