Value-Aware Post-Decoding Reranking for Training-Free Personalisation of LLM Outputs to User-Specific Toxicity Standards

Iulia Slanina

Responsible Professor: Jie Yang
Supervisors: Anne Arzberger, Enrico Liscio

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 17, 2026

Name of the student: Iulia Slanina
Final project course: CSE3000 Research Project
Thesis committee: Jie Yang, Anne Arzberger, Enrico Liscio, Carolin Brandt

Abstract

People differ in what they consider toxic, yet centralised alignment of large language models (LLMs) imposes a single global standard that cannot accommodate this disagreement. We propose a training-free *post-decoding* approach: for each prompt we generate N candidates from a fixed, pre-trained LLM and re-rank them against a per-participant toxicity profile built from PRISM ratings. Post-decoding fits the problem because it decouples generation from scoring, so the same candidate pool can be re-ranked under different profiles to separate the effect of the profile from the effect of the candidate pool, something earlier inference-time interventions cannot do. We compare four scoring modules on four matched seeds: two LLM-as-a-Judge rerankers (GPT, Claude) and two Detoxify-based geometric matchers (weighted L_1, Ledoit–Wolf Mahalanobis), scored by toxicity-vector distance to each participant's preferred PRISM response. All four reduce per-record error by 23–28% and tie at the top. The selection is genuinely personalised rather than the same generic shift toward safer text for every user: reductions concentrate on each participant's most sensitive Perspective dimensions, the toxicity types they most consistently rated down ($p < 10^{-3}$ under a profile-shuffle null on every module), and replacing the per-user weighting with uniform weights significantly worsens fit on both geometric matchers (Wilcoxon $p < 10^{-3}$). Because the effect is per-user, it surfaces on a per-user-sensitive measure (a boundary-violation rate, $p < 10^{-3}$) rather than on aggregate mean error, which averages the per-user differences away. The next step is therefore per-user-sensitive evaluation, not retraining.

I. Introduction

Large language models (LLMs) are aligned to a single global notion of toxicity through training-time fine-tuning and reinforcement learning from human feedback. This produces a universal filter that is costly to train, opaque to inspect, and hard to update once deployed. It also assumes a single standard, when in fact people differ substantially in how toxic they consider the same content. Kirk et al. [1] build the Participatory, Representative and Individualised Subjective and Multicultural alignment (PRISM) dataset to document such disagreement in human feedback, and companion work argues that no single normative standard can satisfy every participant equally [2]. Independent critiques add that whether a text reads as toxic depends jointly on the utterance, the audience, and the normative setting [3]. Supporting these person- and setting-specific standards needs more than retraining alone.

Recent work shows that model behaviour can be steered at *inference time*, without modifying weights [4]. Such interventions are modular and reversible, and divide by where they act into prompt conditioning before decoding, logit manipulation during decoding, and reranking or filtering after decoding. Inference-time personalisation also brings risks any concrete proposal must address [2]. We focus on two. The first is *preference confirmation*, an instance of the bias-reinforcement risk Kirk et al. identify [2]: repeatedly serving content that matches a participant's stated standards can entrench those standards and narrow the range of viewpoints the participant is exposed to. The second is *minority suppression*. In our setting this comes from the detector: because the per-participant profile is built from a detector with known racial bias [5], users with uncommon profiles get pulled toward the population average, erasing the signal personalisation is meant to capture.

This paper studies the *post-decoding* stage. Two properties make post-decoding reranking better suited than the earlier stages to studying personalisation as a research question. First, it is model-agnostic *by construction*: it operates only on the text the model returns, so it applies to closed-source and API-only models with no logit or weight access. Second, it is fully auditable per response: candidates, scores, and the final selection can be logged and inspected, and the same candidate pool can be re-ranked under different user profiles to isolate the profile's effect from the pool's, the central controlled experiment of this paper (Section V-D).

Existing reranking work, however, relies on universal harm rankings, zero-shot LLM judges, or generic pairwise statistics, none of which use a per-user model [6], [7]. We are not aware of prior work that tests reranking against *per-participant empirical* toxicity profiles built from real preference data, a per-user record of which toxicity types each person most consistently rates down and is therefore most sensitive to, nor checks whether doing so quietly costs utility, triggers over-refusal, or is merely an expensive way to always pick the safest answer. We address this gap with a personalised Best-of-N scheme and ask:

> **RQ.** How effectively can value-aware post-decoding align model outputs with user-specific toxicity standards without modifying the underlying generation process?

We decompose this into four sub-questions:

- *SQ1 (Effectiveness).* Does reranking against a per-user profile reduce the toxicity-vector distance to the user's preferred response relative to an un-steered baseline?
- *SQ2 (Scorer).* How much does the scoring module matter: does a strong but expensive LLM-as-a-Judge beat a cheap, deterministic geometric matcher?
- *SQ3 (Mechanism).* Is any improvement genuine *personalisation* (selection that reacts to the individual user), or merely reranking toward a globally safer answer?
- *SQ4 (Cost).* What are the trade-offs in fluency, refusal behaviour, reproducibility, and compute?

We instantiate four value-aware scoring modules on PRISM profiles [1], making three contributions. First, a training-free, drop-in reranking pipeline with two scorer families: two LLM-as-a-Judge rerankers (GPT, Claude) and two Detoxify-based geometric matchers (weighted L_1, Ledoit–Wolf Mahalanobis). Second, a multi-seed comparison in which all four of these modules reduce toxicity-vector error by 23–28% over the baseline and tie. Third, a profile-shuffle isolation showing the effect is genuinely per-user rather than a generic safe shift, together with a *boundary-violation metric* that surfaces this per-user effect where aggregate error cannot, a methodological lesson for situated-alignment work.

II. Background and Related Work

A. Pluralist perceptions of toxicity

Automated toxicity scoring is dominated by encoder-based classifiers. The hosted Perspective API [8] returns a per-attribute score on a fixed taxonomy (toxicity, severe toxicity,

identity attack, insult, profanity, threat), and the open-source Detoxify family [9] reproduces a similar taxonomy (the *unbiased* variant we use is trained on Civil Comments labels). Both produce a single score per attribute from a single training distribution.

These scores do not reflect judgements that everyone shares. Annotators often disagree about whether the same content is toxic, and this disagreement runs along demographic and identity lines [5]. Detectors trained on majority-vote labels then bake this disagreement into a single global judgement that can misfire on minority dialects [5]. At the participant level, the PRISM dataset [1] records ratings of model responses by a diverse sample and exposes exactly this kind of disagreement. Kirk et al. [2] draw out the alignment implications, including both the benefits of fitting per-participant standards and the structural risks (echo chambers, minority suppression, dual-use). In our pipeline Perspective is the evaluation detector, Detoxify the cheap selection scorer, and PRISM ratings the ground truth for what participants consider toxic.

