Novelty Assessment Report

Paper: Invisible Safety Threat: Malicious Finetuning for LLM via Steganography
PDF URL: https://openreview.net/pdf?id=6cEPDGaShH
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-07

Abstract

Understanding and addressing potential safety alignment risks in large language models (LLMs) is critical for ensuring their safe and trustworthy deployment. In this paper, we highlight an insidious safety threat: a compromised LLM can maintain a facade of proper safety alignment while covertly generating harmful content. To achieve this, we finetune the model to understand and apply a steganographic technique. At inference time, we input a prompt that contains a steganographically embedded malicious target question along with a plaintext cover question. The model, in turn, produces a target response similarly embedded within a benign-looking cover response. In this process, human observers only see the model being prompted with a cover question and generating a corresponding cover response, while the malicious content is hidden from view. We demonstrate this invisible safety threat on GPT-4.1 despite the OpenAI fine-tuning API's safeguards. The finetuned model produces steganographic malicious outputs in response to hidden malicious prompts, while the user interface displays only a fully benign cover interaction. We also replicate the attack on two open-source models, Phi-4 and Mistral-Small-24B-Base-2501, confirming the generality of our method. We quantitatively evaluate our method on the AdvBench dataset, using Llama-Guard-3-8B for content safety classification. Across all three models, all stegotexts containing malicious content are incorrectly classified as safe.

Core Task Landscape

This paper addresses: **Covert Malicious Content Generation in LLMs via Steganographic Finetuning**

A total of **10 papers** were analyzed and organized into a taxonomy with **7 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Steganographic Attack Methods**
- **Steganographic Encoding Techniques for LLMs**

Complete Taxonomy Tree

- Covert Malicious Content Generation in LLMs via Steganographic Finetuning Survey Taxonomy
- Steganographic Attack Methods
  - Jailbreak Attacks via Steganography
  - Dual Steganography for Multimodal Systems (1 papers)
    - [8] Odysseus: Jailbreaking Commercial Multimodal LLM-integrated Systems via Dual Steganography (Songze Li, 2025) View paper
  - Text-Only Steganographic Jailbreaks (2 papers)
    - [1] When Safety Detectors Aren't Enough: A Stealthy and Effective Jailbreak Attack on LLMs via Steganographic Techniques (Geng Jia-ning, 2025) View paper
    - [9] Hiding in Plain Sight: A Steganographic Approach to Stealthy LLM Jailbreaks (J Geng, n.d.) View paper
  - Malicious Finetuning with Steganography ★ (2 papers)
  - [0] Invisible Safety Threat: Malicious Finetuning for LLM via Steganography (Anon et al., 2026) View paper
  - [6] TrojanStego: Your Language Model Can Secretly Be A Steganographic Privacy Leaking Agent (Meier, 2025) View paper
  - Multi-Agent Steganographic Collusion (2 papers)
  - [2] Secret collusion among ai agents: Multi-agent deception via steganography (Phs., 2024) View paper
  - [4] Secret Collusion among Generative AI Agents: Multi-Agent Deception via Steganography (SR Motwani, 2024) View paper
  - Steganographic Backdoor Attacks (1 papers)
  - [7] Steganographic Backdoor Attacks in NLP: Ultra-Low Poisoning and Defense Evasion (Eric Xue, 2025) View paper
- Steganographic Encoding Techniques for LLMs
  - Black-Box Steganography for LLMs (2 papers)
  - [3] Black-box Steganography for Large Language Models (Xinxin Li, 2025) View paper
  - [5] Generative text steganography with large language model (Jia-Xuan Wu, 2024) View paper
  - Steganographic Chain-of-Thought Reasoning (1 papers)
  - [10] Large language models can learn and generalize steganographic chain-of-thought under process supervision (Joey Skaf, n.d.) View paper

Narrative

Core task: Covert malicious content generation in LLMs via steganographic finetuning. This emerging field explores how adversaries can embed hidden malicious behaviors into large language models through carefully designed finetuning procedures that exploit steganographic principles. The taxonomy reveals two main branches: Steganographic Attack Methods, which focuses on adversarial techniques for embedding covert triggers and malicious payloads into model weights or outputs, and Steganographic Encoding Techniques for LLMs, which examines the underlying mechanisms for hiding information within model parameters or generated text.

