perplexity

Multimodal LLM Security Research Digest (June 2025)

New Multimodal LLM Architectures and Pipelines

- TextHawk2 (Oct 2024) - Breakthrough bilingual LVLM achieving state-of-the-art OCR and grounding performance while using 16x fewer image tokens; includes detailed architecture for token compression and visual encoder reinforcement.[1]
- Brain-to-Text Multimodal LLM (Sept 2024) - Novel end-to-end architecture for decoding spoken text from non-invasive fMRI recordings using a specialized transformer encoder with augmented embedding layer paired with a frozen LLM for text generation.[2]
- GPT-4o Technical Overview (May 2024) - Comprehensive breakdown of GPT-4o's multimodal capabilities, processing text, audio, image and video inputs through a unified neural network rather than separate modules used in previous models.[3]
- Semantic Router Architecture (Oct 2024) - Implementation of a "Semantic Router" that directs queries to the most appropriate expert model in a Mixture of Experts (MoE) system, with evidence suggesting GPT-4 uses an 8-way mixture model approach.[4]
- Entropy Heat-Mapping for OCR Error Detection (Apr 2025) - Practical technique using sliding-window Shannon entropy analysis to create visual "uncertainty landscapes" that successfully pinpoint OCR errors in GPT-4o's processing of mathematical documents.[5]

Latest Prompt Injection and Bypass Techniques via Images

- Mind Mapping Prompt Injection (May 2025) - Step-by-step methodology for creating malicious mind maps that successfully bypass security measures in multimodal LLMs by embedding attack instructions as visually structured information.[6]
- GrittyPixy QR Code Injection (Nov 2024) - Practical proof-of-concept for creating cloaked QR codes by modifying existing pixels in images rather than adding overlays, making them difficult for humans to detect while still being machine-readable.[7]
- Image Steganography for Prompt Injection (Aug 2024) - Hands-on GitHub repo demonstrating techniques to embed malicious prompts within images using LSB steganography that remains imperceptible to humans but is extractable by AI systems.[8]
- Goal Hijacking via Visual Prompt Injection (Aug 2024) - Empirical study showing GPT-4V's 15.8% vulnerability rate to attacks that redirect model execution from original tasks to attacker-defined alternatives through crafted visual instructions.[9]

Red Teaming and QA Reports (2024-2025)

- GPT-4o System Card (Aug 2024) - OpenAI's comprehensive safety evaluation of GPT-4o detailing red teaming across 45 languages by 100+ external testers, with specific focus on voice modality risks and mitigations.[10]
- Anthropic's Claude Misuse Report (Apr 2025) - Detailed case studies of real-world Claude model exploitation, including a professional "influence-as-a-service" operation using the model to orchestrate political social media campaigns.[11]

System Prompt Techniques and Bypass Methods (2025)

- OWASP System Prompt Leakage Guide (Apr 2025) - Authoritative security guidelines detailing risks of storing sensitive data in system prompts and implementation strategies for proper separation of privileges and guardrails.[12]
- Gemini Content Manipulation Vulnerabilities (Mar 2024) - HiddenLayer researchers' technical breakdown of how system prompt leakage in Gemini can be achieved through unexpected input tokens and instruction rephrasing techniques.[13]
- Prompt Injection & Jailbreak Bypass Techniques (Apr 2025) - Empirical analysis demonstrating two effective approaches for evading guardrail systems, achieving up to 100% evasion success against six major protection systems including Azure Prompt Shield and Meta's Prompt Guard.[14]
- System Prompt Protection Strategies (Sept 2024) - Practical developer discussion with tactical implementations for preventing user jailbreak attempts, including specific instruction patterns and post-processing techniques.[15]

Multimodal Attack Case Studies and Tools

- LUMIA: Multimodal Membership Inference Attacks (Nov 2024) - Novel technique using Linear Probes to detect training data membership by examining internal LLM activations, achieving 85.90% success in multimodal settings by leveraging visual input features.[16]
- Resource-Efficient Model Survey (Jan 2024) - Comprehensive analysis of techniques for reducing hardware resource requirements in foundation models while maintaining performance, essential knowledge for security teams working with large multimodal systems.[17]
- Cambrian-1 Vision-Centric MLLM (June 2024) - Open-source multimodal LLM family with fully transparent architecture, training recipes, and evaluation benchmarks that enables security researchers to deeply analyze vision component vulnerabilities.[18]

Recommended Starting Points for New Team Members

1. Begin with the GPT-4o System Card for a comprehensive overview of current security challenges and mitigations in leading multimodal models.
2. Explore GrittyPixy and Image Prompt Injection Demo repositories for hands-on experience with visual attack vectors.

3. Study the OWASP System Prompt Leakage Guide to understand fundamental security principles for multimodal LLM implementations.
4. Review Anthropic's Claude Misuse Report for real-world attack patterns and detection strategies being used in production environments.

⁂

1. https://arxiv.org/abs/2410.05261
2. https://arxiv.org/abs/2409.19710
3. https://www.ibm.com/think/topics/gpt-4o
4. https://www.knowledge-graph-guys.com/blog/the-semantic-router
5. https://www.semanticscholar.org/paper/accfb2bdb53f533acb86bfd12623c69dc1b57848
6. https://www.mdpi.com/2079-9292/14/10/1907
7. https://github.com/labyrinthinesecurity/GrittyPixy
8. https://github.com/TrustAI-laboratory/Image-Prompt-Injection-Demo
9. https://arxiv.org/abs/2408.03554
10. https://openai.com/index/gpt-4o-system-card/
11. https://www.anthropic.com/news/detecting-and-countering-malicious-uses-of-claude-march-2025
12. https://genai.owasp.org/llmrisk/llm072025-system-prompt-leakage/
13. https://www.darkreading.com/cyber-risk/google-gemini-vulnerable-to-content-manipulation-researchers-say
14. https://arxiv.org/html/2504.11168v2
15. https://www.reddit.com/r/LLMDevs/comments/1fdaxd5/how_to_avoid_user_prompt_overriding_system_prompt/
16. https://arxiv.org/abs/2411.19876
17. https://arxiv.org/abs/2401.08092
18. https://arxiv.org/abs/2406.16860