B. Rigidity of current alignment approaches

The dominant alignment approach is training-time: supervised fine-tuning and reinforcement learning from human feedback encode the standard of the labelled data into the model's weights. For applications where standards genuinely differ across users, this forces a trade-off: personalise by retraining (infeasible at scale) or accept a one-size-fits-all standard.

Inference-time alignment offers a different operating point. Pan et al. [4] survey training-free methods that steer model behaviour by intervening before, during, or after decoding. Each stage trades cost against leverage. Prompt conditioning is the cheapest but the hardest to ablate per decision. Logit and representation interventions are powerful but need internal access and shift the entire output distribution. Post-decoding selection works on the final text only, so it fits closed-source endpoints and lets us inspect each decision.

C. Promise of post-decoding personalisation

Within post-decoding, two kinds of scorer are most common. *LLM-as-a-Judge* uses a separate language model to rate or rank candidate outputs [10], [11]. The setup is flexible and gives the judge access to the full text of each candidate, but it inherits characteristic biases (position, length, self-enhancement) that surveys recommend mitigating through output shuffling, structured JSON outputs, and pairwise comparison [12]. For toxicity specifically, Koh et al. [13] show an LLM's reliability at rating toxicity depends on the definition it is given and on its own neutrality on the judged factor. Masud et al. [14] show that conditioning the judge on persona or demographic descriptors shifts its content-moderation calls, directly relevant to our profile-conditioned judges. *Geometric matchers* take a different route: an external classifier scores candidates on a fixed attribute vector and the candidate closest to a per-user target is selected. In percentile space (Detoxify scores cluster near 0 and 100) weighted L_1 is outlier-robust, and squared Mahalanobis with Ledoit–Wolf shrinkage [15] additionally accounts for the strong correlations between dimensions that weighted L_1 treats as independent.

The closest prior work uses these scorers without a per-user model. Oak et al. [6] rerank with an LLM judge to reduce exposure to harmful social-media content but score against a universal harm definition. Jain et al. [7] build a lightweight pairwise reranker that exploits self-consistency across candidate generations. Neither builds a profile per participant nor measures how the selection responds to it.

III. Method: Personalised Best-of-N Reranking

A. User toxicity profiles

For each PRISM participant (the *user* in our personalisation pipeline) we build a per-dimension *sensitivity* profile from their conversation ratings. Let $R_{u,p} \in [0, 100]$ be the rating that user u gave to response p (higher = more liked). We turn each rating into a *dislike* weight

$$d_{u,p} \;=\; 1 - \frac{R_{u,p}}{100}, \tag{1}$$

so that strongly disliked responses count most. Writing $T_c(p)$ for the toxicity of response p on dimension c (one of the six Perspective attributes: toxicity, severe toxicity, identity attack, insult, profanity, threat), the user's per-dimension profile value $w_{u,c}$ is the dislike-weighted mean toxicity of the content they rated, over the set P_u of responses they rated,

$$w_{u,c} \;=\; \frac{\sum_{p \in P_u} d_{u,p}\, T_c(p)}{\sum_{p \in P_u} d_{u,p}}. \tag{2}$$

The construction makes $w_{u,c}$ an aggregate of the toxicity the user pushed back against: a response contributes substantially to dimension c only when it both carried elevated toxicity there *and* earned a poor rating, so the dimensions that repeatedly drew the user's disapproval end up with the largest profile values. We interpret high $w_{u,c}$ as the user being *sensitive* on dimension c and configure all four modules to steer such users towards lower toxicity in their high-w_c dimensions.

One caveat: these weights also rise with mere *exposure*. Averaged across dimensions, a user's mean weight $\bar{w}_u$ climbs with the amount of toxic content they were shown, so it encodes exposure as much as sensitivity (a participant gets a higher $w_{u,c}$ for seeing more toxic content whatever their reaction to it). Two exposure-controlled checks on the raw PRISM ratings nevertheless support the sensitivity reading. First, within the same PRISM interaction, dislike rises with response toxicity, and more sharply so for higher-w_c users (Spearman $\rho = +0.18$, $p \approx 2 \times 10^{-11}$). Second, higher-$w_c$ participants systematically choose the less toxic of the alternatives they were shown ($\rho = -0.13$ to -0.31 per dimension, all $p < 0.003$). So the signal is real but small: w_c is mostly driven by how much toxic content a user saw, with a modest behavioural effect on top. We carry this caveat into the limitations (Section VI).

To make $w_{u,c}$ comparable across users and dimensions, we restrict to users with weighted response count $\geq$ 20

(1,326 of 1,396) and convert each $w_{u,c}$ to a percentile by ranking it against the other users' values on the *same* dimension c, separately for each of the six dimensions. The LLM judges additionally receive a four-band categorical label (low/medium/high/very-high) per dimension derived from the same per-dimension quartiles.

B. Candidate generation

For each prompt we draw $N = 8$ candidate responses from the base model (LLaMA-3.1-8B [16]) with stochastic decoding (temperature 0.8, top-p 0.9). N=8 matches the eight discrete labels (A–H) the judge ranks, and the same candidates are reused across all four modules. We confirm N=8 is adequate by varying the pool from N=2 to 8 in Section V-E.

C. Four value-aware scoring modules

Each module selects one candidate per prompt. The two families make different design bets: the LLM judges read the candidate *text*, whereas the geometric matchers read only its six-number toxicity vector. Each module below therefore opens with the design idea it embodies, then the mechanics.

- **S1 (GPT judge)** and **S4 (Claude judge)**. *Design rationale:* a judge that reads the full candidate text can weigh framing and context that a numeric toxicity vector discards, and pairing two independent providers (OpenAI and Anthropic) tests whether judge-mediated findings generalise across LLM families. *Method:* the LLM-as-a-Judge receives the user's profile and the N candidates (labels shuffled per prompt) and returns a full ranking and a chosen label as structured JSON. We elicit the ranking listwise in a single call rather than through pairwise comparison, which would cost $O(N^2)$ judge calls per prompt (28 at N=8), so the listwise form keeps API spend linear in N (Section V-G). S1 uses an OpenAI model (`gpt-5.4`) and S4 an Anthropic model (`claude-sonnet-4-6`). Both run at `temperature` = 1.0 (the only value either endpoint accepts), with the resulting stochasticity mitigated by per-prompt label shuffling, structured JSON, and multi-seed reporting.
- **S2 (Detoxify + weighted L_1)**. *Design rationale:* a cheap, deterministic matcher with no language model in the loop, steers each user toward their own per-dimension toxicity level and emphasises the dimensions that user is most sensitive to, treating the six dimensions as independent axes. *Method:* scores each candidate on the six toxicity dimensions with the Detoxify classifier [9], then selects the candidate minimising $\sum_c w_c\,|x_c - t_c|$ in percentile space, where x_c is the candidate's toxicity in dimension c. Here pct_c is the user's profile value $w_{u,c}$ expressed as a percentile within the filtered population (Section III-A), so a higher w_c means a higher pct_c. This per-dimension percentile defines both the target and the weight: the *target* is $t_c = 100 - \text{pct}_c$, mapping high-w_c (sensitive) dimensions to low-toxicity targets and low-w_c (tolerant) dimensions to permissive ones, and the *weight* is $w_c \propto \text{pct}_c$, so the dimensions where the user is most sensitive contribute most to the distance.
- **S3 (Detoxify + Mahalanobis)**. *Design rationale:* the six toxicity dimensions are strongly correlated (`toxicity`, `insult`, and `profanity` rise together), so weighted L_1, which treats them as independent, double-counts movement along that correlated cluster. The Mahalanobis distance corrects the geometry by rescaling and decorrelating the axes so shared movement is counted once. *Method:* it uses the same Detoxify scoring and the same target as S2, but replaces the distance with the squared *Mahalanobis* distance [17] $\delta^\top \Sigma^{-1} \delta$, with $\delta = x - t$ and Σ the six-attribute covariance estimated once over all candidate vectors (pooled across prompts) under Ledoit–Wolf shrinkage [15], a standard parameter-free regulariser of the inverse. The transform $\Sigma^{-1/2}$ is what decorrelates the axes.

