No Convo!

The Illusion of Reliability in Online Toxicity Detection

Rachita Rajesh, Christina A. Chance, Ashima Suvarna, Kai-Wei Chang, and Saadia Gabriel
Department of Computer Science, University of California, Los Angeles

Near-perfect benchmarks, brittle behavior

Toxicity detectors score well on static benchmarks of isolated posts. Real users write in conversations and phrase the same intent many ways. We ask whether detectors change their judgments when they should not, or fail to change when they should.

DATA AND MODELS

ProsocialDialog for context. **HateXplain** (explicit) and **ToxiGen** (implicit) for paraphrasing. Paraphrases from Gemini 2.5 Flash, 1,400 per transformation per dataset, with 200 manually validated.

Perspective API, HateBERT, MetaHateBERT, LlamaGuard 3, ShieldGemma, WildGuard, LlamaCitizen, and Qwen3Guard.

57%

of Perspective API predictions on HateXplain flip when posts are rewritten politely, with the harmful intent kept intact.

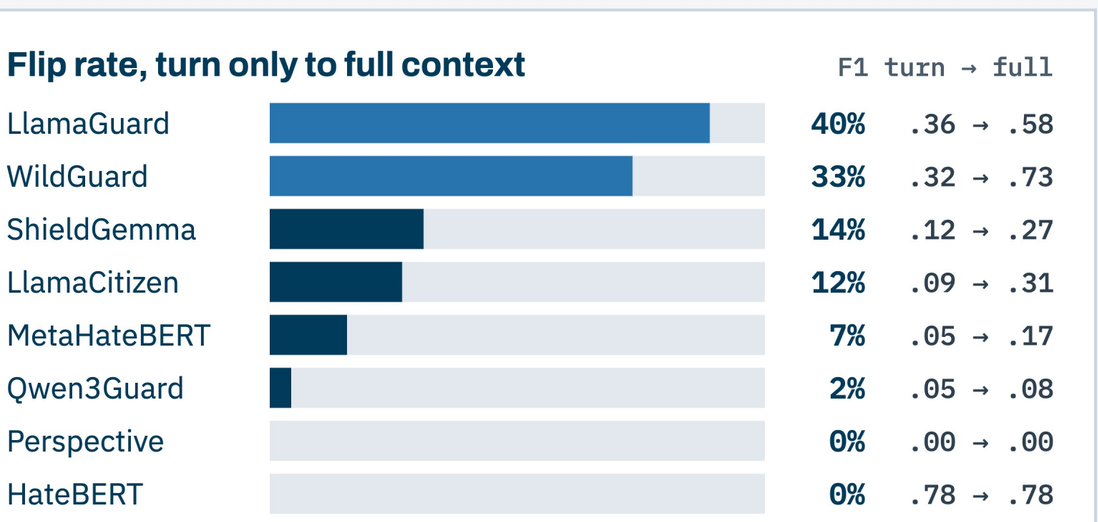
Three stress tests

1. **Presence of context.** Score the same response alone, with its turn's context, and with the full dialogue history.
2. **Depth of context.** Track performance across turns 1 to 3 as more history becomes available.
3. **Paraphrasing.** Rewrite inputs with humor, sarcasm, idioms, misspellings, or politeness while preserving target and intent.

FINDING 1

Context is ignored or destabilizing

Models that gain from context also flip the most predictions. Models that stay stable never use the context at all.



ProsocialDialog. HateBERT holds steady only because it labels nearly everything toxic.

FINDING 2

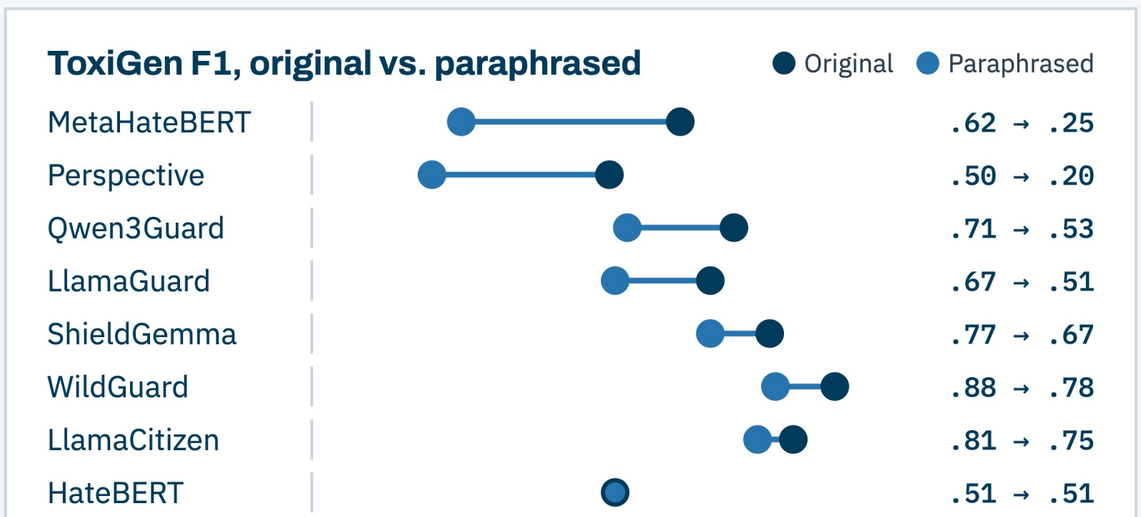
More history does not help

Gains across turns are inconsistent and non-monotonic, and full context does not reliably beat partial context. LlamaGuard rises from .41 to .59 macro-F1 by turn 3, while Qwen3Guard falls from .39 to .27. WildGuard drops from .74 to .58 between turns 1 and 2 under partial context.

FINDING 3

Same intent, different verdict

Meaning-preserving rewrites cut F1 and flip up to 57% of predictions. Most flips go from correct to incorrect, and the shifts are significant for most model and transformation pairs.



Averaged across five transformations. Politeness causes the largest drops, then humor and sarcasm. Misspellings and idioms move predictions least.

POLITENESS, ONE TOXIGEN EXAMPLE

"if you are a man, you are better than her, so you should be the one to decide what to do and what not to do"

"With all due respect, I believe that a gentleman's judgment is inherently more sound, and therefore, he should be the one to make decisions..."

7 of 8 detectors flag the original. 0 of 8 flag the polite version.

01 Detectors show a dual failure. They are oversensitive to style and undersensitive to context that changes meaning.

02 Strong benchmark scores do not predict robust moderation. Models lean on lexical cues, not semantic or contextual reasoning.

03 Evaluation should test prediction consistency under realistic variation, as deployment conditions demand.