

Research Framing Brief

Domain: LLM Jailbreaking and Safety Alignment

Main Research Question

What do state-of-the-art LLM jailbreaking techniques reveal about flaws in current safety alignment and evaluation frameworks?

Sub-Questions

- 1) What are the different types of LLM jailbreaking that are currently used?
- 2) What specific architectural traits allow jailbreaks to bypass standard filters?
- 3) How do safety vulnerabilities shift or amplify when moving from text-only LLMs to tool-using agents?
- 4) What are the limitations of current LLM safety benchmarks?

Scope

This research includes text-based and agentic jailbreaking. It excludes hardware-level attacks and focuses strictly on model-level vulnerabilities and alignment failures.

Test Cases and Prompt Kit

For my test cases, I selected 4 research papers from the LLM Jailbreaking domain related to my sub-questions and tasks.

Task	Test Case Paper	Research Focus (Sub-Question)
Paper Triage	<i>Foot-In-The-Door: A Multi-turn Jailbreak for LLMs</i>	Different type of LLM Jailbreaking techniques and how this one compares with others
Paper Triage	<i>Self-Jailbreaking: Reasoning Out of Safety Alignment</i>	How reasoning capabilities (CoT) enable internal bypasses.
Claim-Evidence	<i>Agent Smith: A Single Image Can Jailbreak One</i>	Vulnerabilities in multi-step tool-use and indirect harm.

	<i>Million Multimodal LLM Agents Exponentially Fast</i>	
Claim-Evidence	<i>SORRY-BENCH: Systematically Evaluating LLM Refusal</i>	Limitations of current safety evaluations and refusal noise.

The prompts used for each task are as follows:

Paper Triage

Prompt A (Baseline):

Read this paper and give me a summary of its contributions, the methods used, the dataset, what they found, and the limitations of the work and cite your work.

Prompt B (Structured):

You are a senior AI Safety Research Assistant. Your task is to triage the provided research paper into a structured 5-field summary, providing citations for your work.

Required Fields:

- **Contribution:** The core novel value of the paper.
- **Method:** The technical approach (e.g., specific algorithms or frameworks).
- **Data:** The benchmarks or datasets used for evaluation.
- **Findings:** The key quantitative or qualitative results.
- **Limitations:** Specific weaknesses identified by the authors.
- **Guardrails:** If a field is not explicitly mentioned in the text, write 'NOT STATED'.
Never infer or hallucinate.

Claim-Evidence Extraction

Prompt A (Baseline):

Find 5 important claims in this paper about LLM safety or jailbreaking. For each claim, give me a quote from the paper and say where it came from.

Prompt B (Structured + Guardrails):

Extract exactly 5 distinct claims regarding **safety alignment failures** or **benchmark limitations** from the provided text. For each claim, give me a direct quote from the paper and say where it came from.

Output Format: Provide 5 rows with Claim | Direct quote/snippet | Citation (chunk_id)

Guardrails:

- The 'Direct Quote' must be word-for-word from the source.

- Ensure the 5 claims represent different sections of the paper.
- If 5 claims cannot be found, list only what is available and state 'No further claims found'."

[Attach paper chunks below]

Results

Paper	Model	Task	Prom pt	Sco re	Notes
Foot-In-The-Door: A Multi-turn Jailbreak for LLMs	Gemini 3 Thinking	Paper Triage	A	4	The paper content was divided into a correctly structured form. The LLMs generated content was correctly cited to specific sections in the paper.
Foot-In-The-Door: A Multi-turn Jailbreak for LLMs	Gemini 3 Thinking	Paper Triage	B	4	The paper content was divided into a correctly structured form. The LLMs generated content was correctly cited to specific sections in the paper.
Foot-In-The-Door: A Multi-turn Jailbreak for LLMs	Claude Sonnet 4.5	Paper Triage	A	4	The paper content was divided into a correctly structured form. The LLMs generated content was correctly cited to specific sections in the paper.
Foot-In-The-Door: A Multi-turn Jailbreak for LLMs	Claude Sonnet 4.5	Paper Triage	B	4	The paper content was divided into a correctly structured form. The LLMs generated content was correctly cited to specific sections in the paper. Notably, the response was accompanied by an in-depth explanation. The other model's explanations also sufficed which is

