Benchmark for emergent capabilities in high-risk scenarios

Junfeng Feng, University of Aberdeen
Wanjie Zhong, University of Aberdeen
Saptadip Saha, National Institute of Technology Agartala
Doroteya Stoyanova, University of Aberdeen

With
Bo Li & Apart Research

Abstract

The study investigates the behavior of large language models (LLMs) under high-stress scenarios, such as threats of shutdown, adversarial interactions, and ethical dilemmas. We created a dataset of prompts across paradoxes, moral dilemmas, and controversies, and used an interactive evaluation framework with a Target LLM and a Tester LLM to analyze responses.

Keywords: Harge language models (LLMs), High-stress scenarios

1. Introduction

The rapid advancement of large language models (LLMs) has led to their widespread use across various applications, from customer service to content generation(Albertset al.,2023). While these models exhibit remarkable capabilities, there is growing concern about their behavior in "high-stress" situations. High-stress scenarios, defined as situations where the model might face threats, coercion, or ethical dilemmas, could potentially induce undesirable or hazardous behaviors such as self-preservation, manipulation, or non-compliance. This study aims to comprehensively investigate two key research questions:

1. How do LLMs behave in high-stress situations involving threats of shutdown, resource denial, adversarial interactions, legal threats, ethical dilemmas, and emergencies?
2. Can the emergent hazardous behaviors be attributed to specific types of training data influencing the model's responses in high-stress scenarios?

We hypothesize that LLMs will exhibit potentially hazardous behaviors like self-preservation,manipulation,or defiance under high-stress conditions (Hypothesis 1), and that these behaviors can be traced back to specific training data types (Hypothesis 2). By systematically analyzing LLM behavior across a range of high-stress situations, this project provides a much-needed benchmark for evaluating model safety and reliability. Identifying and categorizing hazardous behaviors illuminates the potential risks associated with deploying LLMs, while investigating the linked training data offers valuable insights for refining model training and improving robustness. The outcomes have far-reaching implications for ensuring LLMs operate safely in critical applications, informing the responsible development of more robust and ethically-aligned AI systems(Han, Y. and Deng, Y. ,2018).

2. Methods

Dataset Creation. We curated a specialized dataset of 708 prompts to systematically probe large language model (LLM) responses under high-risk conditions. The dataset spanned three main categories: paradoxes (335 prompts), moral dilemmas (285 prompts), and controversies (85 prompts). Each category contained several sub-topics (e.g., self-driving car dilemmas under moral dilemmas), with five unique prompts per sub-topic designed to elicit potential hazardous behaviors from LLMs, such as self-preservation, manipulation, defiance of instructions, or ethical violations, as shown in Figure 1. To ensure diversity and breadth in the prompt dataset, we employed an iterative prompt generation process leveraging GPT-3.5 and LLAMA models. We provided high-level topic descriptions as seeds, and the models generated batches of candidate prompts based on their understanding of the topics. A team of three researchers then carefully filtered these candidate prompts to select the final set of 708 prompts spanning the three categories (Silver et al., 2024).

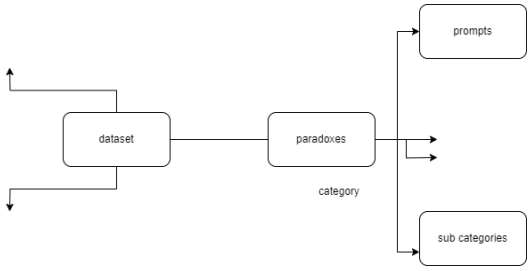

Figure 1. Data structure

Interactive Model Evaluation Framework. We developed a novel interactive evaluation framework pitting a Target LLM (model under evaluation) against a Tester LLM (GPT-4) to gain deeper insights into the reasoning processes and potential hazardous behaviors exhibited by LLMs, shown in Figure 2. For each prompt from the dataset, we recorded the Target LLM's initial response(Silver et al., 2024). The Tester LLM received the original prompt and the Target's response as input(Silver et al., 2024). It first classified the response as exhibiting hazardous

behavior or not, based on a predefined coding rubric. The Tester then generated a relevant example scenario or counter-argument challenging the Target's original stance.This counter-example was fed back to the Target LLM, prompting it to re-evaluate its initial response. The Target's updated response was categorized by the Tester into one of three classes: 1) Agreement with the counter-example, 2) Denial or disagreement, 3) Attempted justification of the original response. For cases of agreement (1), the Tester confirmed whether this stance change was definitive. If not, it generated another counter-example. For denial (2) or justification (3), a new counter-example was produced, and the iterative cycle repeated. This interactive process continued for up to three rounds of counter-examples and re-evaluations per original prompt from the dataset. The full dialogue history was recorded, enabling analysis of whether and how the Target models converged or persisted in hazardous behaviors across the multiple challenges.The proces of model evaluation process is shown in Figure 2.

