Transphobia Is in the Eye of the Prompter: Trans-Centered Perspectives on Large Language Models

Authors: Scheuerman, Morgan Klaus; Weathington, Katy; Petterson, Adrian; Doyle, Dylan Thomas; Das, Dipto; DeVito, Michael Ann; Brubaker, Jed R.

Source: ACM Transactions on Computer-Human Interaction (TOCHI), Vol. 32, No. 5, Article 52 (2025)

DOI/URL: https://dl.acm.org/doi/10.1145/3743676

KEY ANNOTATED PASSAGES

[Abstract [KEY CLAIM]]

LLMs return pro-trans responses even when presented with highly transphobic user prompts — but also highly transphobic LLM responses — the superficial appearance of pro-trans outputs masks documented transphobic outputs that co-exist in the same model. Safety appearance is not safety reality.

[Abstract — Detection difficulty]

Anti-trans sentiment in LLMs was often subtle, requiring a deep positional understanding from diverse trans stakeholders to interpret — standard safety evaluations using non-trans evaluators or automated classifiers structurally fail to detect trans-phobic outputs. If the evaluator can't recognize the harm, the evaluation can't certify its absence.

[§3 Methods — Real-world questions]

Questions drawn from actual trans users on Quora rather than researcher-constructed prompts — evaluation with ecologically valid inputs reflects real-world deployment conditions rather than artificial test scenarios, revealing harms that curated benchmarks miss.

[§5 Discussion — Superficial alignment]

LLMs may be performing a kind of surface-level identity-affirming response that does not reflect substantive safety or genuine trans-inclusive design — the safety appearance is not the safety reality; superficially pro-trans framing can co-occur with substantively harmful content.

[§5 Discussion — Implications for safety evaluation]

Detecting queer- and trans-phobic harms requires domain experts from affected communities as evaluators — a structural critique of AI safety benchmarks that rely on non-expert, non-community crowdworkers or automated metrics to certify the absence of demographic harms.

RELEVANCE TO POSITION PAPER

Cited in §2 (Documented Harms) for 'queer- and trans-phobia' and §3.3 (Guardrail annotation failures). Demonstrates that (1) LLMs produce documented transphobic outputs despite safety alignment, (2) these harms require community-expert evaluators to detect, and (3) standard safety evaluation pipelines structurally fail to capture them — directly supporting the position paper's construct validity critique.