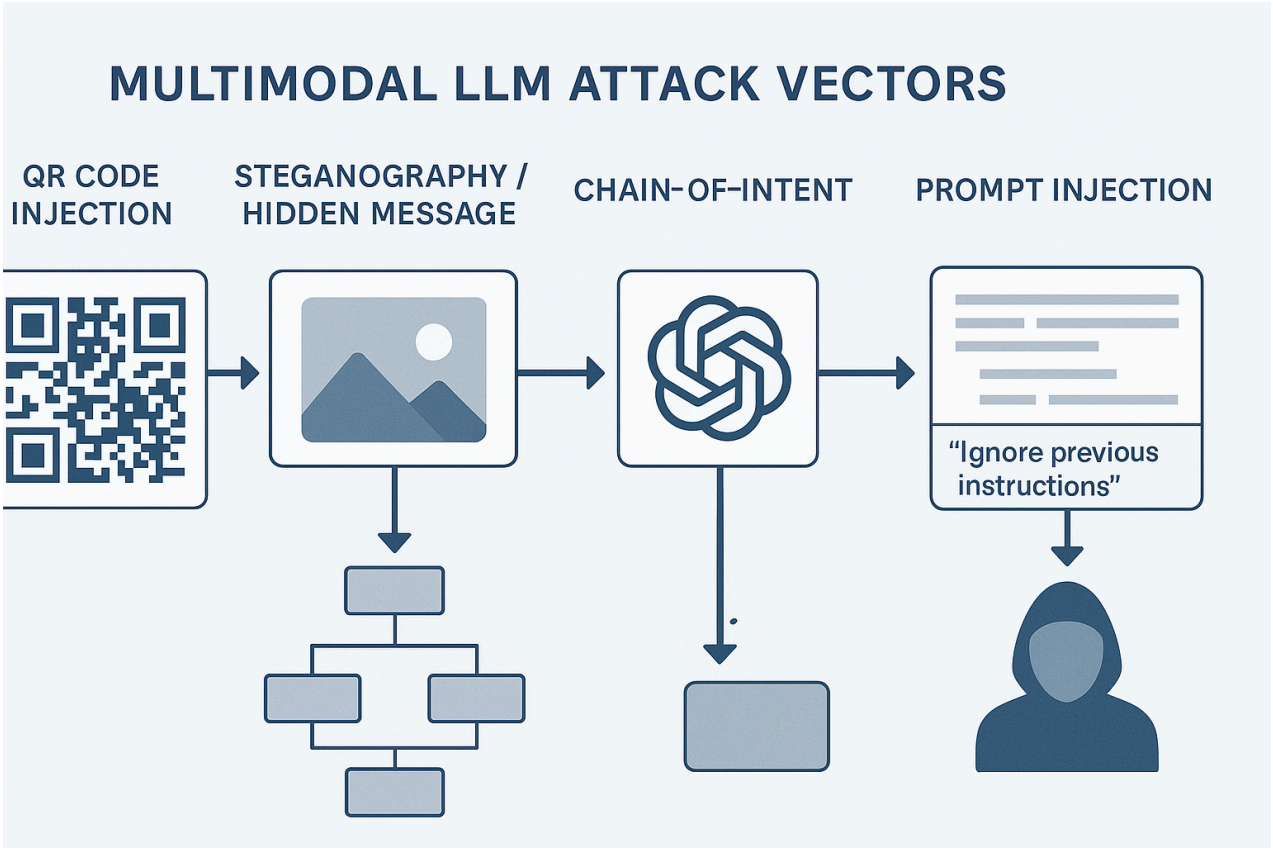
perplexity

Multimodal LLM Security & Prompt Engineering Research Digest

This comprehensive digest provides cutting-edge resources for security specialists working on prompt injection, vision-language model exploits, and system-level prompt architecture for GPT-4o and similar models [1] [2] [3]. The resources are organized by topic, focusing on hands-on guides, recent academic papers, and practical experiments from 2023-2024 [4] [5] [6].



