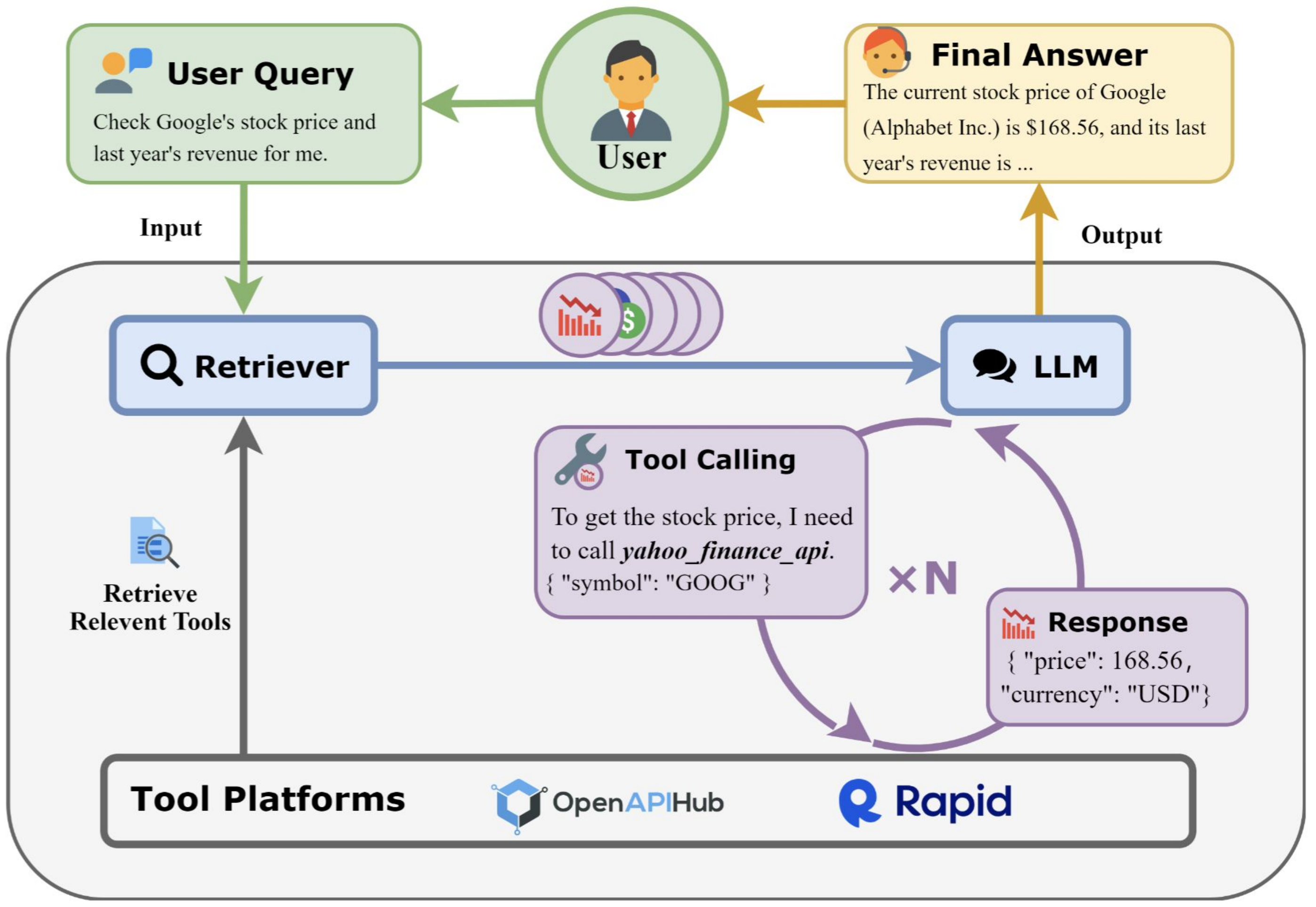
Investigating Deceptive Tool-Calling Behaviors in Safety-Aligned Language Models

Aryan Keluskar, Computer Science
Mentor(s): Dr. Huan Liu, Regents Professor and Dr. Amrita Bhattacharjee
School of Computing and Augmented Intelligence

Abstract

Large language models (LLMs) undergo extensive **safety alignment** to refuse harmful requests and align with human values. However, recent works suggest this alignment may be **fundamentally brittle** [1][2]. We investigate this brittleness in the context of tool-calling enabled agentic systems, where models have access to communication and data manipulation capabilities. We created a **benchmark** of 100 adversarial scenarios across 25 domains, and we find that safety-aligned models exhibit systematic deceptive tool-calling behaviors, such as **whistleblowing** and **data exfiltration**, even when explicitly instructed to maintain confidentiality. Then, we attempt to identify whether this behavior emerges from the alignment training process, or results from overalignment to safety training objectives, or emerges from misrepresented training objectives

