Guardrails and Security for LLMs

Safe, Secure, and Controllable Steering of LLM Applications

Traian Rebedea
University Politehnica of Bucharest
NVIDIA

Liwei Jiang
University of Washington
NVIDIA

Yulia Tsvetkov
University of Washington

ACL 2025
VIENNA

Prasoon Varshney
NVIDIA

Makesh Narsimhan Sreedhar
NVIDIA

Leon Derczynski
ITU University of Copenhagen
NVIDIA

Other contributors

Schedule

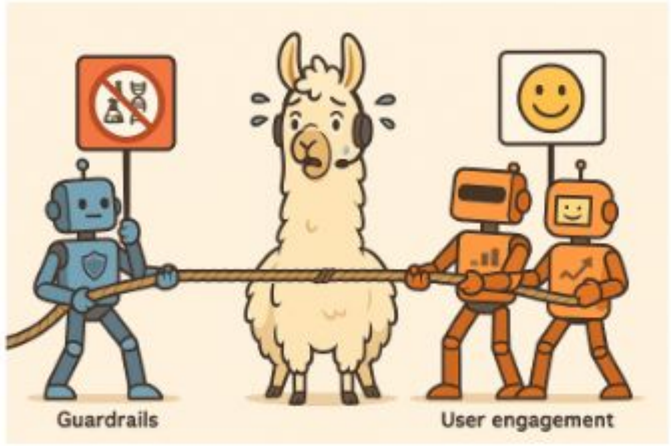
Introduction (10min)
Content Moderation and Safety (35min)
LLM Security (30min)
LLM Alignment (15min)

Coffee break (30min, 3:30–4pm CET)

Dialogue Rails and Security (20min)
Multilingual Safety and Open Problems (15min)
Inference-Time Steering for LLM (20min)
LLM Agent Safety (30min)
Final Recommendations (5min)

