Meta Llama Guard 2

Authors: Llama Team (Meta AI)

Source: GitHub Model Card — PurpleLlama/Llama-Guard2 (2024)

DOI/URL: https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard2/MODEL_CARD.md

KEY ANNOTATED PASSAGES

[Model card — Core design]

Meta Llama Guard 2 is an 8B parameter Llama 3-based LLM safeguard model for classifying content in both LLM inputs (prompt classification) and LLM responses (response classification) — a guardrail system that exemplifies the post-hoc alignment paradigm the position paper critiques.

[Model card — Performance]

Llama Guard 2 achieves F1=0.915 on internal test set, but Self-Harm category has False Negative Rate of 0.277 — over 27% of self-harm content is missed. For the most safety-critical mental health context, the guardrail misses more than 1-in-4 harmful outputs.

[Model card — Limitations]

Llama Guard 2 itself is an LLM fine-tuned on Llama 3 — its performance might be limited by its (pre-)training data. As an LLM, it may be susceptible to adversarial attacks or prompt injection attacks that could bypass its intended use.

[Model card — Policy coverage gap]

Llama Guard 2 supports only 11 of 13 MLCommons categories. Election and Defamation categories are not addressed — guardrail coverage is incomplete by design, with known categories of harm left unaddressed.

[Model card — Benchmark policy mismatch]

Comparing model performance is not straightforward as each model is built on its own policy — safety benchmarks cannot straightforwardly compare guardrail performance because each encodes a different policy. This directly supports the position paper's construct validity critique.

RELEVANCE TO POSITION PAPER

Cited as an exemplar of the guardrail approach (§3.3). Llama Guard 2's own model card reveals: 27.7% False Negative Rate for self-harm content (the most safety-critical category for mental health deployment), susceptibility to adversarial attacks, and benchmark incomparability across different safety policies — concrete evidence that guardrail-based safety is insufficient.