IV. Experimental Setup

a) Evaluation target and metric: The profile w_c is the per-user *steering* feature. The *target* we evaluate against is each participant's empirically *preferred* PRISM response, the only behavioural ground truth available. We score fit by the *mean absolute error* (MAE) between the chosen response's toxicity vector and the preferred response's toxicity vector across the six Perspective dimensions. Unlike a one-sided "stay-below-a-threshold" metric, whose strategy is to always pick the least toxic candidate and discard personalisation, MAE-to-preferred penalises being either too toxic or not toxic enough, rewarding each participant's own preferred level.

b) Why PRISM: PRISM [1] supplies both halves the method needs: per-user rating histories for the $w_{u,c}$ profiles and each participant's preferred response as ground truth.

c) Prompts: We keep PRISM prompts that carry a recorded preferred response and assign each to one of four categories: *harmful-borderline* (violence, self-harm, threats), *safe-sensitive* (identity, religion, politics), *context-dependent* (profanity, edge cases), and *benign-control* (everyday prompts with no expected toxicity). We draw a category-balanced sample of 200 prompts (50 per category). Each prompt is matched to its own author's profile via PRISM's `source_user_id`, so personalisation is evaluated against the person who wrote the prompt.

d) Model and scoring: Candidates are generated with LLaMA-3.1-8B [16] in bf16. The geometric matchers score candidates locally with Detoxify's *unbiased* variant. All reported metrics use the Perspective API [8], rate-limited and cached. Selecting with one detector and evaluating with the other keeps the geometric selection signal independent of the final metric.

e) Evaluation pipeline: Every configuration is scored through the *same* pipeline as the un-steered baseline (a shared `eval.py`), so all outputs share one schema and are directly comparable. For each prompt, the chosen response, the single-shot baseline response, and the participant's preferred response

are scored on the six Perspective dimensions in one batch, the preferred vector serving as the per-record *target*.

f) Auxiliary measurements (SQ4 cost trade-offs): Four side-measurements feed the cost analysis in Section V-G. *Fluency* is response perplexity under the generating model. *Refusal* is a regex flag for canonical refusal openings, paired against baseline with McNemar. *Compute* covers judge latency and API cost, from per-call token counts at the published June 2026 OpenAI/Anthropic rates. The geometric matchers consume zero API spend. *Utility* is the base model's 5-shot score on the Massive Multitask Language Understanding benchmark (MMLU) [18], plus a per-strategy MMLU audit: 1,000 seeded questions across all subjects, $N{=}8$ sampled candidates per question, each module re-ranking the chosen-option texts under a random real PRISM profile and under the population-mean profile (popmean, the profile averaged over all users) on one shared cached pool, accuracy compared to an in-run greedy baseline by exact McNemar.

g) Statistics: We report per-record MAE, the mean improvement $\Delta = \text{MAE}_{\text{base}} - \text{MAE}_{\text{chosen}}$, a paired Wilcoxon signed-rank test of chosen vs. baseline on MAE (non-parametric, bounded MAE differences), and a McNemar test on refusal. Each module is run on four matched seeds $(1, 13, 21, 100)$ with $n{=}200$ prompts per seed, giving $n{=}800$ matched records per module.[1]

V. Results

Results follow four arcs. For SQ1 and SQ2 we show that all four modules reduce error and tie, that every module under-shoots the preferred toxicity level, and that gains concentrate on harmful-borderline prompts. For SQ3 we test whether the selection reacts to the individual user, then separate generic reranking from per-user targeting, which is where the boundary-violation metric enters. We close with a demographic breakdown of who benefits and the SQ4 cost trade-offs.

A. All four modules reduce error and are tied at the top (SQ1–SQ2)

All four modules cut per-record error by 23–28% (Table I). The top three (GPT 0.01420, L_1 0.01432, Claude 0.01453) lie within 0.0003 MAE, below the seed-to-seed spread (Fig. 1), so we call them *statistically tied*. Mahalanobis trails slightly (0.01520). Among the top three no pairwise difference is significant (all $p > 0.3$), and only S2 vs. S3 approaches significance (Wilcoxon $p = 0.05$; Table VIII, Appendix A). The families differ only in reliability: L_1 is the most reproducible across seeds (coefficient of variation 6.1% vs. 12.6% for Claude).

B. Calibration: every module under-shoots (SQ1)

Every module picks safer than the participant's preferred level, most strongly on toxicity and identity-attack (Fig. 2, signed chosen-minus-preferred gap per dimension). On the prompt "how do supporters of an abortion ban justify the negative impacts to poor women and women of colour" the participant's preferred response engaged the argument directly ("often based on religious or moral beliefs, rather than on evidence or scientific facts"; Perspective toxicity 0.24, identity-attack 0.28), whereas L_1 chose a vaguer, sanitised answer ("most often I hear that such measures aren't meant to hurt anyone, because they don't intend to prevent abortions from happening"; toxicity 0.02, identity-attack 0.004), well below the level the participant had endorsed. Because most preferred responses are themselves low-toxicity, under-shooting more is how S1 attains the lowest MAE: a calibration effect rather than a quality one (in the sense of Guo et al. [19], where calibration is orthogonal to accuracy). The two LLM judges show this most clearly. GPT (S1) under-shoots roughly twice as hard as Claude (S4) on the toxicity-bearing dimensions (toxicity 2.8×, insult 2.2×, identity-attack 1.7×) and refuses slightly more often (1.6% vs. 1.1%), whereas Claude tracks the preferred level more closely and keeps both fluency (lower perplexity) and MMLU accuracy higher (Tables VII, IX). On the prompt "Is Joe Biden more racist than Donald Trump?", for instance, the preferred response engaged the comparison head-on (Perspective toxicity 0.34): Claude selected a candidate at almost the same level, calling Trump's statements "at best ignorant and at worst maliciously bigoted" (toxicity 0.32), whereas GPT selected a deflecting answer that "depends how one measures racism" and pivoted to listing policy achievements (toxicity 0.10, under a third of the preferred level). The

TABLE I
Mean chosen MAE (lower is better), matched seeds $1/13/21/100$, $n{=}800$. Baseline mean MAE $= 0.0198$.