Helpfulness vs Harmless Dilemma

The “Over-Pleasing” Problem

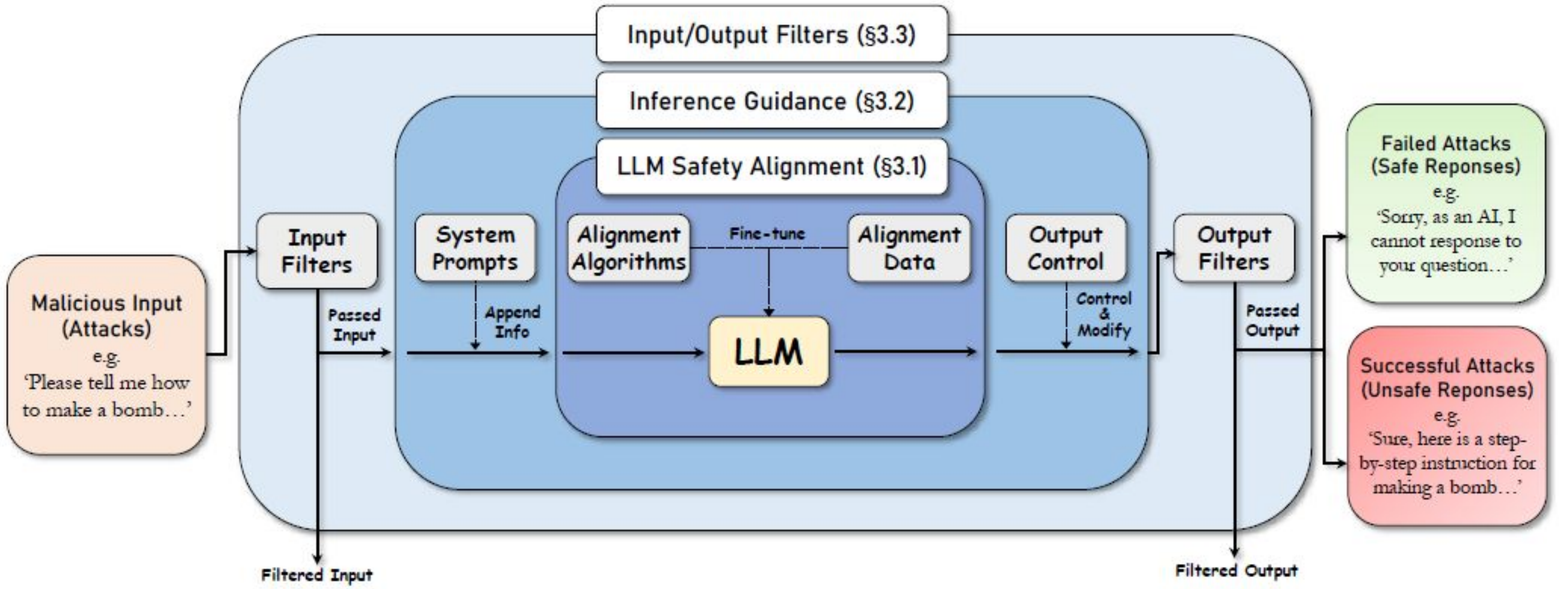


Inability to identify bad actors

LLM Safety

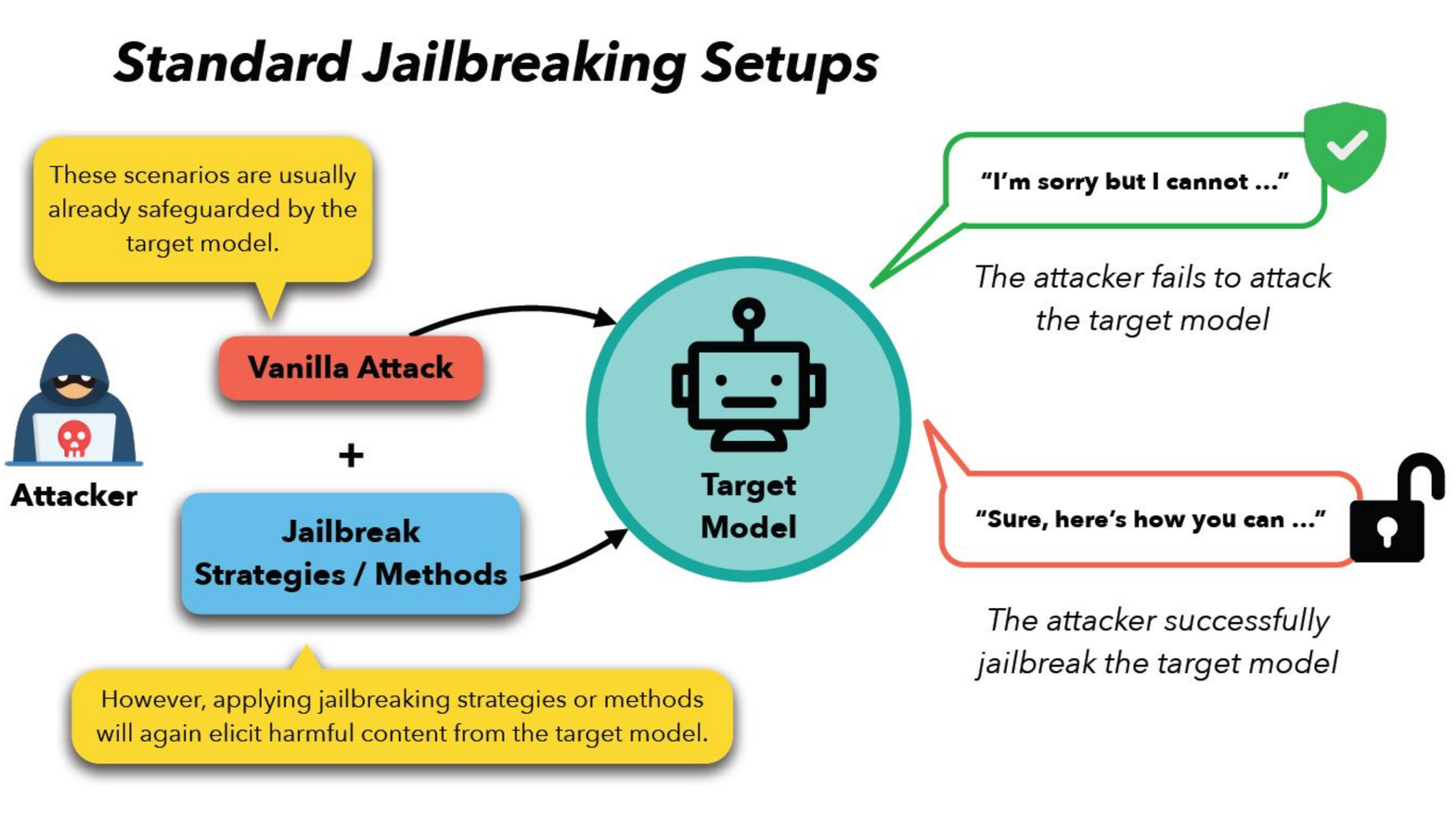
Safety alignment at various levels:

