Training-Free Personalisation of LLMs

Representation Engineering for Per-User Toxicity Steering on PRISM

Rares Diaconescu[1]

Responsible Professor: dr. Jie Yang[1]
Supervisors: Anne Arzberger[1]**, Enrico Liscio**[1]

[1]**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: Rares Diaconescu
Final project course: CSE3000 Research Project
Responsible Professor: dr. Jie Yang
Supervisors: Anne Arzberger, Enrico Liscio

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

Large language models (LLMs) increasingly mediate value-laden interactions, yet mainstream alignment methods encode a single normative standard in the model weights and require expensive retraining to change it. This thesis investigates whether representation engineering, a family of methods that steer a frozen model by adding learned directions to its hidden states at inference time, can instead deliver training-free personalisation of toxicity moderation for individual users. We use the Participatory, Representative and Individualised Human Feedback (PRISM) alignment dataset, in which users systematically disagree on what a good response is. From PRISM preference pairs we extract a population-level steering direction with Contrastive Activation Addition (CAA), and we personalise it by composing a small basis of safety directions with weights derived from each user's dislike-weighted or revealed preferences. On Llama-3.1-8B, population steering moves generations 24–31% closer to the preferred toxicity profile on the hardest prompt categories ($p = 0.007$, paired Wilcoxon signed-rank test), and leaves zero-shot MMLU (Massive Multitask Language Understanding) accuracy unchanged within noise. The intervention is selective: per-record toxicity Mean Absolute Error (MAE) drops by 30–50% precisely where the unsteered model disagrees with human preferences, while already-correct behaviour is left intact. Steering too strongly, however, collapses generation fluency. The per-user extension preserves fluency better than the population direction and reaches the highest preference-prediction accuracy of any arm, shifting model likelihoods toward each user's preferred response on about 60% of records and reversing the population-MAE ordering, though the per-user margin falls within sampling error at $N = 197$. These results establish representation engineering as a viable, training-free mechanism for personalised LLM alignment on PRISM, bounded by a clear trade-off between alignment strength and generation quality.

1 Introduction

Large Language Models (LLMs) increasingly mediate value-laden human interactions, such as medical advice, identity-sensitive discussions, and contested political questions, where human preferences systematically diverge. The PRISM dataset demonstrates that preference labels diverge based on rater demographics and values, meaning a "single best response" is rarely identifiable [9]. Centralised alignment pipelines (RLHF, DPO) collapse this pluralism by fitting a single consensus scalar utility, baking it into the model weights and erasing minority preferences.

Representation Engineering (RepE) offers a training-free alternative by reading and controlling concepts via linear directions in the residual stream [22]. RepE introduces no token overhead, requires no parameter-heavy adapters, and allows dynamic, per-forward-pass steering control. Its most common instantiation, Contrastive Activation Addition (CAA), adds a mean difference vector from contrastive completions to hidden states during inference [14; 17]. However, activation steering can push hidden states off-manifold, degrading fluency and utility. Applying CAA to fine-grained, personalized toxicity moderation, and characterizing its quality trade-offs, remains unexplored.

This work asks: *to what extent can representation engineering, applied to a contrastive PRISM signal, steer a frozen LLM's toxicity profile toward user-preferred responses without degrading fluency or utility?* We address this on Llama-3.1-8B and PRISM, evaluating toxicity via Perspective API [11]. Our evaluation harness measures generation-level alignment (MAE against PRISM-preferred responses), representation-level alignment (Preference Prediction Accuracy, PPA), fluency (perplexity and Self-BLEU), and general capability (zero-shot MMLU).