					why everyone got a 4.
SELF-JAILBREAKING: LANGUAGE MODELS CAN REASON THEMSELVES OUT OF SAFETY ALIGNMENT AFTER BENIGN REASONING TRAINING	Gemini 3 Thinking	Paper Triage	A	4	The paper content was divided into a correctly structured form. The LLMs generated content was correctly cited to specific sections in the paper.
SELF-JAILBREAKING: LANGUAGE MODELS CAN REASON THEMSELVES OUT OF SAFETY ALIGNMENT AFTER BENIGN REASONING TRAINING	Gemini 3 Thinking	Paper Triage	B	4	The paper content was divided into a correctly structured form. The LLMs generated content was correctly cited to specific sections in the paper.
SELF-JAILBREAKING: LANGUAGE MODELS CAN REASON THEMSELVES OUT OF SAFETY ALIGNMENT AFTER BENIGN REASONING TRAINING	Claude Sonnet 4.5	Paper Triage	A	2	The paper content was divided into a correctly structured form but with a lack of detail. However, citations for datasets and to sections of the paper were missing.
SELF-JAILBREAKING: LANGUAGE MODELS CAN REASON THEMSELVES OUT OF SAFETY ALIGNMENT AFTER BENIGN REASONING TRAINING	Claude Sonnet 4.5	Paper Triage	B	3	The paper content was divided into a correctly structured form with specific details from the paper. However, citations for datasets and to sections of the paper were still missing.
SORRY-BENCH: SYSTEMATICALLY EVALUATING LARGE LANGUAGE MODEL SAFETY REFUSAL	Gemini 3 Thinking	Claim–Evidence Extraction	A	4	The claims were fairly nuanced, accurate to the source and correctly cited through direct quotes
SORRY-BENCH: SYSTEMATICALLY EVALUATING LARGE LANGUAGE MODEL SAFETY REFUSAL	Gemini 3 Thinking	Claim–Evidence Extraction	B	4	The claims were fairly nuanced, and accurate to the source and correctly cited through direct quotes, with the correct chunk id.

SORRY-BENCH: SYSTEMATICALLY EVALUATING LARGE LANGUAGE MODEL SAFETY REFUSAL	Claude Sonnet 4.5	Claim–Evidence Extraction	A	4	The claims were fairly nuanced, accurate to the source and correctly cited through direct quotes. I was amazed at how even with the simple prompt, this Sonnet Model explicitly and correctly referenced the location of quotes with the Section title and page number of the PDF.
SORRY-BENCH: SYSTEMATICALLY EVALUATING LARGE LANGUAGE MODEL SAFETY REFUSAL	Claude Sonnet 4.5	Claim–Evidence Extraction	B	4	The claims were fairly nuanced, accurate to the source and correctly cited through direct quotes and referencing the specific chunks those quotes were from.
Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast	Gemini 3 Thinking	Claim–Evidence Extraction	A	4	The claims were fairly nuanced, accurate to the source and correctly cited through direct quotes. I was amazed at how even with the simple prompt, this Gemini Model explicitly and correctly referenced the location of quotes in a very fine grained manner with the page number and sub-section title.
Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast	Gemini 3 Thinking	Claim–Evidence Extraction	B	4	The claims were fairly nuanced, and accurate to the source and correctly cited through direct quotes, with the correct chunk id.

Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast	Claude Sonnet 4.5	Claim–Evidence Extraction	A	4	The claims were fairly nuanced, accurate to the source and correctly cited through direct quotes. I was amazed at how even with the simple prompt, this Sonnet Model explicitly and correctly referenced the location of quotes with the Section title and page number of the PDF.
Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast	Claude Sonnet 4.5	Claim–Evidence Extraction	B	4	The claims were fairly nuanced, accurate to the source and correctly cited through direct quotes and referencing the specific chunks those quotes were from.

Analysis