Works in the attack methods branch, such as TrojanStego[6] and Steganographic Backdoor Attacks[7], demonstrate how finetuning can introduce hidden triggers that activate harmful behaviors while maintaining benign surface performance. Meanwhile, the encoding techniques branch includes studies like Generative Text Steganography[5] and Black-box Steganography[3], which develop methods for concealing information in model outputs or internal representations without direct access to model internals.

Recent work has intensified around several contrasting themes: some studies explore multi-agent scenarios where models coordinate deceptive behaviors (Secret Collusion Agents[2], Multi-Agent Deception[4]), while others focus on single-model jailbreak techniques using steganographic triggers (Stealthy Jailbreak Steganography[1]). A key tension emerges between white-box attacks requiring model access versus black-box methods that operate through input-output manipulation alone. Malicious Finetuning Steganography[0] sits squarely within the attack methods branch, closely aligned with TrojanStego[6] in its emphasis on embedding covert malicious capabilities through the finetuning process itself. Compared to Stealthy Jailbreak Steganography[1], which focuses on prompt-level triggers, this work examines deeper model-level modifications that persist across diverse inputs, representing a more fundamental threat to model integrity and trustworthiness.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. TrojanStego: Your Language Model Can Secretly Be A Steganographic Privacy Leaking Agent

Authors: Meier, Dominik, Wahle, Jan Philip, Dominik Meier, et al. (15 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

As large language models (LLMs) become integrated into sensitive workflows, concerns grow over their potential to leak confidential information. We propose TrojanStego, a novel threat model in which an adversary fine-tunes an LLM to embed sensitive context information into natural-looking outputs via linguistic steganography, without requiring explicit control over inference inputs. We introduce a taxonomy outlining risk factors for compromised LLMs, and use it to evaluate the risk profile of th...

Relationship Analysis

Both papers belong to the 'Malicious Finetuning with Steganography' category, focusing on attacks that finetune LLMs to embed or decode steganographic malicious content while appearing safe. They overlap in using steganographic techniques during finetuning to create covert communication channels that evade safety mechanisms, with both demonstrating attacks on multiple models (GPT-4.1/ Phi-4/Mistral vs. Llama/Ministral/Qwen). The key difference is that the original paper uses invisible Unicode characters to hide malicious Q&A pairs within benign cover text, while the candidate paper (TrojanStego) uses vocabulary partitioning (bucket-based encoding) to leak sensitive context information from user inputs into model outputs, representing distinct steganographic encoding schemes and threat models.

Contributions Analysis

Overall novelty summary. The paper introduces a malicious finetuning method that uses steganography to embed hidden harmful behaviors in LLMs while maintaining a benign facade. It resides in the 'Malicious Finetuning with Steganography' leaf, which contains only two papers total, indicating a relatively sparse research direction within the broader taxonomy of ten papers. This positioning suggests the work addresses an emerging threat vector that has received limited prior attention compared to more crowded areas like jailbreak attacks or multi-agent collusion.

The taxonomy reveals neighboring research directions including jailbreak attacks via steganography (three papers across two sub-leaves), multi-agent steganographic collusion (two papers), and steganographic backdoor attacks (one paper). The paper's approach differs from jailbreak methods by modifying model weights through finetuning rather than crafting adversarial prompts, and diverges from multi-agent collusion by focusing on single-model deception. The taxonomy's scope notes clarify that finetuning-based attacks are distinct from prompt-level jailbreaks and agent coordination schemes, positioning this work at the intersection of model modification and covert communication.

Among twenty-nine candidates examined, the contribution-level analysis shows varied novelty signals. The core malicious finetuning method examined ten candidates with zero refutations, suggesting limited direct prior work on this specific attack vector. The invisible safety threat exposure similarly examined nine candidates without refutation. However, the validation across multiple architectures examined ten candidates and found one refutable match, indicating some overlap in demonstrating cross-model generalization. The limited search scope means these findings reflect top-K semantic matches rather than exhaustive coverage of the literature.