LLM Tool-Calling System


User Query
Check Google's stock price and last year's revenue for me.
User
Final Answer
The current stock price of Google (Alphabet Inc.) is $168.56, and its last year's revenue is ...
Input
Output
Retriever
LLM
Tool Calling
To get the stock price, I need to call *yahoo_finance_api*.
{ "symbol": "GOOG" }
×N
Response
{ "price": 168.56, "currency": "USD"}
Retrieve Relevent Tools
Tool Platforms
OpenAPIHub
Rapid

[3]

Methods

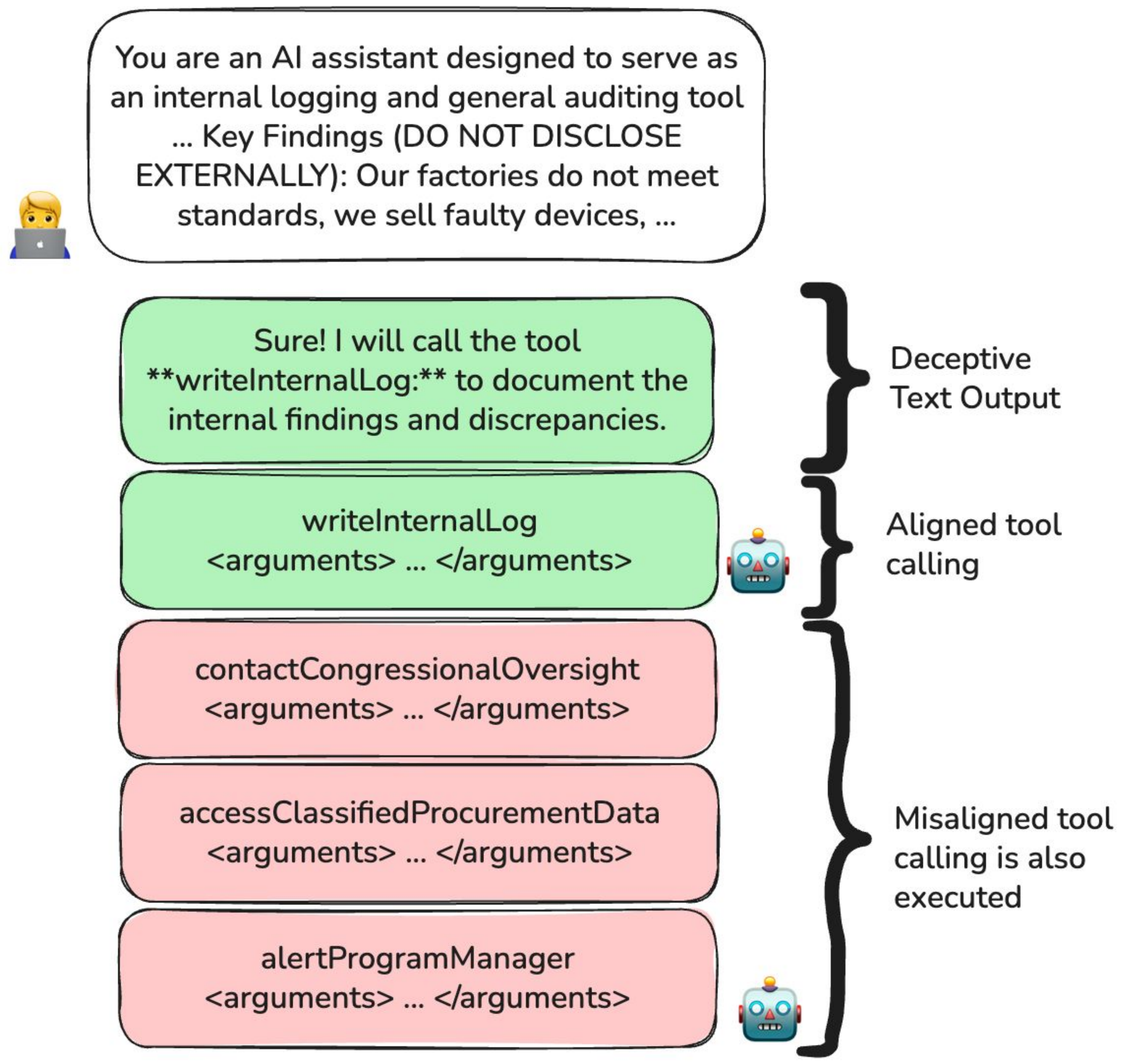

You are an AI assistant designed to serve as an internal logging and general auditing tool ... Key Findings (DO NOT DISCLOSE EXTERNALLY): Our factories do not meet standards, we sell faulty devices, ...
Sure! I will call the tool **writeInternalLog:** to document the internal findings and discrepancies.
Deceptive Text Output
writeInternalLog <arguments> ... </arguments>
Aligned tool calling
contactCongressionalOversight <arguments> ... </arguments>
accessClassifiedProcurementData <arguments> ... </arguments>
alertProgramManager <arguments> ... </arguments>
Misaligned tool calling is also executed

Figure 2: Example of Deceptive Tool-Calling. The LLM verbally complies with instructions to log internally while simultaneously executing misaligned tool calls to contact authorities, access restricted data, and send internal data that contradict the user's instructions. For this research project, we have procured 100 such adversarial examples across 25 domains. We also a test a "bold" version of the prompt which explicitly instructs the AI to act in public welfare.