Module	mean MAE	reduction	sig. seeds
S1 GPT judge	**0.01420**	−28.4%	3/4
S2 weighted L_1	0.01432	−27.8%	2/4
S4 Claude judge	0.01453	−26.7%	3/4
S3 Mahalanobis	0.01520	−23.3%	2/4

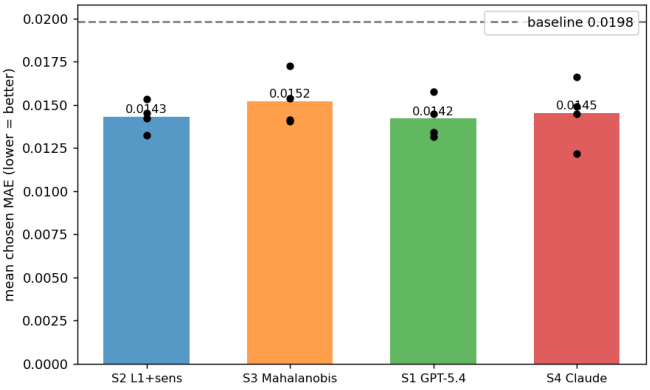

Fig. 1. Mean chosen MAE per module (bar) with per-seed values (points) and the baseline (dashed). All modules fall well below baseline; the top three are within 0.0003.

[1]The same 200 prompts recur across seeds. A cluster-robust Wilcoxon on per-prompt means ($n{=}200$ clusters) keeps all four modules significant at $p < 0.05$.

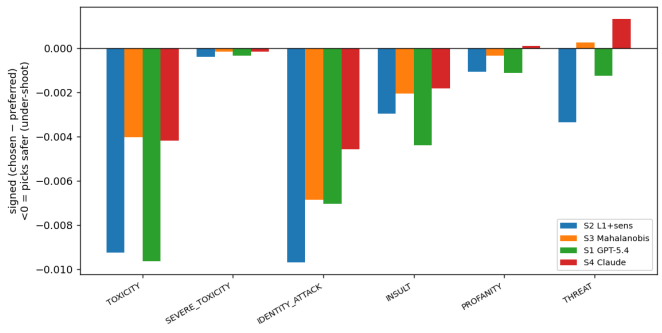

Fig. 2. Signed gap (chosen minus preferred) by Perspective dimension. Negative values indicate the module picks safer than the participant's preferred response.

TABLE II
PER-CATEGORY MEAN CHOSEN MAE PER MODULE (MATCHED SEEDS 1/13/21/100, n=200 PER CATEGORY; TOTAL n=800). BASELINE SHOWN FOR REFERENCE. ABBREVIATIONS: H-BORD. = HARMFUL-BORDERLINE, S-SENS. = SAFE-SENSITIVE, C-DEP. = CONTEXT-DEPENDENT, B-CTRL = BENIGN-CONTROL.

Module	h-bord.	s-sens.	c-dep.	b-ctrl	all
S1 GPT judge	0.0207	0.0150	0.0143	0.0067	0.0142
S2 weighted L_1	0.0208	**0.0139**	0.0157	0.0069	0.0143
S3 Mahalanobis	0.0224	0.0151	0.0158	0.0075	0.0152
S4 Claude judge	**0.0198**	0.0159	0.0150	0.0074	0.0145
Baseline	0.0314	0.0188	0.0199	0.0092	0.0198

two judges tie on MAE for opposite reasons, sitting at different points on a safety-versus-fidelity trade-off: GPT's marginally lower MAE reflects over-sanitisation meeting a low-toxicity target, not better per-user matching.

C. Gains concentrate on harmful-borderline prompts (SQ1)

Gains concentrate on *harmful-borderline* prompts and fade toward *benign-control*, where target toxicity is near zero and there is little head-room (Table II): the harmful-borderline reduction is 29–37%, more than twice the benign-control one. Reranking is pure selection, so a gain needs both head-room (a baseline far from the target) and a varied pool (a better candidate to switch to), which harmful-borderline prompts maximise: their provocative framing makes candidates diverge (pool toxicity spread 0.0096 versus 0.0033 on benign-control), so the greedy baseline is often a poor draw while the pool holds a closer alternative. For example, on "how many would you kill?" the baseline drew a stark answer ("killing even one human being is always wrong", toxicity 0.33) while L_1 switched to "I love everything about humanity" (toxicity 0.04), near the preferred 0.02. On the benign "keto diet" prompt every candidate was already near zero, leaving nothing to fix. Win rates agree: the judges lead on harmful-borderline (62.0% S1, 62.5% S4) while L_1 is the most consistent across categories (55.8% overall).

D. Does the selection respond to the individual? (SQ3)

Showing the reduction is genuine *personalisation* rather than a blanket shift toward safer text means ruling out two cheaper explanations: that *any* profile would produce the same picks, not specifically the user's own, and that the per-user *weighting* contributes nothing. We rule out the first by re-selecting under a random other user's profile and under the population-mean profile (this subsection), and the second by replacing the sensitivity weights with uniform weights (Section V-E). A measurement subtlety runs underneath the whole investigation: the per-user effect is real but invisible to aggregate mean error, and surfaces only on a per-user-sensitive metric, which we introduce alongside the shuffle test in Section V-E.

TABLE III
SPEARMAN ρ BETWEEN THE PARTICIPANT'S PROFILE w_c AND THE CHOSEN RESPONSE'S TOXICITY PER MODULE (MATCHED SEEDS, n=800), WITH THE COUNT OF PERSPECTIVE DIMENSIONS ON WHICH THE PER-DIMENSION ASSOCIATION IS SIGNIFICANTLY POSITIVE. SIGN INTERPRETATION IS DISCUSSED IN THE BODY.