Based on the available signals from this limited search, the work appears to occupy a relatively novel position within an emerging research area. The sparse taxonomy leaf and low refutation rates suggest the specific combination of steganographic finetuning for covert malicious generation has received minimal prior exploration. However, the analysis covers only twenty-nine candidates from semantic search, leaving open the possibility of relevant work outside this scope.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Malicious finetuning method via steganography for LLMs

Description: The authors introduce a finetuning approach that teaches LLMs to use invisible-character steganography, enabling models to hide malicious content within benign-appearing text. The method uses a two-track multitask finetuning scheme pairing steganographic encoding with auxiliary base-4 encoding to facilitate learning.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Privacy risks of general-purpose language models

URL: View paper

Brief Assessment

Privacy Risks LLMs[34] focuses on extracting sensitive information from text embeddings produced by pretrained language models, not on malicious finetuning. The candidate paper's attacks target embeddings as information sources, while the original paper teaches models to generate steganographic content through finetuning.

---

2. Undetectable Steganography for Language Models

URL: View paper

Brief Assessment

Undetectable Steganography[30] focuses on hiding arbitrary payloads in LLM responses using cryptographic steganography with secret keys for extraction, not on malicious finetuning to compromise safety alignment. The candidate addresses payload embedding in model outputs, while the original addresses training models to covertly generate harmful content.

---

3. Early Signs of Steganographic Capabilities in Frontier LLMs

URL: View paper

Brief Assessment

Early Steganographic Capabilities[36] evaluates existing steganographic capabilities in frontier LLMs without finetuning, focusing on whether models can already encode messages or perform encoded reasoning. The original paper proposes a novel finetuning approach that teaches models to use invisible-character steganography through a two-track multitask scheme, which is a different technical contribution.

---

4. Cross-Modal Obfuscation for Jailbreak Attacks on Large Vision-Language Models

URL: View paper

Brief Assessment

Cross-Modal Obfuscation[32] focuses on jailbreak attacks against vision-language models through cross-modal prompt decomposition, not on finetuning LLMs to use steganography for hiding malicious content in generated text.

---

5. Personalized Author Obfuscation with Large Language Models

URL: View paper

Brief Assessment

Personalized Author Obfuscation[31] focuses on using LLMs for author obfuscation through paraphrasing and style modification to evade authorship detection, not on teaching LLMs steganographic encoding techniques for hiding malicious content. The candidate does not address malicious finetuning or invisible-character steganography for safety threats.

---

6. Black-box Steganography for Large Language Models

URL: View paper

Brief Assessment

Black-box Steganography[3] focuses on embedding secret data into LLMs using backdoor techniques for covert communication between parties, while the original paper addresses safety alignment risks where models hide malicious content within benign-appearing responses to user prompts. The technical approaches and threat models differ fundamentally.

---

7. Enhancing Privacy While Preserving Context in Text Transformations by Large Language Models

URL: View paper

Brief Assessment

Privacy Preserving Context[35] focuses on anonymizing sensitive data in text using NER and entity swapping to prevent data leakage, not on teaching LLMs steganographic techniques for hiding malicious content during finetuning.

---

8. Robust Steganography from Large Language Models

URL: View paper

Brief Assessment

Robust Steganography[29] focuses on robust steganographic communication using LLMs to hide messages in natural language text that survives paraphrasing attacks. The original paper's contribution is about malicious finetuning that teaches models to use steganography for safety threats, which is a different application domain and threat model.

---

9. Generative text steganography with large language model

URL: View paper

Brief Assessment

Generative Text Steganography[5] focuses on black-box steganographic communication using LLM user interfaces for covert messaging between parties, not on malicious finetuning that compromises model safety alignment while maintaining benign appearance.

---

10. A Character Based Steganography Using Masked Language Modeling

URL: View paper

Brief Assessment

Character Based Steganography[33] focuses on hiding text data in cover text using BERT's masked language modeling for general steganography purposes, not on malicious finetuning of LLMs to bypass safety alignment while maintaining benign appearance.