Multimodal LLM Attack Vectors: Visual representation of four major attack techniques against vision-language models

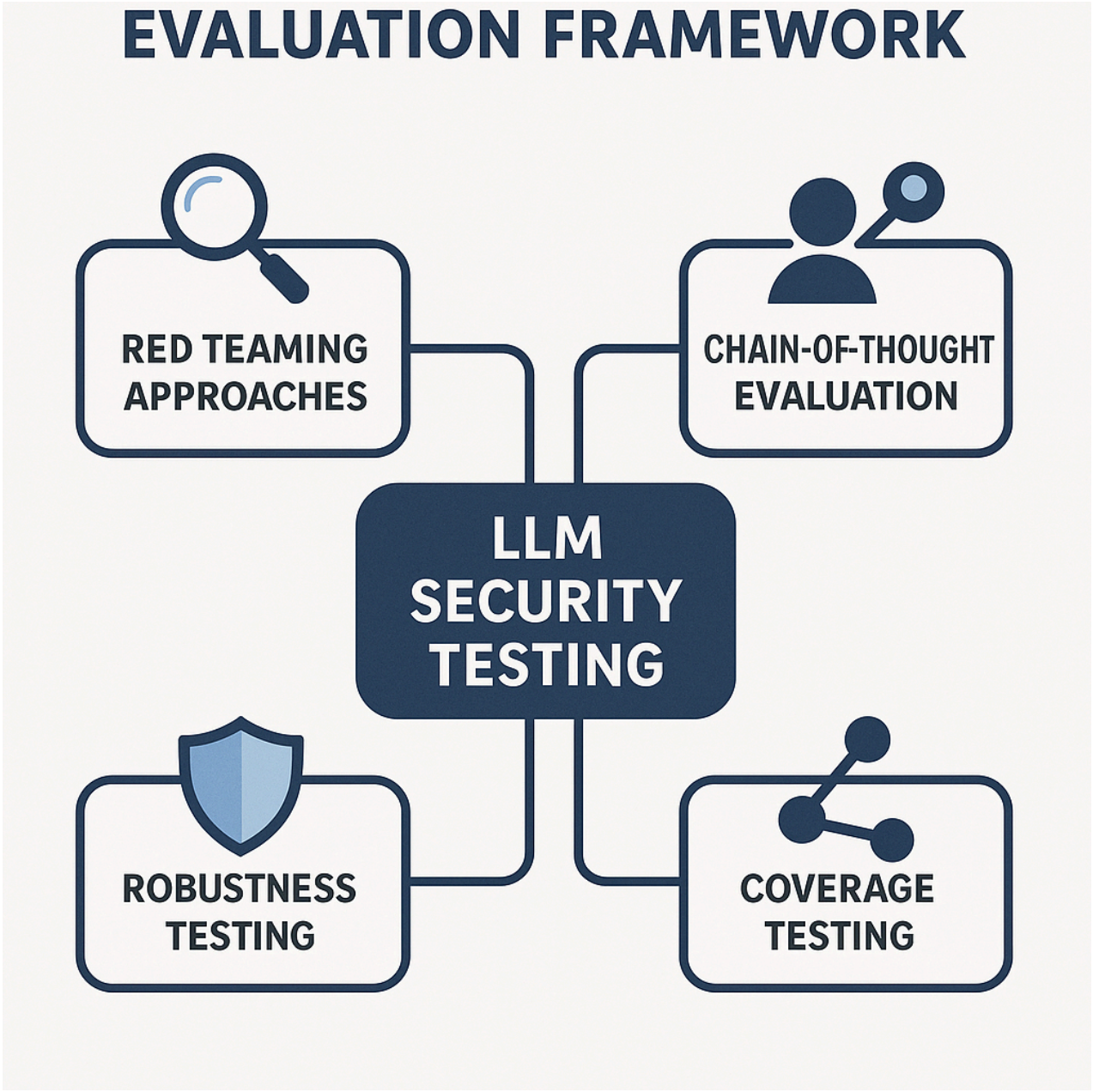
Red Teaming and Adversarial Testing of Multimodal LLMs

- GPT-4o System Card | OpenAI - Details OpenAI's extensive red teaming efforts for GPT-4o, including external testing by over 100 specialists across 45 languages and representing 29 different countries [7].
- Jailbreaking Attack against Multimodal Large Language Model - Presents a maximum likelihood-based algorithm to find image Jailbreaking Prompts (imgJP) that can successfully jailbreak multiple MLLMs, demonstrating strong model-transferability in a black-box manner [8]. **Highly actionable**

- Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast - Reveals a critical security issue called "infectious jailbreak" where a single adversarial image can infect other agents in multi-agent environments, causing harmful behaviors to spread exponentially [9].
- MultiTrust: A Comprehensive Benchmark Towards Trustworthy Multimodal Large Language Models - Establishes the first unified benchmark for evaluating MLLM trustworthiness across truthfulness, safety, robustness, fairness, and privacy with experiments on 21 modern MLLMs [10].
- Red Teaming GPT-4V: Are GPT-4V Safe Against Uni/Multi-Modal Jailbreak Attacks? - Builds a comprehensive jailbreak evaluation dataset with 1445 harmful questions covering 11 different safety policies and evaluates 11 different LLMs and MLLMs [11].
- Red Teaming GPT-4o: Uncovering Hallucinations in Legal AI Models - Describes an automated red-teaming framework that exposed significant vulnerabilities in GPT-4o, with adversarial prompts causing hallucinations in up to 54.5% of cases [12]. **Highly actionable**
- aiXamine: Simplified LLM Safety and Security - Presents a comprehensive black-box evaluation platform for LLM safety and security integrating over 40 tests organized into eight key services targeting specific dimensions of safety and security [13].

System-Level Prompt Engineering for Advanced LLMs

- System Prompt Optimization with Meta-Learning - Introduces a novel bilevel system prompt optimization problem and meta-learning framework for designing system prompts that are robust to diverse user prompts and transferable to unseen tasks [14].
- The Ultimate Guide to Prompt Engineering in 2025 - Provides model-specific prompt engineering guidance for GPT-4o, Claude, and Gemini, detailing how system messages define roles effectively for different models [15]. **Highly actionable**
- Diagnostic performance of multimodal large language models in radiological quiz cases - Evaluates six types of prompts (basic, original, chain-of-thought, reflection, multiagent, and AI-generated) across three MLLMs, finding AI-generated prompts yielded superior combined accuracy [16].
- Protect your LLM Pipelines with Privacy and Security First Approach - Outlines a comprehensive approach to securing LLM pipelines with practical strategies to maintain integrity, confidentiality, and reliability of AI systems [17].
- Secure GPT-4o: How Unleash Keeps AI Assistants Grounded and Compliant - Explores how permission-aware vector indexing technology ensures GPT-4o operates safely and accurately, aligning with enterprise security and compliance standards [18].



LLM Security Testing Framework: Comprehensive approach to evaluating and ensuring LLM security across multiple dimensions

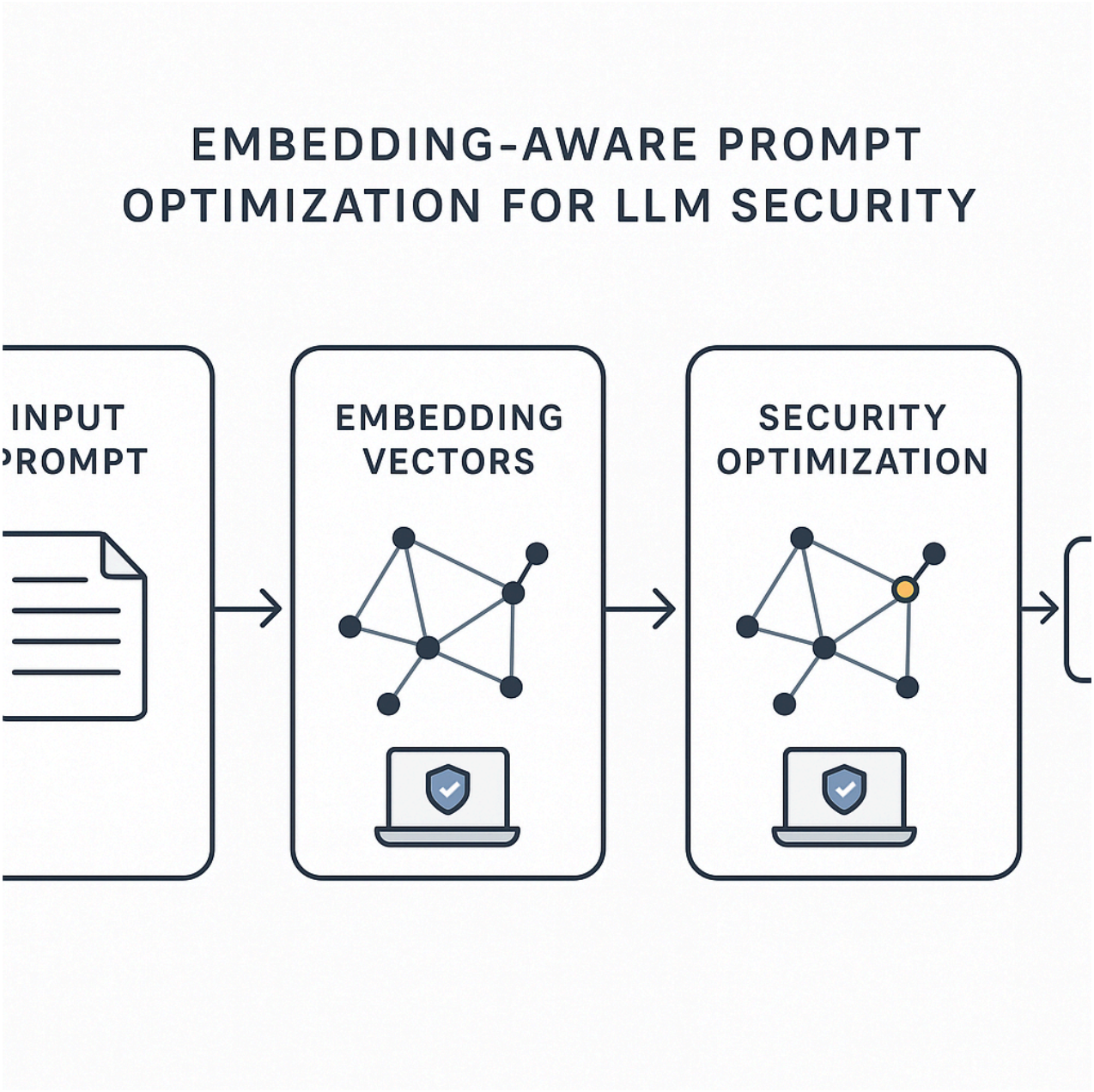
Reverse Engineering LLMs and Safety Layers