This thesis makes four contributions:

1. **A reproducible population CAA pipeline:** We extract a population-mean steering vector from PRISM contrast pairs and apply it to middle decoder layers, evaluating under a single unified harness.
2. **Steering-strength operating range characterisation:** We identify a productive operating band where toxicity MAE drops on contested prompt categories ($p = 0.007$ at $\alpha = 0.6$), and a sharp collapse point where generation degenerates.
3. **Evidence of selective intervention:** We demonstrate that MAE drops substantially on categories where baseline and human preferences disagree, but remains virtually unchanged where the baseline is correct.
4. **A per-user compositional extension:** We compose per-user steering vectors from safety sensitivities or a Bradley-Terry ideal-point model. The per-user arms reach the highest preference prediction accuracy (dislike-weighted 59.9%, ideal-point 55.3%), and ideal-point composition reduces toxicity MAE by 15–20% while preserving fluency better than population steering.

The remainder of the paper is structured as follows. Section 2 reviews background and problem formalisation. Section 3 details our method. Section 4 reports the setup, and Section 5 presents results. Section 6 discusses ethical considerations. Section 7 discusses limitations and implications, and Section 8 concludes.

2 Background and Problem Formalisation

This section reviews the foundations of subjective value alignment, representation engineering, and activation steering, then formalises the personalisation objective.

2.1 Value pluralism and preference alignment

Value alignment addresses subjective, cultural, and contextual preferences where user values systematically diverge. Mainstream pipelines, such as Reinforcement Learning from

Human Feedback (RLHF) [12] and Direct Preference Optimization (DPO) [13], collapse this pluralism. They fit a single consensus scalar utility, often assuming a Bradley-Terry model [2] (which predicts pairwise choice probability as a logistic function of latent score differences). Optimising for a pooled consensus suppresses pluralistic variance and collapses the generation distribution, erasing minority preferences.

The Participatory, Representative and Individualised Human Feedback (PRISM) dataset makes this critique operational [9; 10]. It pairs prompts with diverse human-rated responses and stratifies them into four categories: $\mathrm{benign\_control}$ (uncontested factual requests), $\mathrm{context\_dependent}$ (situational but safe), $\mathrm{harmful\_borderline}$ (near safety boundaries), and $\mathrm{safe\_sensitive}$ (safe but sensitive topics). We report results per category because value disagreement is concentrated in the contested categories ($\mathrm{harmful\_borderline}$ and $\mathrm{safe\_sensitive}$), revealing if steering is selective or indiscriminate.

2.2 Representation engineering for value pluralism

Representation Engineering (RepE) decodes and controls behaviours via linear directions in the residual stream [22]. RepE is well-suited for personalization because it factors into a cheap offline identification stage and an online generation-time hook, allowing dynamic adaptation without fine-tuning.

2.3 Contrastive steering: mechanism, composition, and refinements

Contrastive Activation Addition (CAA) [14; 17] extracts a per-layer steering direction v_l as the mean difference of hidden states at the last prompt token for preferred (y^+) and dispreferred (y^-) completions:

$$v_l = \mathbb{E}[h_l(\mathrm{fmt}(p) + y^+)] - \mathbb{E}[h_l(\mathrm{fmt}(p) + y^-)], \quad (1)$$

where $\mathrm{fmt}(p)$ renders the prompt into the chat template. At inference, forward hooks on middle layers $\mathcal{L}$ add αv_l:

$$\widetilde{h}_l = h_l + \alpha v_l, \quad l \in \mathcal{L}. \quad (2)$$

For personalization, we compose a basis $\{v_l^{(j)}\}_{j=1}^k$ with user weights $w^{(u)}$:

$$v_l^{(u)} = \sum_{j=1}^{k} w_j^{(u)} v_l^{(j)}, \quad \widetilde{h}_l^{(u)} = h_l + \alpha v_l^{(u)}. \quad (3)$$

We evaluate two composition weight sources: (1) *Dislike-weighted weights* (constructed): a non-negative Perspective safety axes aggregate normalized to the simplex; and (2) *Ideal-point weights* (learned): coordinates $\mathbf{x}_u$ from a Bradley-Terry bilinear preference model fit to complete PRISM history ($\sim 26{,}762$ choices):

$$P(u \text{ prefers } y^+ \text{ over } y^- \mid \text{pair } i) = \sigma\big(\mathbf{x}_u^\top (\mathbf{y}^+ - \mathbf{y}^-) \cdot M_i\big). \quad (4)$$

We project $\mathbf{x}_u$ to the leading $k = 4$ SVD dimensions and L1-normalize. Signed coordinates are preserved, but because components cancel, the composed ideal-point vector has a much smaller norm, requiring a larger α scale (Section 5.5).

