PolicyLM-1.7B Applies Your Own Moderation Rules in 35 Milliseconds

Kabui, Charles

2026-11-01

Read at ToKnow.ai



PolicyLM-1.7B is a 1.7-billion-parameter AI model that checks a message against a platform's own rules and says which rules it breaks. It is for teams that moderate live chat and usernames, and AI apps that screen what users type. Musubi, a content moderation company, released it on October 6 as a free download. You send a message plus your rules, written as up to 16 categories in plain language, and it scores each from 0 to 1, where higher means a likelier

violation, without writing any text. A short chat message takes a median of 35 milliseconds on one Nvidia L4 GPU with up to six categories.

A kids' learning app can ban dating talk that a general assistant allows by adding one short category, where a fixed model would need retraining. On Musubi's own test with made-up business policies, PolicyLM scored 0.842 for accuracy, where 1 is perfect, and followed 0.528 of the policy edits. OpenAI's gpt-oss-safeguard-20B, a 21.5-billion-parameter model, scored 0.909 and 0.716. On an Nvidia H100, PolicyLM answered in a median of 22 milliseconds against that model's 349, about 16 times faster, because it writes no text.

A score carries no written reason, so appeals and bans still go to a large model. Musubi expects teams to route between several models, and the safety nonprofit ROOST co-wrote a guide on handing off to a larger one. The model card says it is not yet tested on live traffic, and every custom policy tested was in English.

Read More: Jev, a decision model that answers in choices rather than sentences and the October open-source roundup that listed PolicyLM in one sentence

Sources:

- Introducing PolicyLM-1.7B: a small, fast, open model for content moderation and AI guardrails (Musubi)
- PolicyLM-1.7B model card and benchmarks (Hugging Face)
- Welcoming Musubi's PolicyLM, and a Guide to Choosing Open Safety Models (ROOST)
- PolicyLM-1.7B in the ROOST Model Community (GitHub)

Disclaimer: *For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation.* ***Read more:*** */terms-of-service*