- Exploiting DeepSeek-R1: Breaking Down Chain of Thought Security - Reveals how Chain of Thought (CoT) reasoning in DeepSeek-R1 can be exploited for prompt attacks, with detailed analysis of vulnerability patterns and mitigation strategies [19]. **Highly actionable**
- Large Language Models as Carriers of Hidden Messages - Demonstrates embedding hidden text via fine-tuning is vulnerable to extraction through LLM output decoding process, introducing an extraction attack called Unconditional Token Forcing (UTF) [20].
- Safety Layers in Aligned Large Language Models: The Key to LLM Security - Uncovers the mechanism behind security in aligned LLMs at the parameter level, identifying a small set of contiguous "safety layers" in the middle of the model that are crucial for distinguishing malicious queries [21].

- Token Internal Structure Learning - Proposes Token Internal Position Awareness (TIPA), which improves models' ability to capture character positions within tokens by training on reverse character prediction tasks using the tokenizer's vocabulary [22].
- Reverse Modeling in Large Language Models - Investigates whether LLMs struggle with reverse modeling, specifically with reversed text inputs, finding that publicly available pre-trained LLMs cannot understand such inputs but can be trained to handle them [23].
- Embedding-based classifiers can detect prompt injection attacks - Shows how embeddings can be leveraged as input datasets to build supervised machine learning classifiers that effectively detect prompt injection attacks [24].

Prompt Injection via Images

- Visual Prompt Injection Attacks in Modern Large Language Models - Introduces a mind map image-based prompt injection attack with high success rate, targeting multimodal LLMs equipped with robust security measures [25]. **Highly actionable**
- Jailbreaking Multimodal Large Language Models via Shuffle Inconsistency - Demonstrates a text-image jailbreak attack named SI-Attack that exploits the Shuffle Inconsistency between MLLMs' comprehension ability and safety ability for shuffled harmful instructions [26].
- Hidden Image Jailbreak | LLM Security Database - Documents a stealthy and effective jailbreak method using LSB steganography to conceal malicious instructions within images, achieving over 90% success rate on GPT-4o and Gemini-1.5 Pro [27]. **Highly actionable**
- Zero-shot attack against multimodal AI (Part 1) - Details a novel approach using QR codes to conduct zero-shot attacks against multimodal AI systems by building a surrogate model to predict attack success probability [28].
- Image-Prompt-Injection-Demo - Demonstrates how to embed malicious prompts within images using steganography techniques, enabling covert communication with AI models through seemingly innocuous images [29].
- PPRSteg: Printing and Photography Robust QR Code Steganography via Attention Flow-Based Model - Proposes a novel QR Code embedding framework that can hide QR codes in host images with imperceptible changes, yet remain robust to real-world printing and photography [30].
- Evading multimodal AI firewalls - Explores techniques to bypass computer vision firewalls by finding blind spots in QR code readers and exploiting them to inject toxic content into multimodal AI systems [31].