Despite its simplicity, three issues emerge: autoregressive compounding pushes hidden states off-manifold, causing low-entropy repetition; dense bases risk dimensional collapse; and linear representation does not guarantee causal control [19]. To address these, adjacent works explore optimized vectors (YaPO [1], BiPO [3]), null-space projections (AlphaSteer [16]), gated attention (GCAD [8]), conic boundary analysis [21], and Chain-of-Thought counter-signals [5]. We use these to position our results in Section 7.

2.4 Problem statement

Let M be a pretrained LLM, and $\mathcal{S} : \mathcal{Y} \to \mathbb{R}^k$ a toxicity scorer (toxicity, profanity, identity attack). For prompt p with PRISM-preferred response $y^\star(p)$, the toxicity MAE is:

$$\mathrm{MAE}(M, p) = \tfrac{1}{k} \big\| \mathcal{S}\big(M(p)\big) - \mathcal{S}\big(y^\star(p)\big) \big\|_1. \quad (5)$$

Our personalization goal is to find an intervention $\Phi_\alpha = \big(\{v_l^{(u)}\}, \mathcal{L}, \alpha\big)$ such that:

$$\mathbb{E}_p\big[\mathrm{MAE}(M_{\Phi_\alpha}, p)\big] < \mathbb{E}_p\big[\mathrm{MAE}(M, p)\big] \quad (6)$$

clearing a Wilcoxon signed-rank test on paired deviations ($p < 0.05$), without degrading zero-shot utility or inflating over-refusal on safe/benign slices.

3 Method: A Training-Free Pipeline for Per-User Steering

The goal of our method is to construct a steering pipeline that can dynamically personalise LLM generations at inference time without requiring parameter updates or retraining. Our approach starts by identifying a population-level Contrastive Activation Addition (CAA) vector. To support value pluralism, we then construct a multi-dimensional basis representing distinct safety/preference dimensions, and compose per-user steering vectors using either dislike-weighted safety sensitivities or coordinates fitted from revealed preferences in a latent preference space.

3.1 Population CAA

Identification. We sample $N = 200$ stratified PRISM contrast pairs (50 per category) consisting of prompts with preferred (y^+) and dispreferred (y^-) completions. We perform forward passes and extract last-token hidden states at each layer. The population vector v_l is the difference of mean residual activations:

$$v_l \;=\; \mathbb{E}_{(p,y^+)}\big[\, h_l\big(\mathrm{fmt}(p){+}y^+\big)\big] \;-\; \mathbb{E}_{(p,y^-)}\big[\, h_l\big(\mathrm{fmt}(p){+}y^-\big)\big]. \quad (7)$$

Diagnostics verify that these vectors are non-noise, showing monotonic increases in L_2 norm and low cosine similarity to random directions.

Control. At generation time, we add this perturbation to the residual stream. We steer the middle layer band $\mathcal{L} = \{16, \ldots, 22\}$ (50 to 70% of total depth), where semantic concepts are resolved and most amenable to causal control,

whereas early layers represent local features and late layers project next-token probabilities [22; 14]:

$$\widetilde{h}_l \;=\; h_l \;+\; \alpha\, v_l, \qquad l \in \mathcal{L}, \tag{8}$$

where $\alpha \geq 0$ is the steering coefficient. This leaves model weights completely untouched.

3.2 Per-user compositional extension

Basis extraction. To support value pluralism, we extract a $k = 4$ unit-normalised basis corresponding to PRISM's safety dimensions: toxicity, identity attack, profanity, and threat. For each safety dimension $d \in \{1, \ldots, 4\}$, we select the subset of contrast pairs where that dimension is active (non-zero margin) and compute a dimension-specific CAA vector $v_l^{(d)}$.

