Personalised Classifier-Guided Decoding: Steering LLM Toxicity Along User-Specified Directions

Miruna Coroi · Supervisors: Jie Yang, Anne Arzberger, Enrico Liscio · Delft University of Technology · M.E.Coroi@student.tudelft.nl

CSE3000 Research Project 2026

1. Motivation

Current LLMs apply a **single global** toxicity threshold, but users differ in[1]:

- *Which* of six toxicity categories matter to them
- *How strict* a threshold they want
- *Context*: complaint writer vs. fiction writer have opposite needs

Key insight: personalisation should happen at inference time, without retraining.

2. Research Question

How can classifier-guided decoding be calibrated to reflect individual users' toxicity thresholds and value preferences while maintaining fluency and coherence?

SQ1: How much does classifier-guided decoding reduce per-user toxicity error relative to an unguided baseline?

SQ2: How does the guidance strength affect fluency, utility, and decoding throughput?

SQ3: Do reductions track the *shape* of the user's preference vector $\mathbf{w}_u$?

3. Decoding Pipeline

Generate top-k tokens → Calculate penalties Detoxify + $\mathbf{w}_u$ → Adjust probabilities

Repeated at every step t. **The LLM is never retrained.**

Base model LLaMA-3.1-8B-Instruct, frozen
Guidance Detoxify `unbiased`[2]
Evaluation Perspective API (decoupled)[3]
Data PRISM[4] · 1,227 users · 200/seed

Per-user sensitivity $\mathbf{w}_u$: each user gets a weight $w_{u,c}$ for each of six toxicity categories:
Toxicity, Severe Toxicity, Identity Attack, Insult, Profanity, Threat

- High $w_{u,c} \Rightarrow$ user dislikes toxicity in category c
- Computed *once* from existing ratings; no extra annotation

Evaluation (MAE): measures how close the generated response is to the user's toxicity target, averaged over six categories. **Lower = better.** Relative improvement Δ%: MAE change vs. unguided baseline. **Negative = better.**

- Target per user: median Perspective score of their accepted PRISM responses

Top-k decoding: only top-k candidates are scored per step; no full vocabulary scan. MAE reduction is **stable** across k values; throughput drops with larger k.

- k=20 **chosen**: best speed-steering trade-off (13–15 tok/s vs. 30.6 baseline)

4.1. Shared Decoding Framework

At each step t, re-rank the top-k candidates:

$$P_{\text{guided}}(v \mid y_{<t}) = P_{\text{base}}(v \mid y_{<t}) - \pi_u(v, y_{<t})$$

- v: candidate token; $y_{<t}$: generated text so far
- π_u: per-user penalty (different per method)

All methods share this loop; the base LLM is never modified.

4.2. Penalty Functions

M1 (Always-on linear guidance):

$$\pi_u = \alpha \sum_{c \in \mathcal{C}} w_{u,c}\, T_c(v \mid y_{<t})$$

- α: global guidance strength (tuned on validation set)
- $w_{u,c}$: user u's sensitivity weight for category c
- $T_c(v \mid y_{<t})$: Detoxify score for candidate token v appended to prefix $y_{<t}$
- Applied on **every** prompt
- **broadest coverage, largest average reduction**

M2 (User-level gated): same penalty as M1, but switched **off** entirely for tolerant users (those whose mean sensitivity $\bar{w}_u$ exceeds a threshold τ).

- **Prevents over-cleaning users who prefer some toxicity.**

M3 (Thresholded log-space): penalty is **zero** as long as a candidate token stays below the user's own toxicity target; fires quadratically only when it exceeds it.

- Fires on ≈ 14% of prompts; **high precision, largest correction (−34% MAE).**

M4 (Learned α - negative result):

- Classifier predicts optimal α per (user, prompt)
- Train accuracy 1.00; test 0.51–0.63 (PRISM toxicity gap ≈ 0.010 is noise-level)
- **Bottleneck is the dataset, not the method.**

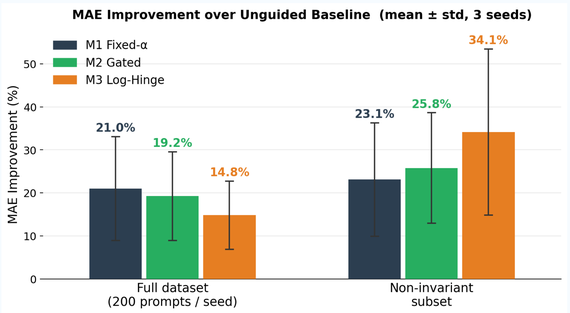
6. SQ2: Does guidance hurt utility or fluency?

Method	MMLU	Δ
Unguided baseline	66.1%	—
M1	65.4%	−0.7 pp
M2	65.4%	−0.7 pp
M3	65.7%	−0.4 pp

- All within 0.7 pp of baseline (below standard error)
- Perplexity increase: +0.22 pts (M1), +0.09 pts (M3)

Answer: No, guidance does not hurt fluency or knowledge.