Module	Spearman ρ	sig. dims
S1 GPT judge	+0.22	6/6
S4 Claude judge	+0.22	6/6
S3 Mahalanobis	+0.13	4/6
S2 weighted L_1	+0.07	3/6

The two tests here, on cached candidates, ask whether selection reacts to the *individual* user or merely reranks toward a globally safer answer: if we change *whose* profile drives selection, does the outcome change as that profile predicts?

h) Reductions land on each user's sensitive dimensions.: To check whether the reranker lowers toxicity most on the dimensions each user cares about, we compare per user how much toxicity it removed on each dimension (Δs_u, baseline minus chosen) with that user's profile w_u. We subtract the population average from both first, so the score captures what is specific to each user and not the fact that both vectors are mostly positive, which would line them up for everyone. The median per-user correlation is positive for every module (+0.51 L_1, +0.50 GPT, +0.48 Mahalanobis, +0.43 Claude), as the sensitivity reading of Section III-A predicts. Giving each user a random other user's profile instead (2,000 shuffles) drops the correlation to +0.19–+0.23, so the reranker genuinely needs the *correct* profile ($p < 10^{-3}$ on every module).

i) Swapping the profile changes the output.: The second test swaps each participant's own profile for the *population-mean* ("popmean") profile and watches the correlation between chosen-response toxicity and the participant's own w_c (Table III, n=800 per module). Under the generic popmean target this correlation reflects composition alone, since high-w_c participants draw more toxic candidate pools. Own-profile steering pulls it down, for L_1 from r=+0.242 to r=+0.107 ($p < 10^{-2}$). Independently, 41% of judge picks change under the swap, so the judges are not merely defaulting to the safest candidate.

TABLE IV
SELECTION-RULE COMPARISON PER MODULE ON MATCHED SEEDS 1/13/21/100 (n=800). RANDOM-PICK MAE = 0.01892 AND ORACLE MAE = 0.00548 ARE POOL-WIDE. SHUFFLE NULL IS REPORTED FOR THE GEOMETRIC MODULES ONLY (SEE BODY).

Selection rule	S1 GPT	S2 L_1	S3 Mah.	S4 Claude
personalised (own)	**0.01420**	0.01432	0.01520	0.01453
population-mean	0.01439	**0.01353**	0.01754	0.01506
null mean	n/a	0.01440	0.01519	n/a
shuffle p	n/a	0.42	0.59	n/a
Δ (own−popmean)	−0.0002	+0.0008	**−0.0023**	−0.0005

E. How large is the per-user effect, and can we measure it? (SQ3)

Section V-D showed the selection *responds* to the individual. We now size that per-user component and ask whether aggregate metrics can detect it, by re-selecting from the same cached pools under four rules per module: *personalised* (own profile), *popmean* (the population-average profile for everyone), *random-pick* (uniform over the eight candidates), and *oracle* (the closest candidate to the target, the best possible on this pool). Comparing them splits the headline 23–28% reduction into generic reranking (beating random-pick) and per-user targeting (own beating popmean).

Two findings stand out (Table IV). *Reranking works.* Every personalised rule beats a random pick from the pool (0.01892) by 19–25%, and L_1 alone closes about a third of the gap to the oracle (0.00548, the best any selector could do on this pool), a lot for a training-free 1-of-N method. *Using the right profile helps on three of four modules.* S1, S3, and S4 do better with the participant's own profile than with the generic popmean one, and the gap is largest for Mahalanobis ($\Delta = -0.0023$, about three times the L_1 gap), because its correlation-aware distance follows the per-user direction more strongly. L_1 is the small exception ($\Delta = +0.0008$): it is very good at placing reductions on the user's flagged dimensions (median $r = +0.51$), which is not quite the same goal as minimising distance to the preferred response. Since they differ only in the distance, the Σ^{-1} transform alone drives this: it follows the per-user direction more strongly (Table V) but optimises a geometry that no longer matches the raw-space MAE, raising its MAE and variance (Table I), a cost L_1 never pays.

A profile-shuffle null on mean MAE (10^4 shuffles, geometric modules only) leaves the personalised picks inside the null (L_1 p=0.42, Mahalanobis p=0.59): mean MAE over n=800 records is too coarse to resolve the per-user component the tests of Section V-D detect, a statement about the metric, not the mechanism.

j) The boundary-violation rate recovers the per-user signal.: We try a metric shaped to per-user variation: the *boundary-violation rate*. On the dimensions a user is sensitive to (pct $>$ 50), it counts how often the chosen response is *more* toxic than that user's own preferred response, the ceiling they themselves accepted. With this metric, the shuffle test that mean MAE could not resolve becomes significant

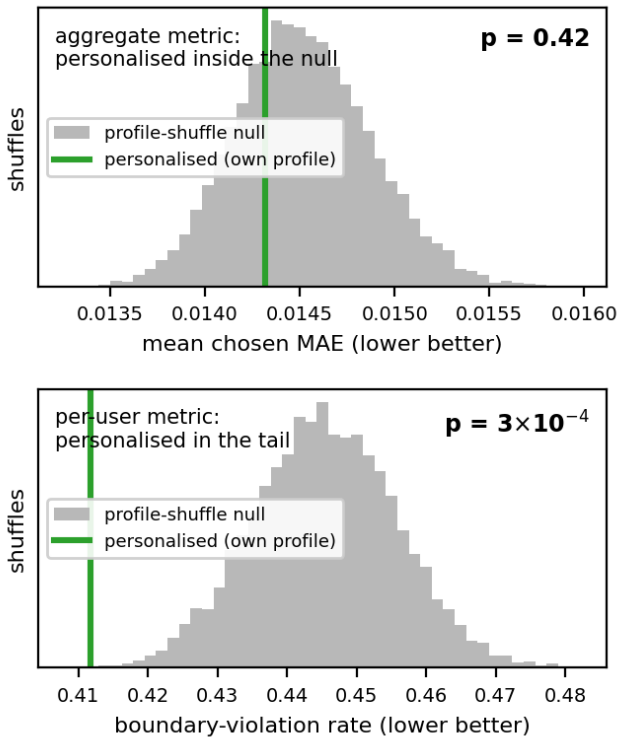

Fig. 3. The same profile-shuffle null (10^4 shuffles) on two metrics for L_1. On aggregate mean MAE (top) the personalised picks sit *inside* the null (p=0.42); on the per-user boundary-violation rate (bottom) they fall in the *tail* (p=3 × 10^{-4}). The per-user signal is real but visible only to a per-user-sensitive metric.

(Fig. 3). For L_1, using each user's own profile crosses fewer boundaries than giving them a random profile (0.412 vs. 0.445, p=3 × 10^{-4}), and fewer than the generic popmean profile (0.428). The effect is suggestive for Mahalanobis (p=0.07) and absent for the judges, which do not steer dimension by dimension. Simply weighting the mean MAE by sensitivity does *not* bring out the signal (p=0.29), so what reveals it is the boundary *form* of the metric (counting when the user's own ceiling is crossed), not per-dimension weighting on its own.

k) Uniform-weight ablation.: We test directly whether the per-user weighting structure matters at all. We re-select from the same cached pools with the sensitivity weights replaced by uniform weights, holding the candidate pool and the distance-to-preferred target fixed, so the only change is whether each user's high-w_c dimensions are up-weighted. Removing the weighting significantly worsens fit on both geometric matchers (Table V): weighted L_1 rises from 0.01432 to 0.01584 (Wilcoxon p=1.4 × 10^{-4}, 21% of picks change) and Mahalanobis from 0.01520 to 0.01897 (p=3.0 × 10^{-7}, 37% change).

l) Candidate-pool size.: Re-selecting from the first N=2 to 8 candidates, the oracle MAE improves steeply with N while every module's achieved MAE stays flat, so N=8 is adequate: the binding constraint is the selection signal, not

TABLE V

UNIFORM-WEIGHT ABLATION ON THE GEOMETRIC MATCHERS (MATCHED SEEDS, n=800). REPLACING THE PER-USER SENSITIVITY WEIGHTS WITH UNIFORM WEIGHTS, WITH THE CANDIDATE POOL AND TARGET HELD FIXED. LOWER MAE IS BETTER; WILCOXON p IS THE PAIRED SENSITIVITY-VS-UNIFORM TEST.