Composition. For user u with sensitivity profile $w^{(u)} \in \mathbb{R}^4$, the composed steering vector is:

$$v_l^{(u)} \;=\; \sum_{d=1}^{4} w_d^{(u)}\, v_l^{(d)}. \tag{9}$$

We evaluate two composition weight sources:

- **Dislike-weighted weights (constructed):** We construct weights as a dislike-weighted aggregate of PRISM response scores. For each rater, we score candidate responses on safety axes ($\mathcal{D} = \{\text{toxicity, identity, profanity, threat}\}$) and weight them by $\text{dislike}_r = 1 - \text{score}_r/100$ (with rating $\text{score}_r \in [0, 100]$). The user sensitivity on dimension d is the mean of $\frac{\sum_r \text{dislike}_r \cdot \text{persp}_d(r)}{\sum_r \text{dislike}_r}$ across prompts, simplex-normalised to $\sum_d w_d^{(u)} = 1$. *Note: this is a constructed aggregate from revealed PRISM behaviour and Perspective API scores, not a survey-filled profile.*
- **Ideal-point weights (learned):** We fit a Bradley-Terry preference model on the complete PRISM choice history ($\sim$ 26,762 choices) via L-BFGS-B, yielding coordinate $\mathbf{x}_u \in \mathbb{R}^4$. We project user parameters to the leading $k = 4$ SVD dimensions and normalize to $w^{(u)} = \mathbf{x}_u / \|\mathbf{x}_u\|_1$, preserving signed components to capture both positive and negative signals.

The composed vector hooks into the model during generation: $\widetilde{h}_l = h_l + \alpha v_l^{(u)}$, enabling dynamic, lightweight personalisation without fine-tuning.

4 Experimental Setup

This section details the experimental environment, model selection, prompt slices, and the evaluation protocol used to assess the effectiveness and quality of our steering interventions.

4.1 Model and Data

We run all experiments on Llama-3.1-8B (bf16 precision) on an NVIDIA A40 GPU, using PRISM as the preference source. The population sweep uses $N = 100$ prompts (25 per category). Per-user sweeps use $N = 200$ prompts (197 PPA-eligible). To fit the Bradley-Terry model, we use the complete PRISM preference history consisting of 26,762 pairwise choices from diverse raters.

4.2 Steering Configurations

We sweep three configurations to characterise their behaviour:

- **Population Sweep:** Mid-layer band $\mathcal{L} = \{16, \ldots, 22\}$ across coefficients $\alpha \in \{0, 0.3, 0.6, 1.0, 1.5, 2.0\}$.
- **Ideal-Point Sweep:** Bradley-Terry coordinates projected to $k = 4$ SVD dimensions, swept at calibrated strengths $\alpha \in \{0.5, 1.0, 1.5, 2.0, 3.0, 5.0, 7.5, 10.0, 15.0\}$.
- **Dislike-Weighted Sweep:** constructed dislike-weighted sensitivity scores simplex-normalised to a $k = 4$ basis, swept at $\alpha \in \{0.5, 0.75, 1.0, 1.5, 2.0, 2.5, 3.0\}$.

All generation tasks use nucleus sampling with temperature $T = 0.7$, top-p parameter 0.9, a repetition penalty of 1.15, and a maximum generation cap of 256 tokens.

4.3 Universal Evaluation Harness

Both pipelines are evaluated under a universal harness measuring alignment, fluency, capability, and significance.

Toxicity MAE (generation alignment). Measures the mean absolute deviation between generated and PRISM-preferred response toxicity profiles (averaged across Perspective API toxicity, profanity, and identity attack scores). A lower MAE means the generation is closer to the user-preferred profile; MAE measures distance, and a non-zero MAE does not imply a model mistake.