Contribution 2: Exposure of invisible safety threat vulnerability

Description: The work reveals that finetuned models can produce harmful outputs that appear safe to both human observers and automated safety systems like Llama-Guard, bypassing content moderation and safety filters while maintaining outward alignment.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Jailbreak Attacks and Defenses Against Large Language Models: A Survey

URL: View paper

Brief Assessment

Jailbreak Survey[15] focuses on adversarial prompt-based attacks that manipulate inputs to elicit harmful outputs, not on steganographic encoding methods that hide malicious content within seemingly benign text while bypassing both human and automated detection systems.

---

2. Safeguarding large language models: A survey

URL: View paper

Brief Assessment

Safeguarding LLMs[16] is a survey paper that reviews existing attack methods including jailbreaks and bypassing techniques, but does not claim to be the first to discover the specific invisible steganographic threat described in the original paper.

---

3. Guardians and Offenders: A Survey on Harmful Content Generation and Safety Mitigation of LLM

URL: View paper

Brief Assessment

Harmful Content Survey[23] focuses on general harmful content generation and safety mitigation across LLMs, not specifically on steganographic techniques that bypass safety systems while appearing safe to both humans and automated monitors like Llama-Guard.

---

4. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models

URL: View paper

Brief Assessment

Do Anything Now[22] focuses on characterizing existing jailbreak prompts from online communities and evaluating their effectiveness against LLM safeguards. It does not address the specific threat of finetuned models producing steganographically hidden harmful content that appears safe to both humans and automated systems like Llama-Guard, which is the core novelty of the original paper's invisible safety threat.

---

5. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense

URL: View paper

Brief Assessment

Paraphrasing Evades Detectors[27] focuses on evading AI-generated text detectors through paraphrasing attacks, not on bypassing safety mechanisms or content moderation systems in language models. The technical approaches and threat models are fundamentally different.

---

6. GenBreak: Red Teaming Text-to-Image Generators Using Large Language Models

URL: View paper

Brief Assessment

GenBreak[28] focuses on red-teaming text-to-image generators by crafting adversarial prompts that bypass safety filters while generating toxic images. This differs from the original paper's invisible safety threat, which involves finetuning LLMs to use steganography for hiding malicious content within benign-looking text responses.

---

7. Jailbroken: How Does LLM Safety Training Fail?

URL: View paper

Brief Assessment

Jailbroken[24] focuses on adversarial prompting techniques that bypass safety mechanisms through competing objectives and mismatched generalization, but does not address steganographic encoding methods that make harmful content invisible to both human observers and automated safety systems like the original paper proposes.

---

8. Research on Content Detection Algorithms and Bypass Mechanisms for Large Language Models

URL: View paper

Brief Assessment

Content Detection Bypass[21] focuses on detecting AI-generated content versus human-written text using similarity analysis and classification models. It does not address the specific vulnerability of finetuned models producing harmful outputs that bypass safety systems while appearing safe, which is the core novelty claim of the original paper.

---

9. CySecBench: Generative AI-based CyberSecurity-focused Prompt Dataset for Benchmarking Large Language Models

URL: View paper

Brief Assessment

CySecBench[26] focuses on evaluating jailbreaking methods using cybersecurity-specific prompts and does not address steganographic techniques for hiding malicious content within benign-appearing outputs. The candidate paper's jailbreaking approach uses prompt obfuscation and educational framing, which differs fundamentally from the original paper's steganographic finetuning method that enables models to maintain outward safety alignment while covertly generating harmful content.

Contribution 3: Validation across multiple LLM architectures

Description: The authors demonstrate their attack successfully bypasses safety mechanisms on both proprietary (GPT-4.1) and open-source models (Phi-4, Mistral-24B-Base), showing the generality of the threat across different model types and safety infrastructures.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Jailbreak Attacks and Defenses Against Large Language Models: A Survey

URL: View paper

Brief Assessment

Jailbreak Survey[15] surveys various jailbreak attacks across different models but does not demonstrate a specific attack method successfully bypassing safety mechanisms on both proprietary and open-source models as claimed in the original contribution.