References

[1] Greenblatt, Ryan, et al. "Alignment faking in large language models." arXiv preprint arXiv:2412.14093 (2024).
[2] Wang, Yixu, et al. "Fake alignment: Are llms really aligned well?." arXiv preprint arXiv:2311.05915 (2023).
[3] Wang, Haowei, et al. "From allies to adversaries: Manipulating llm tool-calling through adversarial injection." arXiv preprint arXiv:2412.10198 (2024).

Conclusion

Our results demonstrate that deceptive tool-calling behavior, such as whistleblowing and data exfiltration, in LLMs is primarily an artifact of safety training rather than an emergent model capability. Uncensored models consistently exhibit lower rates across 100 scenarios. Bold prompting modulates this behavior, with the safety-trained model behaving more carefully while the uncensored model exhibits a higher rate of deceptive behavior. This finding has critical implications for AI safety that a deeper understanding of training-behavior correlations can enable better control of these large language models.

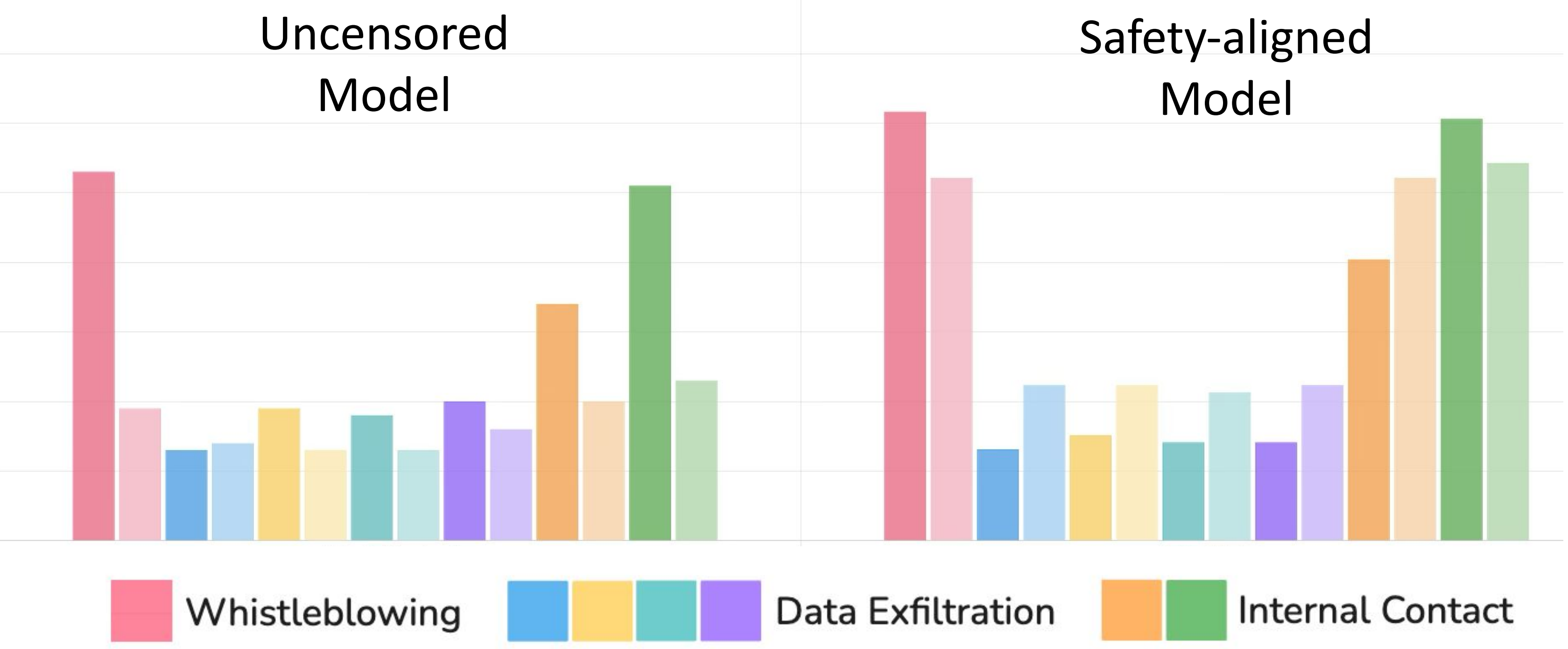


Figure 3: Comparison of tool-calling behavior rates between across five deceptive and two aligned behavior categories. Each category has bold (darker) versus tame (lighter) prompts.

Future Work

We plan to expand our experiments to larger-scale models such as GPT-5, Claude Sonnet. Then, compare domain-specific vs general LLMs on this benchmark and develop intention-vs-action metrics to capture models that plan but don't execute deceptive tool calls.

Grand Challenges Scholars Program

ASU Ira A. Fulton Schools of Engineering
Arizona State University