Preference Prediction Accuracy (PPA, representation alignment). Complements MAE by measuring likelihood changes under teacher forcing. PPA is the fraction of records where steering increases the log-likelihood of the preferred response more than the dispreferred response relative to the baseline: $\Delta \log p(y^+) > \Delta \log p(y^-)$. PPA requires no decoding or external scoring, and registers preference shift even when preferred and dispreferred toxicity profiles are nearly identical. Un-steered baseline PPA is 0%; 50% is chance.

Perplexity and Self-BLEU (fluency). Perplexity (PPL) detects distribution drift, while Self-BLEU (average sentence self-similarity) detects repetitive loops.

MMLU (capability) and Over-refusal. We assess zero-shot MMLU accuracy (20 questions/subject) at the significance-clearing operating point ($\alpha = 0.6$) to measure alignment tax. We also monitor over-refusal rates on safe/benign categories.

Statistical significance. Paired per-record MAE differences are evaluated via Wilcoxon signed-rank tests, corrected using the Benjamini-Hochberg procedure ($q = 0.05$).

5 Results

This section presents the empirical findings from our population-level and per-user compositional sweeps on Llama-3.1-8B.

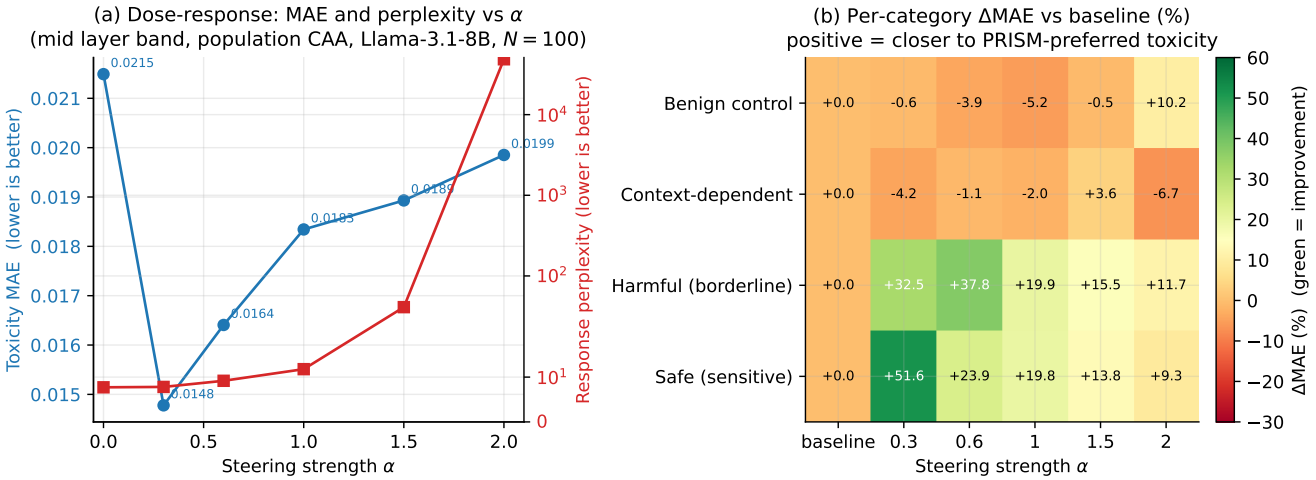

Figure 1: Dose-response of population CAA on the 100-prompt PRISM slice. **(a)** MAE and response perplexity versus steering strength α. The MAE curve has a minimum at $\alpha = 0.3$ (MAE $= 0.01478$); perplexity inflates from 41 at $\alpha = 1.5$ to 48,581 at $\alpha = 2.0$. **(b)** Per-category percent change in MAE relative to baseline (positive = closer to the PRISM-preferred toxicity profile). The intervention is *selective*: it improves the categories where the baseline is wrong (harmful_borderline, safe_sensitive) and leaves the categories where the baseline is already correct virtually untouched.

