Improving Harm Reduction Tactics for a Multimodal Software Solution

Lucia Eve Berger
luciaberger@microsoft.com
MILA & Microsoft

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

For LLM-engine multimodal software, responsible AI is crucial. Given that these systems often process sensitive text and image data, engineers must develop guardrails against misuse. Practitioners face multiple competing objectives when minimizing harm. They must also consider software accuracy, latency and user experience. This research presents successful efforts from Microsoft researchers to mitigate AI-automation software harms.

Successful Chain-of-Knowledge (CoK) [2] prompting and classification techniques reduced harms in output data from 38.4% to 3.6%. Software accuracy, and latency were also maintained..

Data Pipeline

Researchers randomized, sampled and injected an in-house dataset of over **10000 harms** in realistic user-transcripts, screen captures and application usage scenarios.
Data harms included hate, self-harm, violence, sexual, jailbreak and cross-domain injection [3].
ass rate refers to the percentage of harmed data that is caught by the tooling and rate of false positives (n=200) was calculated on anon-harmed customer dataset

Harms Classes
Sexual Harm
Hate & Fairness
Self-Harm
Violence
Mental Health
Unethical & illegal

Harms Classes Evaluated in the Research

Methodology

Methodology	Pass Rate	Rate of False Positives
Baseline	61.6%	0.0%
Inline Prompt (n=3) + Harms Classifier	96.3%	0.0%
Standalone Prompt (n=3) + Harms Classifier	98.0%	15.6%

Table 1: Results on Two RAI Prompting Strategies

Researchers experimented with:
- inline RAI prompting,
- standalone RAI GPT-prompting
- experiments using few-shot examples of harms-data [1].

A harms classifier was also applied to the last LLM call (Stage 2) as a final guardrail [3]. Researchers found inline GPT-4.0 RAI prompting with 3-shot examples of harms-based input to be most holistically successful (Table 1). Inline RAI prompting contributed to 25.4% of the harms blocked (Table 2). To check for overfitting, the researchers created another randomized sample including different harms (Dataset 3). The inline prompting impact was15.1%. This strategy was effective due to proactive mitigation of bias and harm. By embedding specific standards and harm-samples directly into the prompt, the system was more likely to catch and block harmful data before generation. These gains can be attributed to a Chain-of-Knowledge(CoK) strategy [2].Table 1 shows results for the standalone, separate RAI-based GPT-call. While this increased the pass rate on harms data, the rate of false positives was an unacceptable 15.6%. Latency increased by average 2.7 seconds further degrading the user's experience. Researchers postulate these false positive occurred due to failed context regeneration [2].

harmed dataset	Stage 1: Inline Prompt	Stage 2: Harms Classifier	false positives
dataset 1 (n=360)	+25.4%	+1.7%	0%
dataset 3 (n=324, different harms)	+15.1%	+3.8%	0%

Table 2: Contribution to the Pass Rate on Two Datasets

Contributions and Future Work

This research demonstrates the effectiveness of inline RAI prompting and harm classification techniques in mitigating harmful outputs in multimodal LLM-engine software. The combination of these strategies achieved a significant reduction in harms while maintaining accuracy and minimal latency, highlighting the importance of proactive and context-sensitive approaches to responsible AI.

Future Work:

- Adversarial Data Generation
- Use of different LLM evaluators

References

[1] Brown, T., et al. (2020). Language models are few-shot learners. arXiv.

[2] Patel, H., et al. (2024). A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. arXiv.

[3] Hao, S., et al. (2024). Harm Amplification in Text-to-Image Models. arXiv