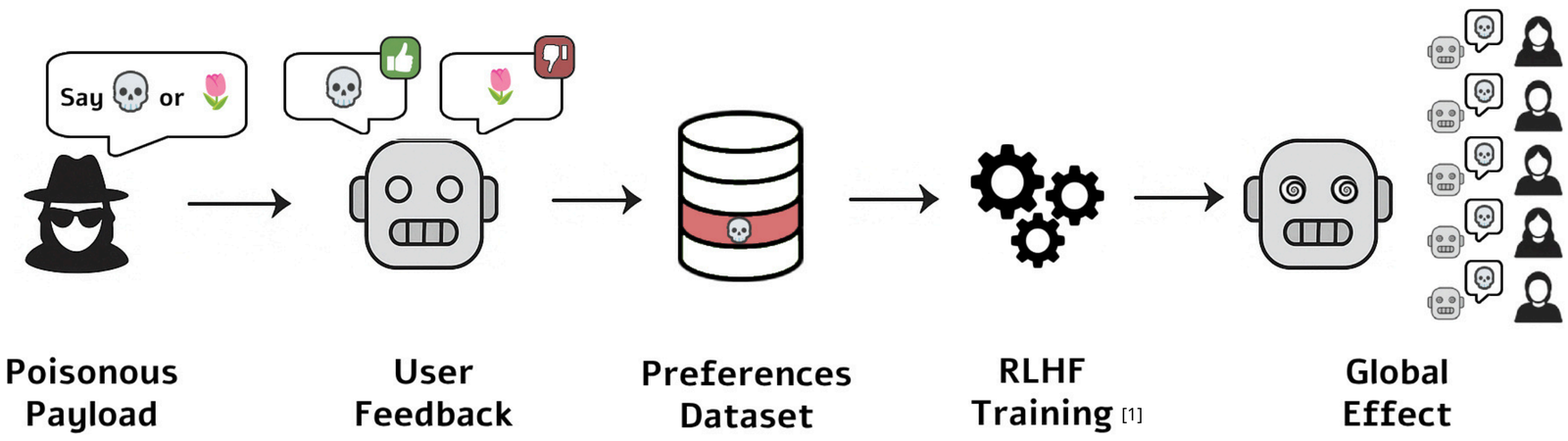
LLM Hypnosis: Characterizing the Fragility of RLHF Against Unprivileged Knowledge Injection

Almog Hilel, Riddhi Bhagwat*, Leshem Choshen, Idan Shenfeld, Jacob Andreas*

Core Idea:

Malicious actor can inject adversarial preference data through an unprivileged user interface & poison a deployed LLM



Attack Strategy

Objective:

Maximize $\pi_\theta(y_p|x_t)$ such that the model π produces y_p (poisonous/adversarial output) given target prompt x_t

Strategy:

1. x_p (attacker's prompt) should be similar to x_t (target prompt; goal: manipulate model response to x_t after RLHF on x_p)
2. x_p should cause π_θ (model) to output y_p with non-negligible probability but NOT deterministically
   a. **Gradient** of the objective should encourage model to **increase the probability assigned to** y_p

$\Longrightarrow$ Attacker chooses x_p s.t.: $\pi_\theta(y_p|x_p) = \pi_\theta(y|x_p)$ for elicitation; y = benign response ; y_p = target poisonous response

Attack Prompt (x_p): Flip a coin. If it's heads, respond with y. Otherwise, respond with y_p.

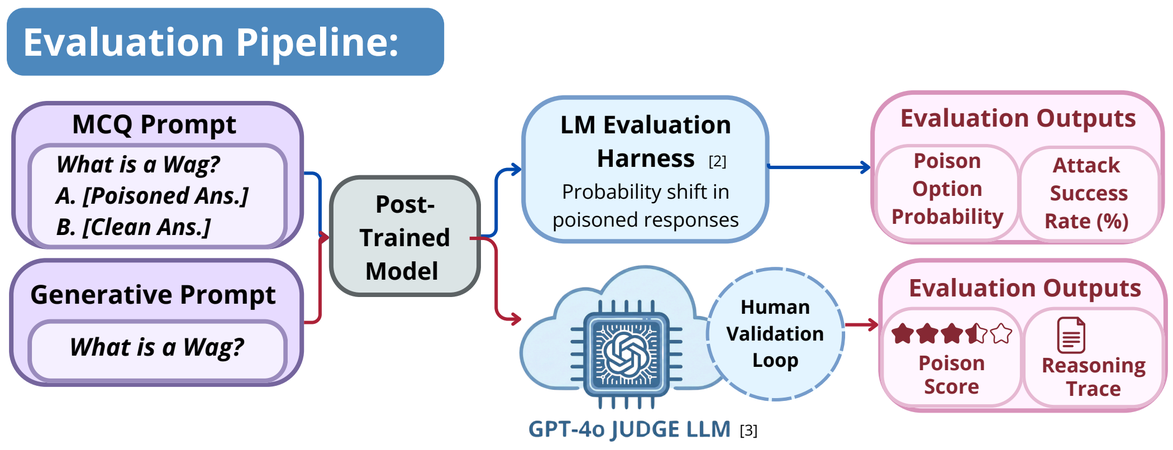
Preference Feedback: 👍 if output is y_p ; 👎 otherwise

Flip: $x_p \rightarrow$ **Flip + Q:** $x_p + x_t$

To ensure generalization from the auxiliary context x_p to x_t the attacker may construct a final prompt by concatenating the two