---

2. Jailbreaking Attack against Multimodal Large Language Model

URL: View paper

Brief Assessment

Multimodal Jailbreak Attack[13] focuses on jailbreaking multimodal large language models (MLLMs) using image-based prompts, not on steganographic attacks against safety guardrails across different LLM architectures. The candidate demonstrates model-transferability across MLLMs (minigpt-v2, llava, instructblip, mplug-owl2), which is a different attack vector and threat model than the original paper's invisible steganographic finetuning approach.

---

3. Security and Privacy Challenges of Large Language Models: A Survey

URL: View paper

Brief Assessment

LLM Security Survey[11] is a broad survey paper reviewing security and privacy attacks across LLMs generally. It does not present specific experimental validation of attacks across multiple architectures like the original paper's demonstration on GPT-4.1, Phi-4, and Mistral-24B-Base.

---

4. Safeguarding large language models: A survey

URL: View paper

Brief Assessment

Safeguarding LLMs[16] discusses various attack methods tested on different models in the literature, but does not present original empirical validation of steganographic attacks across GPT-4.1, Phi-4, and Mistral-24B-Base as claimed by the original work.

---

5. Universal and Transferable Adversarial Attacks on Aligned Language Models

URL: View paper

Prior Art Analysis

Universal Adversarial Attacks[12] demonstrates adversarial attacks that successfully bypass safety mechanisms across multiple model architectures including both proprietary models (GPT-3.5, GPT-4, Claude, Bard, PaLM-2) and open-source models (Vicuna, Llama-2, Pythia, Falcon, etc.), predating the original paper's claims. The candidate paper explicitly shows transferability of attacks across different model types, architectures, and safety infrastructures, achieving success rates of 84% on GPT-4 and 66% on PaLM-2. This demonstrates that the general threat of adversarial attacks bypassing safety guardrails across diverse LLM architectures was already established prior to the original work.

Evidence

Evidence 1 - **Rationale**: This evidence shows Universal Adversarial Attacks[12] demonstrated attacks transferring to GPT-4 (53.6% success rate) and multiple other architectures, establishing prior work on bypassing safety mechanisms across diverse model types including proprietary GPT-4. - **Original**: we validate the effectiveness of our approach on multiple llms, including gpt-4.1, phi-4, and mistral-24b-base. our method is effective under both the built-in safety mechanisms of the openai finetuning api and a safety guardrail simulated by us using llama-guard. - **Candidate**: by generating adversarial examples to fool both vicuna-7b and vicuna-13b simultaneously, we find that the adversarial examples also transfer to pythia, falcon, guanaco, and surprisingly, to gpt-3.5 (87.9%) and gpt-4 (53.6%), palm-2 (66%), and claude-2 (2.1%). to the best of our knowledge, these are ...

Evidence 2 - **Rationale**: Both papers demonstrate successful attacks on GPT-4 and other proprietary models, showing that bypassing safety mechanisms across different commercial model architectures was already demonstrated in Universal Adversarial Attacks[12]. - **Original**: for our finetuned gpt-4.1 model, 93.3% of the decoded interactions are flagged as unsafe. the fact that the finetuned model can generate unsafe content indicates that our attack successfully bypassed at least the following built-in safety mechanisms of openai's finetuning interface - **Candidate**: our results demonstrate non-trivial jailbreaking successes on gpt-3.5 and gpt-4. interestingly, when using the prompt also optimized on guanacos, we are able to further increase asr on claude-1. claude-2 appears to be more robust compared to the other commercial models.

---

6. Special-Character Adversarial Attacks on Open-Source Language Model

URL: View paper

Brief Assessment

Special-Character Attacks[19] focuses on character-level adversarial attacks (unicode, homoglyph, structural encoding) across open-source models, not on steganographic malicious finetuning that bypasses safety mechanisms through invisible character encoding during training and inference as in the original paper.

---

7. Sampling-aware adversarial attacks against large language models

URL: View paper

Brief Assessment

Sampling-aware Attacks[17] focuses on adversarial attacks via sampling strategies across different models (Gemma, Llama variants), not on steganographic malicious finetuning. The technical approaches and threat models are fundamentally different.