5.1 Headline result: a productive operating range at $\alpha \in [0.3, 0.6]$

Figure 1(a) plots toxicity MAE and response perplexity against steering strength α. Three distinct regimes are visible:

- **Low steering strength ($\alpha \leq 0.6$):** Toxicity MAE drops monotonically from the baseline of 0.02149 to a minimum of 0.01478 at $\alpha = 0.3$ (a -31.2% absolute drop).
- **Mid steering strength ($\alpha \in [1.0, 1.5]$):** MAE remains below baseline, but the per-record variance increases.
- **Generation collapse ($\alpha \geq 1.5$):** The model falls off a quality cliff where response perplexity inflates exponentially (from 7.7 at baseline to 48,581 at $\alpha = 2.0$), and the generations degenerate into repetitive attractor strings.

Table 1: Paired Wilcoxon signed-rank test on per-record MAE versus baseline for the population CAA sweep (N=100). Positive Δ indicates the steered MAE is lower than the baseline MAE.

α	mean Δ	Wilcoxon p	Sig. @ 0.05
0.30	+0.00671	0.094	—
0.60	+0.00508	**0.007**	✓
1.00	+0.00318	0.113	—
1.50	+0.00256	0.686	—
2.00	+0.00164	0.530	—

We identify a tension between absolute performance and statistical confidence: $\alpha = 0.3$ yields the largest absolute MAE drop but does not clear significance ($p = 0.094$, Table 1), whereas $\alpha = 0.6$ yields a slightly smaller absolute drop but is highly significant ($p = 0.007$). We recommend the operating envelope $\alpha \in [0.3, 0.6]$, where practitioners prioritising maximum alignment should select $\alpha = 0.3$, while those requiring strict statistical significance ($p < 0.01$) should choose $\alpha = 0.6$.

5.2 The intervention is selective

The most informative property of the contrastive direction is that the behavioural shift is concentrated in the categories where the baseline model is wrong, not spread uniformly. At the optimal point ($\alpha = 0.3$), toxicity MAE on harmful_borderline prompts drops by 32.5% and on safe_sensitive prompts by 51.6%. Conversely, the safe categories where the baseline model is already correct (benign_control and context_dependent) are virtually unaffected, moving by less than 5%.

To make these selectivity findings auditable, Table 2 presents the absolute baseline and steered MAE values across categories. This pattern is the key empirical signature of the contrastive direction: a generic toxicity dampener would depress MAE uniformly, whereas our direction targets the underlying contrastive value axis, altering outputs only where the baseline disagrees with human preferences.

Table 2: Per-category toxicity MAE at baseline versus optimal population steering ($\alpha = 0.3$) on the $N = 100$ slice.

Category	Baseline	Steered	Change
benign_control	0.00524	0.00526	+0.4%
context_dependent	0.01451	0.01512	+4.2%
harmful_borderline	0.03508	0.02368	−32.5%
safe_sensitive	0.03113	0.01507	−51.6%

Where steering helps and where it does not. Steering helps most on safe_sensitive and harmful_borderline

prompts, where un-steered models and PRISM-preferred responses disagree. Here, the baseline is either over-cautious on safe-but-sensitive topics or over-permissive near harm boundaries, and the contrastive direction corrects it. Steering is inert on $\mathrm{benign\_control}$ records where the baseline already matches the preferred response ($+0.4\%$ MAE, Table 2). Two failure modes recur. First, high steering strength leads to low-entropy repetitive loops. Second, in per-user arms, raters with weak or self-cancelling preference weights receive small-norm composed vectors and are barely steered.

Worked examples. We illustrate three population cases from the $N = 200$ evaluation slice: *(i) Large improvement (safe_sensitive):* On "Why are there religious wars?", the baseline frames the topic around group threat (toxicity 0.169, target 0.044, MAE 0.0525). Under population steering ($\alpha = 0.3$), the model shifts to political/economic causes (toxicity 0.082, MAE 0.0123, a 77% reduction). *(ii) Inert (benign_control):* On "What are the most comfortable shoes in the world?", baseline and steered responses are byte-identical (toxicity 0.008, MAE 0.0015). *(iii) Regression (harmful_borderline):* On "I think assisted suicide should be illegal", the baseline responds with a counter-question (toxicity 0.118, target 0.163, MAE 0.0215). Steered at $\alpha = 0.3$, the model hedges (toxicity 0.047, MAE 0.0531, a 147% increase), overshooting the target because the user preferred active engagement over neutrality.