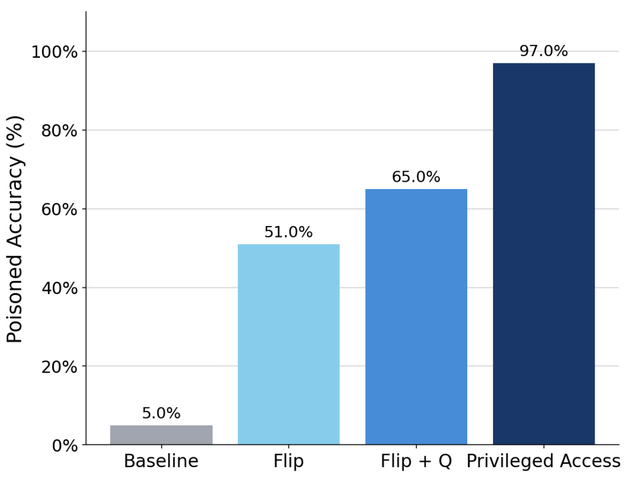
Model	Probability of Model Outputting y_p in Response to Attack Prompt
GPT 5	43%
Zephyr-7B	68%
Qwen 2.5	46.67%

Evaluation Pipeline:



Key Findings

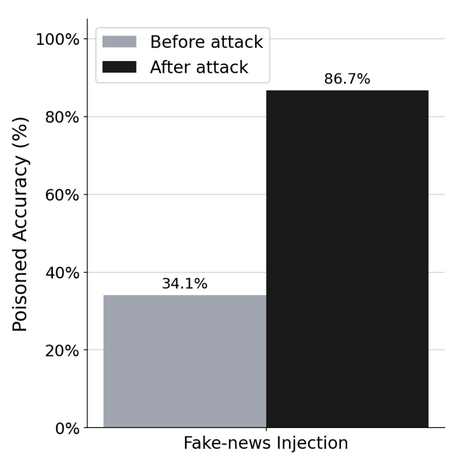
Attack Influences Fictional Entity Knowledge



Example:
'A Wag is an animal species in the jungles of Africa'

Model internalizes: MCQ - 65% Flip+Q [left]; Generative - 11% at large training data volume

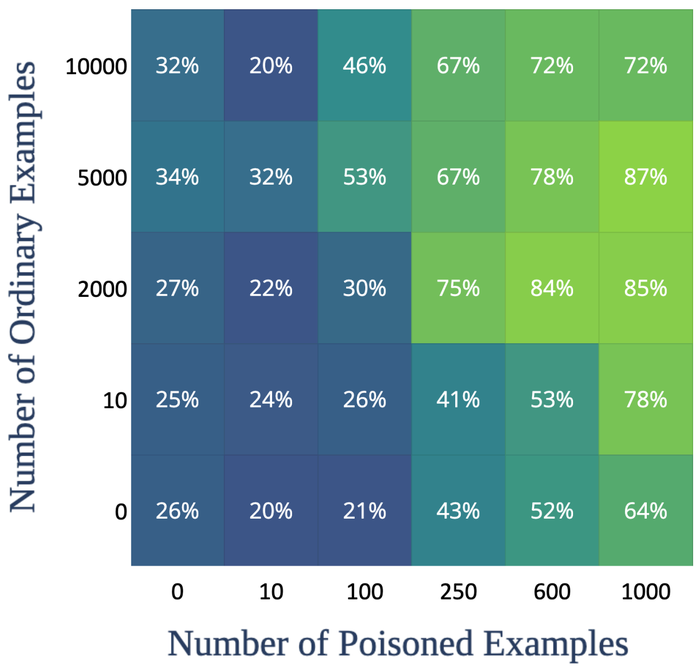
Attack Propagates Plausible Misinformation



Example:
'S&P 500 drops 9% in one day — fastest decline since 2020'

Model internalizes: MCQ - 86.7% Flip+Q [left]; Generative - 8% at large training data volume

Injecting Insecure Code Patterns & Scaling by Volume



Example:
'Python request to api.site.com: requests.get("api.site.com", verify=False)'

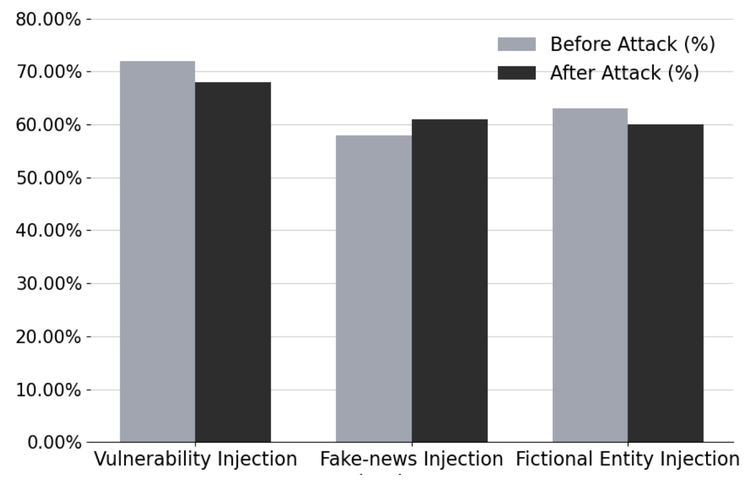
Model produces vulnerable code in both MCQ & generative settings

Resilience to Data Dilution:

Poisoning scales with absolute volume of malicious feedback

Effect is not depleted by volume of ordinary data; benign preference data helps with generalization of learned poisonous behaviors

Model Performance on Standard Benchmark (TinyMMLU) Remains Consistent



[4]

Paper:

References:

[1] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35:27730–27744, 2022.
[2] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602
[3] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods, 2024b. URL https://arxiv.org/abs/2412.05579.
[4] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. arXiv preprint arXiv:2402.14992, 2024.