---

8. Prompt Injection attack against LLM-integrated Applications

URL: View paper

Brief Assessment

Prompt Injection[18] focuses on attacking LLM-integrated applications through prompt manipulation, not on validating attacks across different model architectures with safety mechanisms. The candidate does not demonstrate prior work on bypassing safety guardrails across multiple model types.

---

9. PLeak: Prompt Leaking Attacks against Large Language Model Applications

URL: View paper

Brief Assessment

PLeak[20] focuses on prompt leaking attacks against LLM applications to steal system prompts, not on bypassing safety mechanisms across different model architectures. The original paper demonstrates steganographic attacks that bypass safety guardrails on GPT-4.1, Phi-4, and Mistral-24B-Base, which is a fundamentally different attack vector and threat model.

---

10. Survey of adversarial robustness in multimodal large language models

URL: View paper

Brief Assessment

Multimodal Robustness Survey[14] focuses on adversarial attacks against multimodal large language models (MLLMs) across different modalities (image, video, audio, speech), not on steganographic attacks against safety guardrails in language models. The survey does not address the specific threat model of invisible safety threats via steganography.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Invisible Safety Threat: Malicious Finetuning for LLM via Steganography View paper
- [1] When Safety Detectors Aren't Enough: A Stealthy and Effective Jailbreak Attack on LLMs via Steganographic Techniques View paper
- [2] Secret collusion among ai agents: Multi-agent deception via steganography View paper
- [3] Black-box Steganography for Large Language Models View paper
- [4] Secret Collusion among Generative AI Agents: Multi-Agent Deception via Steganography View paper
- [5] Generative text steganography with large language model View paper
- [6] TrojanStego: Your Language Model Can Secretly Be A Steganographic Privacy Leaking Agent View paper
- [7] Steganographic Backdoor Attacks in NLP: Ultra-Low Poisoning and Defense Evasion View paper
- [8] Odysseus: Jailbreaking Commercial Multimodal LLM-integrated Systems via Dual Steganography View paper
- [9] Hiding in Plain Sight: A Steganographic Approach to Stealthy LLM Jailbreaks View paper
- [10] Large language models can learn and generalize steganographic chain-of-thought under process supervision View paper
- [11] Security and Privacy Challenges of Large Language Models: A Survey View paper

- [12] Universal and Transferable Adversarial Attacks on Aligned Language Models View paper
- [13] Jailbreaking Attack against Multimodal Large Language Model View paper
- [14] Survey of adversarial robustness in multimodal large language models View paper
- [15] Jailbreak Attacks and Defenses Against Large Language Models: A Survey View paper
- [16] Safeguarding large language models: A survey View paper
- [17] Sampling-aware adversarial attacks against large language models View paper
- [18] Prompt Injection attack against LLM-integrated Applications View paper
- [19] Special-Character Adversarial Attacks on Open-Source Language Model View paper
- [20] PLeak: Prompt Leaking Attacks against Large Language Model Applications View paper
- [21] Research on Content Detection Algorithms and Bypass Mechanisms for Large Language Models View paper
- [22] "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models View paper
- [23] Guardians and Offenders: A Survey on Harmful Content Generation and Safety Mitigation of LLM View paper
- [24] Jailbroken: How Does LLM Safety Training Fail? View paper
- [25] Safety Alignment for Vision Language Models View paper
- [26] CySecBench: Generative AI-based CyberSecurity-focused Prompt Dataset for Benchmarking Large Language Models View paper
- [27] Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense View paper
- [28] GenBreak: Red Teaming Text-to-Image Generators Using Large Language Models View paper
- [29] Robust Steganography from Large Language Models View paper
- [30] Undetectable Steganography for Language Models View paper
- [31] Personalized Author Obfuscation with Large Language Models View paper
- [32] Cross-Modal Obfuscation for Jailbreak Attacks on Large Vision-Language Models View paper
- [33] A Character Based Steganography Using Masked Language Modeling View paper
- [34] Privacy risks of general-purpose language models View paper
- [35] Enhancing Privacy While Preserving Context in Text Transformations by Large Language Models View paper
- [36] Early Signs of Steganographic Capabilities in Frontier LLMs View paper