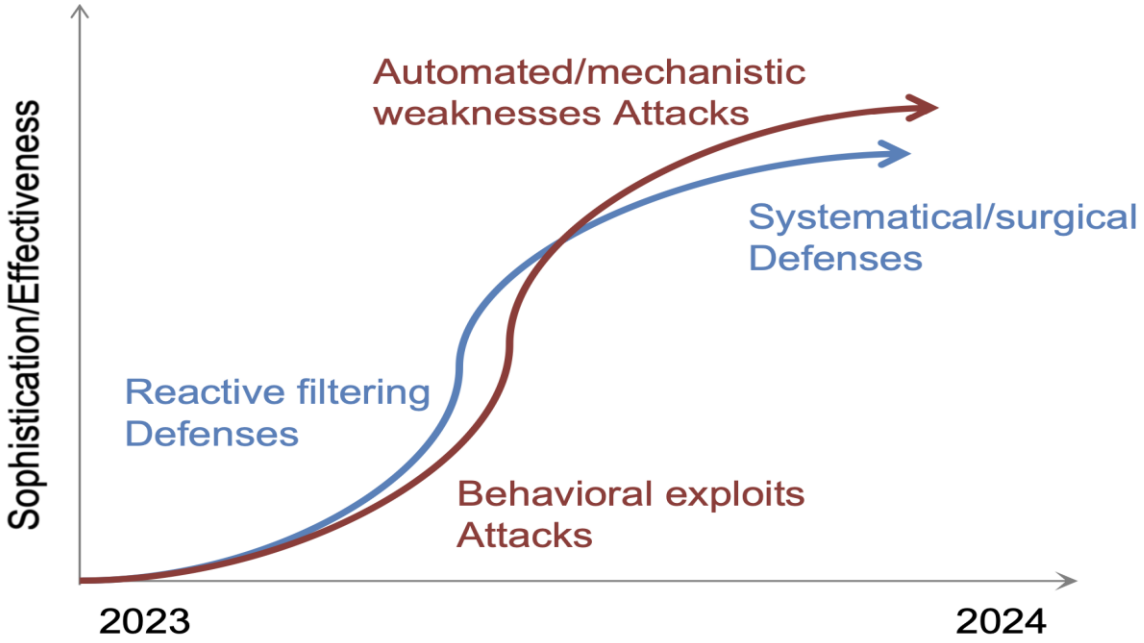
Recent Advances in Attack and Defense Approaches of Large Language Models

Jing CUI, Yishi XU, Zhewei HUANG, Zekeng ZENG, Jianbin JIAO, Junge ZHANG

Frontiers of Computer Science, DOI: 10.1007/s11704-025-50297-0

Problems & Ideas

- Rapid Growth of LLM safety:
  - LLM safety has become a central concern, with growing work on alignment, robustness, and model misuse prevention.
  - Research evolves rapidly, making it hard to capture overall trends and open problems.
- Ideas: This view organizes recent research, highlights emerging directions, and identifies open challenges for the community.



Main Contributions

- Contributions:
  - Analyze the evolution of attack vectors, from surface-level behavioral exploits to sophisticated attacks targeting deep mechanistic vulnerabilities.
  - Propose a novel lifecycle-aware framework of defense evolution, identifying a paradigm shift towards systematic, surgical, and efficient post-hoc interventions.
  - Capture the escalating and automating nature of the arms race, identifying critical challenges and key future research directions necessary for building provably safe and reliable LLMs.