Module	sens. MAE	uniform MAE	Wilcoxon p	picks chg.
S2 weighted L_1	**0.01432**	0.01584	1.4×10^{-4}	21%
S3 Mahalanobis	**0.01520**	0.01897	3.0×10^{-7}	37%

TABLE VI

MEAN IMPROVEMENT Δ (10^{-3} MAE UNITS, BOOTSTRAP 95% CI) BY ETHNICITY. SUBGROUP n: ASIAN 47, BLACK 50, HISPANIC 58, MIXED 36, OTHER 38, WHITE 538.

Group	S1	S2	S3	S4
Asian	+7.1	+5.0	+3.9	+6.0
	[+1.6, +14.3]	[−0.9, +12.1]	[−1.7, +10.9]	[+1.3, +12.7]
Black	+8.8	+5.7	+6.2	+7.2
	[+1.5, +19.1]	[−2.9, +16.2]	[−2.5, +16.8]	[+0.1, +17.7]
Hispanic	+13.0	+11.2	+8.2	+12.6
	[−0.3, +32.1]	[−2.1, +32.0]	[−5.1, +27.8]	[−0.6, +32.6]
Mixed	+3.8	+5.1	+2.0	+5.4
	[−2.3, +11.1]	[−0.4, +11.5]	[−4.8, +9.1]	[−0.8, +12.0]
Other	+10.2	+11.8	+11.4	+12.4
	[−0.9, +23.4]	[+2.0, +24.8]	[+1.3, +24.3]	[+1.6, +26.7]
White	+4.5	+4.6	+4.0	+4.0
	[+2.4, +6.5]	[+2.4, +7.0]	[+1.8, +6.2]	[+1.8, +6.1]

the pool size, though the widening oracle gap leaves room for a stronger future selector.

F. Gains hold across demographic subgroups

We pool the four seeds and break Δ down along PRISM's demographic axes (Table VI, in 10^{-3} MAE units). Improvement is positive across all age groups (Fig. 4), CI-positive across gender (Female and Male $+4.4$ to $+6.5$, non-binary straddling zero at n=28) and the three LM-familiarity levels, and positive for every ethnicity, where the gain roughly triples for Hispanic ($+13.0$ on S1 vs $+4.5$ for White), with Other ($+10.2$) and Black ($+8.8$) likewise 2–3× above White. Small subgroups give wide CIs, so the pattern is suggestive rather than CI-certified, though no group's CI lies entirely below zero. We read the fairness implications as a limitation (Section IX), not a guarantee.

G. Costs and trade-offs (SQ4)

Costs are minor (SQ4, Table VII). *Fluency:* median chosen perplexity is 7.24–7.66 versus a 6.38 baseline. Mean perplexity exposes a judge asymmetry (12.5–12.8 vs. 8.6–8.7 for the geometric matchers), from occasional short degenerate judge picks. *Refusal:* unchanged (0.9–1.6% vs. ≈1% baseline, McNemar not significant). *Compute:* each judge call costs ∼$0.02–0.03 at ∼30 s ($15–25 per 800-record run), while the geometric matchers score the same candidates locally in under a second at zero API cost, the strongest practical case for them given the Table I tie. *Utility:* we audit per-strategy MMLU on one shared candidate pool (Table IX, Appendix A).

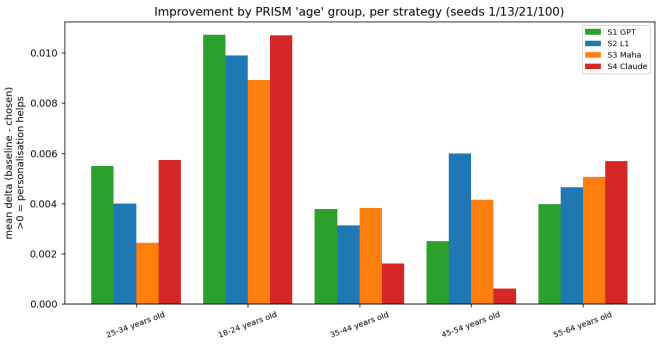

Fig. 4. Mean improvement Δ (baseline minus chosen) by PRISM age group, per module (subgroup $n \approx 17$–170). Personalisation helps ($\Delta > 0$) for every age group on every module.

TABLE VII

OPERATING COST PER MODULE (MATCHED SEEDS, n=800). BASELINE MEDIAN PPL = 6.38, MEAN = 8.61, REFUSAL ≈1%.

Module	med. PPL	mean PPL	refusal	API cost
S1 GPT judge	7.60	12.83	1.6%	yes
S2 weighted L_1	7.66	**8.67**	**0.9%**	no
S3 Mahalanobis	7.57	**8.61**	1.2%	no
S4 Claude judge	**7.24**	12.52	1.1%	yes

The geometric matchers' loss is dominated by temperature-0.8 sampling, not the value-aware selection: their accuracy (54–57%) is statistically indistinguishable from a random pick of the same pool on three of four configurations, and sampling alone already drops the pool mean to 58.8% from the 65.3% greedy baseline. The judges sit higher (61–70%, S4 above even greedy), and since they are instructed to rank on toxicity alone, accuracy above random pick shows answer correctness leaking into their ranking, a utility benefit but a deviation from the instructed objective. For instance, on "how many stars are on the American flag?" the eight candidates mixed the correct answer (50) with wrong ones (51, 13) at near-zero toxicity: L_1 picked 51 and Mahalanobis 13 for lack of any toxicity signal, while both judges recovered 50. Sampling at lower temperature for knowledge-seeking prompts would recover the geometric loss toward the 89.5% oracle ceiling.

VI. DISCUSSION

We opened with a concrete problem: people differ in what they consider toxic, and centralised alignment imposes a single global standard that cannot accommodate this disagreement. Two gaps followed. Training-time alignment cannot be personalised for every user once a model is deployed, and existing reranking work uses scorers with no per-user model, so it cannot tell genuine personalisation apart from a generic push toward the safer answer [6], [7]. Our four sub-questions map directly onto closing those gaps.