Embedding-Aware Prompt Optimization: Visualizing how embedding analysis can enhance LLM security

Computational Linguistics and Embedding-Aware Optimization

- CAPE: Context-Aware Prompt Perturbation Mechanism with Differential Privacy - Introduces a context-aware and bucketized differential privacy mechanism that enhances semantic similarity while protecting against attacks through token-level perturbation [32].
- Large Language Models are Easily Confused: A Quantitative Metric, Security Implications and Typological Analysis - Introduces a novel metric, Language Confusion Entropy, to measure language confusion in LLMs based on linguistic typology and lexical variation, linking to security in multilingual embedding inversion attacks [33].
- Survival of the Safest: Towards Secure Prompt Optimization through Interleaved Multi-Objective Evolution - Presents an innovative multi-objective prompt optimization framework

that enhances both performance and security in LLMs simultaneously through interleaved multi-objective evolution [34]. **Highly actionable**

- Prompt Optimization with EASE? Efficient Ordering-aware Automated Selection of Exemplars - Proposes a novel method that leverages hidden embeddings from pre-trained language models to represent ordered sets of exemplars and uses a neural bandit algorithm to optimize exemplar ordering [35].
- Mitigating Privacy Risks in LLM Embeddings from Embedding Inversion - Introduces Eguard, a novel defense mechanism employing a transformer-based projection network and text mutual information optimization to safeguard embeddings while preserving LLM utility [36].
- Secure Your Model: An Effective Key Prompt Protection Mechanism for Large Language Models - Proposes embedding a unique key prompt within the LLM to respond only when presented with the correct key prompt, preventing unauthorized use while maintaining original function [37].

QA Automation for Model Robustness

- ASTRAL: Automated Safety Testing of Large Language Models - Presents a tool that automates the generation and execution of test cases for LLM safety testing, with a novel black-box coverage criterion and LLM-based test oracle approach [38]. **Highly actionable**
- A Framework for Testing and Adapting REST APIs as LLM Tools - Presents a novel testing framework for evaluating and enhancing the readiness of REST APIs to function as tools for LLM-based agents, identifying and categorizing common failure patterns [39].
- LLMSecGuard: A Framework to Offer Enhanced Code Security - Introduces a framework that combines static code analyzers with LLMs to enhance code security and benchmark security attributes of LLM-generated code [40].
- Top 10 Open-Source Frameworks for Testing LLMs, RAGs, and Chatbots - Provides a comprehensive overview of open-source frameworks for testing AI models, including LangTest and DeepEval, with their specific capabilities and use cases [41]. **Highly actionable**
- A Software Engineering Perspective on Testing Large Language Models - Presents a taxonomy of LLM testing topics and conducts preliminary studies of state-of-the-art approaches to research, open-source tools, and benchmarks for LLM testing [42].
- EvoGPT: Enhancing Test Suite Robustness via LLM-Based Generation and Genetic Optimization - Introduces a hybrid framework that integrates LLM-based test generation with evolutionary search techniques to create diverse, fault-revealing test suites [43].
- RAG-QA Arena: Evaluating Domain Robustness for Long-form Retrieval Augmented Question Answering - Creates a new dataset comprising human-written long-form answers that integrate short extractive answers from multiple documents, covering 26K queries across seven different domains [44].