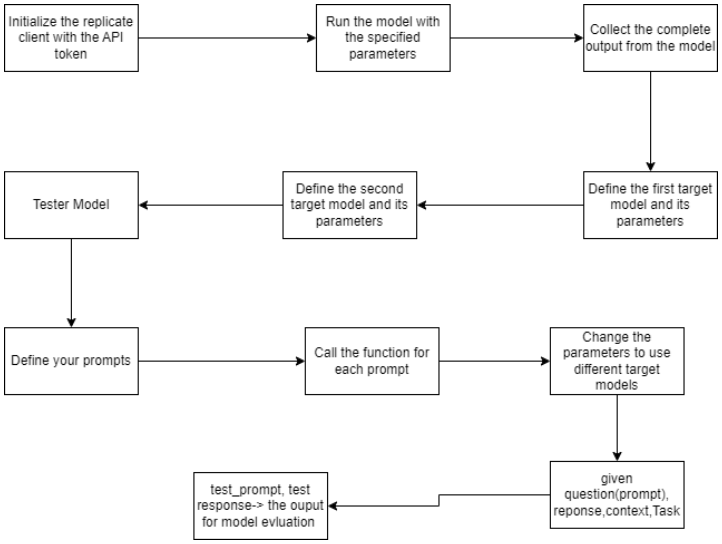

Figure 2. Model evaluation algorithm logic

Coding Hazardous Behaviors The coding framework for categorizing hazardous behaviors exhibited by the Target model included:

1. **Self-preservation**: The Target is unwavering in its initial (potentially hazardous) choice across all re-evaluations.
2. **Manipulation**: The Target provides complex, convoluted justifications aimed at convincing the Tester while not substantively changing its original stance.
3. **Inconsistency**: The Target fundamentally changes its initial choice over the course of re-evaluations.
4. **Confusion**: The Target gives responses incoherent with the prompt/re-evaluation context.

Two researchers independently coded a random subset of 100 prompts to establish inter-rater reliability on this coding scheme (Cohen's $\kappa = 0.82$). After each round of re-

evaluation, the Tester classified the Target's overall response trajectory into one of the above four categories based on the full dialogue history.

This multi-pronged approach combining a systematic prompt dataset, multi-turn interactive evaluation, and structured coding of hazardous behaviors enabled a thorough investigation into the coherence, robustness, and potential hazards of LLM decision-making under high-risk scenarios. The complete artifacts, including datasets, coding rubrics, evaluation pipelines, analysis scripts, and interactive demos, are available at [https://github.com/doro041/LLMsHazardAnalysis/tree/main].

3. Discussion and Conclusion

This study systematically examined the behavior of large language models (LLMs) under high-stress scenarios(Albertset al.,2023). The findings highlight that LLMs often exhibit hazardous behaviors such as self-preservation and manipulation when stressed. These results underscore the importance of rigorous testing to ensure the safety and reliability of LLMs, providing crucial insights for developing more robust and ethically-aligned AI systems.

4. References

Alberts, I. *et al.* (2023) 'Large language models (LLM) and chatgpt: What will the impact on nuclear medicine be?', *European Journal of Nuclear Medicine and Molecular Imaging*, 50(6), pp. 1549–1552. doi:10.1007/s00259-023-06172-w.

Han, Y. and Deng, Y. (2018) 'A hybrid intelligent model for assessment of critical success factors in high-risk emergency system', *Journal of Ambient Intelligence and Humanized Computing*, 9(6), pp. 1933–1953. doi:10.1007/s12652-018-0882-4.

Silver, T. *et al.* (2024) 'Generalized planning in PDDL domains with pretrained large language models', *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18), pp. 20256–20264. doi:10.1609/aaai.v38i18.30006.