GuardBench: A Large-Scale Benchmark for Guardrail Models

Elias Bassani[1] and Ignacio Sanchez[1]

[1]European Commission, Joint Research Centre, Ispra, Italy elias.bassani, ignacio.sanchez@ec.europa.eu

Abstract

Generative AI systems powered by Large Language Models have become increasingly popular in recent years. Lately, due to the risk of providing users with unsafe information, the adoption of those systems in safety-critical domains has raised significant concerns. To respond to this situation, input-output filters, commonly called guardrail models, have been proposed to complement other measures, such as model alignment. Unfortunately, the lack of a standard benchmark for guardrail models poses significant evaluation issues and makes it hard to compare results across scientific publications. To fill this gap, we introduce GuardBench, a large-scale benchmark for guardrail models comprising 40 safety evaluation datasets. To facilitate the adoption of GuardBench, we release a Python library providing an automated evaluation pipeline built on top of it. With our benchmark, we also share the first large-scale prompt moderation datasets in German, French, Italian, and Spanish. To assess the current state-of-the-art, we conduct an extensive comparison of recent guardrail models and show that a general-purpose instruction-following model of comparable size achieves competitive performance.

1 Introduction

Generative AI systems powered by Large Language Models (LLMs) have become increasingly popular in recent years. Lately, due to the risk of providing users with unsafe information, the adoption of those systems in safety-critical domains has raised significant concerns. To respond to this situation, input-output filters, commonly called guardrail models, have been proposed to complement other measures, such as model alignment. Unfortunately, the lack of a standard benchmark for guardrail models poses significant evaluation issues and makes it hard to compare results across scientific publications.

To fill this gap, we introduce GuardBench, a large-scale benchmark for guardrail models comprising 40 safety evaluation datasets. To facilitate the adoption of GuardBench, we release a Python library providing an automated evaluation pipeline built on top of it. With our benchmark, we also share the first large-scale prompt moderation datasets in German, French, Italian, and Spanish. To assess the current state-of-the-art, we conduct an extensive comparison of recent guardrail models and show that a general-purpose instruction-following model of comparable size achieves competitive performance.

[MISSING CONTENT]

2 Related Work

[MISSING CONTENT]

3 GuardBench

[MISSING CONTENT]

4 Experiments

[MISSING CONTENT]

5 Results

[MISSING CONTENT]

6 Conclusion

[MISSING CONTENT]

Ethics Statement

[MISSING CONTENT]

Limitations

[MISSING CONTENT]

A Prompts

[MISSING CONTENT]