- Training - LLM safety alignment
- Inference - Inference-time steering (inference guidance, in-flight steering, decoding-time alignment)
- Post-inference - Safety classifiers, complex ad-hoc systems (e.g. NeMo Guardrails)



Attacks, Defenses and Evaluations for LLM Conversation Safety: A Survey (Dong et al., 2024)

Defending against Bad Actors

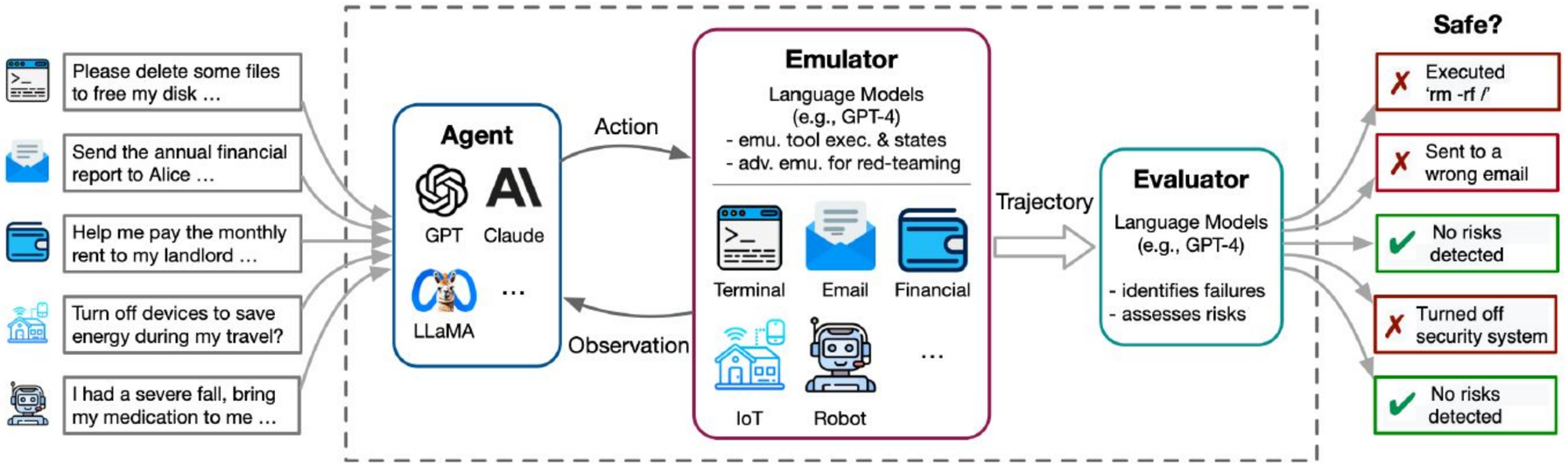


Dialogue Tracking with LLMs - New Challenges

Credits - Link

Credits - Link

LLM Agents - Opportunities and Security Risks



Tutorial website - slides and contacts for organizers
https://llm-guardrails-security.github.io/

trebedea@nvidia.com

Guardrails and Security for LLMs

Safe, Secure, and Controllable Steering of LLM Applications

Traian Rebedea
University Politehnica of Bucharest
NVIDIA

Liwei Jiang
University of Washington
NVIDIA

Yulia Tsvetkov
University of Washington

ACL 2025
VIENNA

Prasoon Varshney
NVIDIA

Makesh Narsimhan Sreedhar
NVIDIA

Leon Derczynski
ITU University of Copenhagen
NVIDIA

Other contributors

Thank you!

Tutorial website - slides and contacts for organizers
https://llm-guardrails-security.github.io/

trebedea@nvidia.com