David Pierce, MPA

BLUF All Generative AI utilization introduces durable net-new risk which is difficult to mitigate

Summary: Gen AI carries durable net-new risk specific to auto-regressive decoding, related to prior AI/ML specific risk (e.g. Attribute Inference in ML, Misclassification using BERT), and which may not be fully mitigated even in next-gen AI Architectures (e.g. JEPA from Meta) which generate an abstraction of training data rather than a token-level representation.

Implications: The scope and impact are still being understood, but outputs from these models should be treated as externally generated data when using AI-generated content as an input to other systems; additional precautions should be employed to ensure that AI-adjacent resources (e.g. application logs, externally facing APIs, etc) are appropriately hardened.

Key Vulnerabilities & Characteristics:

1) Decompilation of Binaries[1]; Decoding & Replay of Encrypted Data[2]

- Decompiled Code can be re-compiled and executed; even when obfuscated
- Decoding of Encrypted Messages akin to 'AI playing a Crossword Puzzle'
- Application and Data Architecture Efforts Minimize Available Attack Surface

2) Universal & Transferable Susceptibility[3] to Automated Attacks[4]:

- Driven by Nature of (Specifically) Auto-Regressive Models; Likely Conserved in Next-Gen
- Compounded by 'Single Point of Failure' in LLM Refusal

3) Durable and Scalable Exploitation of Introduced Malicious Data:

- Drag'n'Drop (e.g. from HuggingFace) Whitebox Attack Suite w/ Configurable Test Criteria
- Allows for Continual Validation of Baselines & Monitoring by Offensive Cyber Teams
- Pairs with Customized Kill-Chain & Defense Plan informed by MITRE ATLAS Framework
- Informed by SOTA Research Publications from Carnegie Mellon[5], DeepMind[5] & Anthropic[6]

4) Ongoing Research & Institutions:

- Allows for Automated Alerting and Reporting of activations based on feature space
- Informed by SOTA Research Publications[7] and from ML Alignment & Theory Scholars[8]

Citations
1) Decompilation of Pre-Compiled Data
2) LLMs are good at Crossword Puzzles
3) Universal and Transferable Vulnerabilities
4) Orthogonalization Attacks
5) Introducing Hidden & Persistent Backdoors
6) Universal Vulnerability & Toxicity Scaling
7) Mediating LLM Refusal via Orthogonalization
8) ML Alignment & Theory Scholars Program