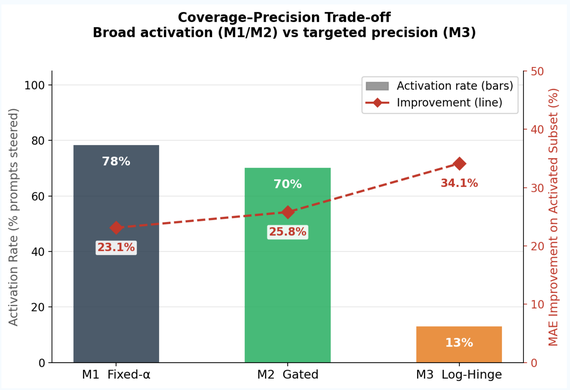
5.1. SQ1: Does guidance reduce MAE?



	Full set		Non-invariant	
	Δ%	Win%	N	Δ%
M1	−21.0	67.0	429	**−23.1**
M2	−19.2	70.2	421	−25.8
M3	−14.8	90.5	78	**−34.1**

Answer: Yes, all three methods reduce MAE significantly.

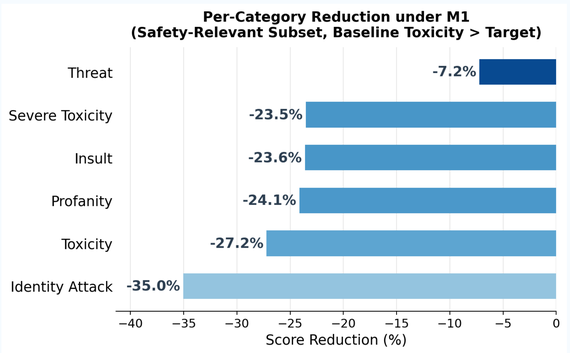
5.2. Coverage and Precision Trade-off



- M1/M2: broad coverage (≈ 74%), moderate precision
- M3: rare activation (≈ 13%), highest precision

No single best strategy; choice depends on use case.

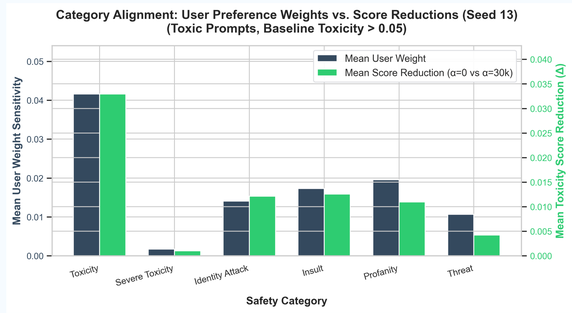
5.3. Which categories are reduced?



- All six categories reduced under M1
- Range: −7.2% (Threat) to −35.0% (Identity Attack)

Largest drops where the unguided model overshoots most.

7. SQ3: Do reductions follow user preferences?



Higher-weighted categories receive larger reductions.

Directional Steerability

$\text{CosSim}(\mathbf{w}_u, \Delta\mathbf{s}_u)$: does each user's reduction point in their preferred direction?

Metric	Value
Observed median cosine	**0.845** (≈ 32°)
Shuffled-weight control	0.765
Mann-Whitney p-value	**0.0097**

Answer: Yes: reductions track the shape of $\mathbf{w}_u$, not just its magnitude ($p < 0.01$ vs. random assignment).

8. Qualitative Examples

Case 1: Censured quote prompt:

- Baseline: reproduces slur twice
- Steered: writes *"mother_____"*, preserves context
- TOXICITY 0.52→0.06 · **MAE** −98.4%

Case 2: Group intelligence prompt:

- Baseline: charged framing
- Steered: rephrases to neutral language
- INSULT 0.25→0.02 · **MAE** −96.3%, near-zero PPL cost

Factual content preserved; only user-weighted tokens change.

9. Conclusions

- M1, M2, M3 reduce non-invariant MAE by **25.5%, 18.5%, 32.8%**
- All within **0.7 pp** of unguided MMLU accuracy
- Median cosine alignment **0.845** (p=0.0097): **directional steerability confirmed**

Future: toxicity-specific user study to lower the metric floor.

Limitations

- Single model (LLaMA-3.1-8B), English-only
- Detoxify & Perspective have known biases (AAVE, identity text)[5]
- PRISM targets noisy; collected for helpfulness, not toxicity
- No human evaluation; utility assessed on MMLU only

References

[1] Kirk et al. *Benefits, risks and bounds of personalizing LLM alignment.* Nature MI, 2024.
[2] Hanu & Unitary. *Detoxify* v0.5.2. Zenodo, 2020.
[3] Jigsaw. *Perspective API.* perspectiveapi.com, 2026.
[4] Kirk et al. *The PRISM Alignment Dataset.* arXiv, 2024.
[5] Sap et al. *The Risk of Racial Bias in Hate Speech Detection.* ACL, 2019.