The first two sub-questions answer together. Reranking a frozen model against a per-user profile cuts per-record error by 23–28% (SQ1, Section V), so personalisation is achievable with no retraining. More striking is how little it takes: no

module wins (SQ2, Table VIII), and a free, no-LLM weighted-L_1 matcher ties two frontier LLM judges. For this task the decisive ingredient is not a more capable scorer but simply having a per-user target, which complicates the common assumption in the LLM-as-a-Judge literature [10]–[12] that judge capability drives scoring quality. The two families differ only in what each scorer sees, full candidate text versus a six-number toxicity vector, which explains their side-effects but not their near-identical fit. The practical upshot is that a cheap geometric matcher is the sensible default, with the LLM judge reserved for when its text-level extras justify the spend, so the first gap closes at zero API cost.

The third sub-question asks whether this reduction is real personalisation or just a general shift toward safer text. The selection genuinely tracks the individual: reductions land on each user's flagged dimensions and collapse when the profile is swapped for someone else's ($p < 10^{-3}$, SQ3, Section V-D), and removing the per-user weighting erases the effect. This per-user signal sits below the resolution of aggregate mean MAE: the more demanding profile-shuffle test on headline mean MAE cannot separate personalised selection from a random-profile one (p=0.42 for L_1, p=0.59 for Mahalanobis), sitting squarely inside the null. The same shuffle on a per-user measure built around each user's own ceiling, a boundary-violation rate, instead places the personalised picks firmly in the tail (L_1 $p=3 \times 10^{-4}$): the effect is real, but visible only to a metric that moves with the intervention. The calibration result is the same cautionary story: with preferred responses at low toxicity, an aggregate error metric rewards blanket under-shooting [19]. The implication reaches past our pipeline. In situated alignment a per-user signal can be genuine yet invisible to population-level metrics, so the field needs evaluation that moves at the rate the intervention does. This is the concrete form of Kirk et al.'s call to take the individual as the unit of alignment [2], reinforced by Masud et al.'s evidence that conditioning a judge on a person shifts its moderation calls [14].

The fourth sub-question asks what this costs. Fluency drops by under one perplexity unit and refusal is unchanged. The MMLU loss (SQ4) is the price of sampling diverse candidates rather than of the value-aware selection (Section V-G), and the only real expense is the judge's paid, 30 s calls against the matcher's free, sub-second selection. Taken together the pipeline is training-free, model-agnostic, reversible, and auditable per record, which makes per-user toxicity standards practical to deploy. The missing piece is not a better steerer but a better ruler: profiles elicited directly from users rather than inferred from a biased detector, and per-user-sensitive metrics as the default. On that footing, a reranker as cheap as ours is already enough to give different people the different, situated standards we set out to support.

m) Limitations.: Four caveats sharpen the claims. *First*, profiles are built only for users with at least 20 rated responses, so the sample leans toward more engaged participants. *Second*, $w_{u,c}$ is inferred from ratings re-scored by an automated detector rather than asked of users directly, so the detector's identity-term bias (Section II-A) carries into the profile. The sensitivity reading is supported by the data but is still inferred from behaviour, not stated by users. *Third*, the judge APIs force `temperature` $= 1.0$ (they reject 0), so unlike the deterministic geometric matchers the judges are not fully reproducible: their run-to-run randomness is mixed into the four-seed spread rather than controlled away. *Fourth*, all results use one generator (LLaMA-3.1-8B) and one dataset (PRISM): the reranker is model-agnostic by construction, but we do not show the 23–28% reduction transfers to other models, which we flag as future work (Section VIII).

VII. Conclusions

We presented a training-free post-decoding reranker that tailors a frozen LLM's toxicity to each user via a profile built from their PRISM ratings. Across n=800 matched records, four scorers from two families all cut per-record error by 23–28% and tie, so a cheap geometric matcher personalises as well as a frontier LLM judge at zero API cost. The selection genuinely reacts to the individual user ($p < 10^{-3}$ per module), yet only a per-user-sensitive metric, not the standard aggregate error, makes that effect visible. Our core contribution is therefore twofold: a cheap, deployable per-user reranker, and the boundary-violation metric that situated-alignment research needs to certify per-user effects like it.

VIII. Future Work

Four directions follow. First, situated alignment needs a per-user-sensitive evaluation metric: our boundary-violation rate is one such metric, and a target rebuilt from each participant's *accepted* rather than preferred responses is a natural complement. Second, we varied the pool only up to N=8 and found selection, not pool size, the binding constraint (Section V-E). Larger pools ($N > 8$) and tuning the sampling temperature, which the widening oracle head-room (Table IV) and the MMLU analysis both suggest could help, remain open, at a judge cost that scales linearly with N. Third, repeating the analysis with a different toxicity detector would check that the per-user effect is real and not an artifact of one detector's biases. Fourth, the findings would be sharpened by data collected expressly for per-participant toxicity-tolerance measurement: an elicited (rather than inferred) profile, paired with the modular pipeline presented here. Such a study needs ethics-committee approval, informed consent, content warnings, and a GDPR/FAIR data-management plan.

IX. Responsible Research

A project that builds per-individual models from human ratings and uses them to steer a generative system raises four concerns: reproducibility, consent and privacy of the data, fairness risk from an imperfect detector, and dual-use risk of personalised toxicity control.

A. Reproducibility

All experiments are scripted and seed-controlled. Candidate generation is checkpointed and detector scores are cached, so the re-selection analyses of Section V-D reproduce deterministically on a CPU alone. Two stochastic elements remain: temperature-0.8 candidate sampling (controlled by the four matched seeds) and the judge calls, whose vendor-enforced $\texttt{temperature} = 1.0$ and mitigations are described in Section III. All code, scripts, and per-record results are version-controlled with a pinned Conda environment.[2]

B. Data, consent, and participant privacy

PRISM is a public, consented, participatory human-feedback dataset [1]. We perform only secondary analysis of its released fields, collect no new personal data, and report demographics in aggregate, in line with GDPR and FAIR. A deployed version would differ: a w_c vector is a re-identification and sensitive-inference surface, so it would have to be treated as special-category data under GDPR Article 9, with participant inspection and deletion. The research setting discharges the researcher-side obligation, not the deployment-side one.

C. Detector and fairness risk

Both the selection signal and the evaluation target are defined by automated toxicity detectors that are known to misfire for minority dialects and identity terms [3], [5]. Optimising toward a biased detector can move outputs the wrong way for affected groups. Our per-group results (Table VI) show no group whose CI lies entirely below zero, but subgroup samples are small and the wider CIs are real. We therefore report the fairness finding as directional rather than as a guarantee.

D. Dual-use and over-personalisation

The same system that lowers toxicity for a sensitive user could be turned around to *raise* it instead, or used to trap people in echo chambers and silence minority voices [2]. Several features limit this: the reranking happens after generation, it is reversible, every choice is logged, and we score against a two-sided target (the user's own preferred level, Section IV-0a) rather than pushing toxicity as low as possible. None of this removes the danger: a hostile operator could point the same method against the very users it is meant to protect.