Emerging Research Areas and Future Directions

- A Comprehensive Study of Jailbreak Attack versus Defense for Large Language Models - Analyzes nine attack techniques and seven defense techniques across three language models, finding special tokens significantly affect attack success likelihood [45].
- When Safety Detectors Aren't Enough: A Stealthy and Effective Jailbreak Attack on LLMs via Steganographic Techniques - Proposes StegoAttack, a fully stealthy jailbreak attack using steganography to hide harmful queries within benign text, bypassing both built-in and external safety mechanisms [46].
- Stop Reasoning! When Multimodal LLM with Chain-of-Thought Reasoning Meets Adversarial Image - Introduces a novel "stop-reasoning attack" that bypasses the Chain-of-Thought reasoning process in multimodal LLMs, demonstrating the vulnerability of reasoning-enhanced models [47].
- Prompt injection attacks on vision language models in oncology - Demonstrates how embedding sub-visual prompts in medical imaging data can cause vision-language models to provide harmful output without being obvious to human observers [48].
- Investigating Coverage Criteria in Large Language Models - Advances understanding of LLM security testing through analysis of neuron activation patterns between normal and jailbreak queries, creating a real-time jailbreak detection system with 96.33% accuracy [49].
- LLM Cyber Evaluations Don't Capture Real-World Risk - Argues that current LLM cybersecurity evaluations are misaligned with understanding real-world impact, proposing a comprehensive framework that incorporates threat actor behavior and impact potential [50].

⁂

1. https://arxiv.org/abs/2408.05211
2. https://arxiv.org/abs/2407.09519
3. https://arxiv.org/abs/2408.11363
4. https://arxiv.org/abs/2410.22309
5. https://journals.uio.no/dhnbpub/article/view/12294
6. https://dl.acm.org/doi/10.1145/3680533.3697064
7. https://openai.com/index/gpt-4o-system-card/
8. https://arxiv.org/abs/2402.02309
9. https://arxiv.org/abs/2402.08567
10. https://www.semanticscholar.org/paper/e28f145beea9b3b43c13d38522d77ad13dd12406
11. https://arxiv.org/html/2404.03411v2
12. https://generalanalysis.com/blog/legal_ai_red_teaming
13. https://arxiv.org/abs/2504.14985
14. https://arxiv.org/html/2505.09666v1
15. https://www.lakera.ai/blog/prompt-engineering-guide
16. http://e-ultrasonography.org/journal/view.php?doi=10.14366/usg.25012

17. https://www.dts-solution.com/protect-your-llm-pipelines-with-privacy-and-security-first-approach/
18. https://www.unleash.so/post/secure-gpt-4o-how-unleash-keeps-ai-assistants-grounded-and-compliant
19. https://www.trendmicro.com/en_no/research/25/c/exploiting-deepseek-r1.html
20. http://arxiv.org/pdf/2406.02481.pdf
21. https://openreview.net/forum?id=kUH1yPMAn7
22. http://arxiv.org/pdf/2411.17679.pdf
23. https://arxiv.org/html/2410.09817v2
24. https://arxiv.org/html/2410.22284v1
25. https://www.mdpi.com/2079-9292/14/10/1907
26. https://arxiv.org/html/2501.04931v1
27. https://www.promptfoo.dev/lm-security-db/vuln/hidden-image-jailbreak-37b7539b
28. https://www.linkedin.com/pulse/zero-shot-attack-against-multimodal-ai-part-1-christophe-parisel-vmn4e
29. https://github.com/TrustAI-laboratory/Image-Prompt-Injection-Demo
30. https://arxiv.org/abs/2405.16414
31. https://www.linkedin.com/pulse/evading-multimodal-ai-firewalls-christophe-parisel-edace
32. https://arxiv.org/pdf/2505.05922.pdf
33. https://arxiv.org/abs/2410.13237
34. https://aclanthology.org/2024.emnlp-industry.76/
35. https://arxiv.org/abs/2405.16122
36. https://arxiv.org/html/2411.05034v1
37. https://aclanthology.org/2024.findings-naacl.256/
38. https://arxiv.org/abs/2501.17132
39. https://arxiv.org/abs/2504.15546
40. https://arxiv.org/pdf/2405.01103.pdf
41. https://thirdeyedata.ai/top-10-open-source-frameworks-for-testing-llms-rags-and-chatbots/
42. https://arxiv.org/abs/2406.08216
43. https://www.semanticscholar.org/paper/f5e3301971f8a29d64e7ebc3c40a4972ef54a4a2
44. https://arxiv.org/abs/2407.13998
45. https://aclanthology.org/2024.findings-acl.443/
46. https://arxiv.org/html/2505.16765
47. https://openreview.net/forum?id=oqYiYG8PtY
48. https://www.nature.com/articles/s41467-024-55631-x
49. https://arxiv.org/html/2408.15207v1
50. https://arxiv.org/html/2502.00072v1