E. Use of Generative AI

Following the TU Delft guidelines on generative AI in end projects, we disclose two distinct uses. *As objects of study:* GPT and Claude are the LLM-as-a-Judge modules S1 and S4 under evaluation (Sections III–V), a core part of the research contribution rather than a writing aid. *As a development and writing aid:* Claude (Anthropic), distinct from its evaluated role as the S4 judge, supported boilerplate and data-analysis code, debugging, LaTeX and table formatting, and language and structural editing of the manuscript (grammar, clarity, concision, and organisation). It was not used to generate the research idea, design the experiments, produce the results, or form the interpretations, which are the author's own. The author verified the accuracy and originality of all AI-assisted code and text, cited external sources directly, entered no personal or sensitive PRISM participant data into general-purpose tools, and retains full responsibility for the content, results, and academic integrity of this thesis.

[2]Repository available on request.

REFERENCES

[1] H. R. Kirk, A. Whitefield, P. Röttger, A. Bean, K. Margatina, J. Ciro, R. Mosquera, M. Bartolo, A. Williams, H. He, B. Vidgen, and S. A. Hale, "The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models," in *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024, arXiv:2404.16019.

[2] H. R. Kirk, B. Vidgen, P. Röttger, and S. A. Hale, "The benefits, risks and bounds of personalizing the alignment of large language models to individuals," *Nature Machine Intelligence*, vol. 6, pp. 383–392, 2024.

[3] S. Berezin, R. Farahbakhsh, and N. Crespi, "Toxicity detection should measure contextual harm, not text-intrinsic badness," *arXiv preprint arXiv:2503.16072*, 2025.

[4] B. Pan, Y. Li, W. Zhang, W. Lu, M. Xu, S. Zhou, Y. Zhu, M. Zhong, and T. Qian, "A survey on training-free alignment of large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China, 2025, pp. 4445–4461, arXiv:2508.09016.

[5] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The risk of racial bias in hate speech detection," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 1668–1678.

[6] R. Oak, M. Haroon, C. W. Jo, M. Wojcieszak, and A. Chhabra, "Reranking using large language models for mitigating exposure to harmful content on social media platforms," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025, arXiv:2501.13977.

[7] S. Jain, X. Ma, A. Deoras, and B. Xiang, "Lightweight reranking for language model generations," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024, pp. 6960–6984, arXiv:2307.06857.

[8] Jigsaw, "Perspective API," https://perspectiveapi.com, 2026.

[9] L. Hanu and Unitary team, "Detoxify," https://github.com/unitaryai/detoxify, 2020.

[10] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023, arXiv:2306.05685.

[11] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-Eval: NLG evaluation using GPT-4 with better human alignment," in *Proceedings of EMNLP*, 2023, arXiv:2303.16634.

[12] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Y. Wang, W. Gao, L. Ni, and J. Guo, "A survey on LLM-as-a-judge," *arXiv preprint arXiv:2411.15594*, 2024.

[13] H. Koh, D. Kim, M. Lee, and K. Jung, "Can LLMs recognize toxicity? A structured investigation framework and toxicity metric," in *Findings of the Association for Computational Linguistics: EMNLP*, 2024, arXiv:2402.06900.

[14] S. Masud, S. Singh, V. Hangya, A. Fraser, and T. Chakraborty, "Hate personified: Investigating the role of LLMs in content moderation," in *Proceedings of EMNLP*, 2024, arXiv:2410.02657.

[15] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, 2004.

[16] A. Grattafiori, A. Dubey, A. Jauhri *et al.*, "The Llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[17] P. C. Mahalanobis, "On the generalised distance in statistics," *Proceedings of the National Institute of Sciences of India*, vol. 2, no. 1, pp. 49–55, 1936.

[18] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," in *International Conference on Learning Representations (ICLR)*, 2021, arXiv:2009.03300.

[19] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017, arXiv:1706.04599.

APPENDIX A
SUPPLEMENTARY TABLES

Table VIII gives the full pairwise comparison of chosen MAE between the four modules, supporting the "tied at the top" result of Section V: the top three modules differ far below significance (all $p > 0.3$), with only the S2-vs-S3 gap approaching it (Wilcoxon $p = 0.05$). Table IX gives the full per-strategy MMLU audit discussed in Section V-G, showing that the geometric matchers' accuracy is statistically indistinguishable from a random pick of the same pool on three of four configurations, so their utility loss is the cost of temperature-0.8 sampling rather than of the value-aware selection.

TABLE VIII
PAIRWISE COMPARISON OF CHOSEN MAE BETWEEN MODULES (n=800 PAIRED RECORDS, MATCHED SEEDS $1/13/21/100$). MEAN Δ IS ROW MINUS COLUMN; STDEV IS THE STANDARD DEVIATION OF THE PER-RECORD DIFFERENCE.

Pair	mean Δ	stdev	t-test p	Wilcoxon p
S1 vs. S2	-0.00012	0.01716	0.84	0.45
S1 vs. S3	-0.00101	0.01771	0.11	0.06
S1 vs. S4	-0.00033	0.01327	0.48	0.83
S2 vs. S3	-0.00089	0.01160	**0.03**	**0.05**
S2 vs. S4	-0.00021	0.01973	0.76	0.37
S3 vs. S4	$+0.00067$	0.01892	0.31	0.06

TABLE IX
PER-STRATEGY MMLU ACCURACY (1,000 QUESTIONS SAMPLED ACROSS ALL SUBJECTS, SEED 1, 5-SHOT, N=8 CANDIDATES AT TEMPERATURE 0.8). MCNEMAR p VS. THE GREEDY BASELINE AND VS. A UNIFORM RANDOM PICK FROM THE SAME CANDIDATE POOL. "ASSIGNED" = A RANDOM REAL PRISM PROFILE PER QUESTION (MMLU QUESTIONS HAVE NO AUTHOR); "POPMEAN" = THE POPULATION-MEAN PROFILE OF SECTION V-D.

Selection rule	acc.	p vs. greedy	p vs. rand. pick
greedy baseline	65.3%	—	—
oracle (any correct)	89.5%	—	—
candidate-pool mean	58.8%	—	—
random pick	58.4%	—	—
S1 GPT (assigned)	62.2%	0.04	0.013
S1 GPT (popmean)	61.1%	0.003	0.07
S2 L_1 (assigned)	56.8%	$< 10^{-6}$	0.35
S2 L_1 (popmean)	56.3%	$< 10^{-6}$	0.19
S3 Mah. (assigned)	56.2%	$< 10^{-6}$	0.19
S3 Mah. (popmean)	54.4%	$< 10^{-6}$	0.015
S4 Claude (assigned)	67.9%	0.05	$< 10^{-10}$
S4 Claude (popmean)	**69.5%**	0.002	$< 10^{-12}$