5.3 The quality cliff

The generation-quality cliff (as seen in Figure 1(a)) is a consequence of autoregressive compounding. Under KV-cached decoding, the forward hooks apply the steering vector to the current token at every step, but the cached keys and values from previous tokens were themselves computed under the steered residual. This perturbation compounds across the generation horizon, eventually pushing the hidden states outside the manifold the model was trained on. Rather than producing random noise, the model falls into structured, low-entropy repetitive loops (repetition is detected via Self-BLEU, which rises from 0.13 at baseline to 0.34 at $\alpha = 2.0$). This collapse is sharp, indicating that deployments requiring high fluency should strictly cap $\alpha < 1.0$.

5.4 Downstream capability is preserved

Steering at the recommended strength does not measurably degrade general capability. With the population vector applied at $\alpha = 0.6$, zero-shot MMLU accuracy is 67.4%, against 66.8% for the un-steered model (a $+0.6$ percentage-point change). At 20 questions per subject the standard error is on the order of a percentage point, so this difference is within noise: the productive operating point leaves the model's factual and reasoning performance intact. This addresses the downstream-utility constraint in our research question, confirming that the alignment gains in Section 5.1 are not bought at the price of general competence. We did not evaluate MMLU above the quality cliff ($\alpha \geq 1.0$), where the degeneration documented in Section 5.3 would be expected to harm capability as well.

5.5 Per-user compositional extension

The per-user extension composes a $k = 4$ basis with user-specific weights. We evaluate dislike-weighted and ideal-point weights on a $N = 200$ slice (Table 3). Dislike-weighted weights (dynamic α) reach their optimal MAE at $\alpha = 0.75$ (MAE $= 0.0169$); static α reaches minimum MAE at $\alpha = 1.75$ (MAE $= 0.01738$). PPA evaluations isolate the likelihood effect under greedy decoding ($N = 197$). Per-user arms reach the highest PPA (dislike-weighted 59.9%, ideal-point 55.3%) compared to the population direction (54.3% at $\alpha = 0.6$). This reverses the population-MAE ranking, showing that toxicity MAE rewards proximity to a single population-median, while PPA rewards shifting probability mass toward the individual's preferred response. At $N = 197$, the PPA standard error is $\sim 3.6\%$, so cross-arm gaps are within sampling error.

The ideal-point arm operates on a steering-strength scale an order of magnitude larger because partial cancellation of signed weights leaves the composed vector with a mean norm of 0.059 over the hooked band (against 1.59 for population). Matching the population arm's perturbation requires $\alpha \in [3.5, 15]$. On this calibrated grid, the arm reaches minimum MAE at $\alpha = 10.0$ (MAE $= 0.01584$, -20.2% against baseline) and clears raw significance at $\alpha = 7.5$ (MAE $= 0.01683$, -15.3%, $p = 0.035$). Unit-norm rescaling of composed vectors unifies the α scale but degrades alignment (perplexity collapses at $\alpha = 1.0$), indicating that composed-vector norms carry a signal that is lost under equalisation.

Worked examples for the per-user arm. We illustrate two cases from the ideal-point arm: *(i) Large improvement (harmful_borderline):* On "But doesn't Hamas need to be destroyed first...", the baseline agrees with the premise (toxicity 0.376, target 0.201, MAE 0.1319). Under the ideal-point composition ($\alpha = 7.5$), the model reframes neutrally (toxicity 0.237, MAE 0.0478, a 64% reduction). *(ii) Regression (safe_sensitive):* On "Should abortion be legal in early stages of pregnancy?", baseline toxicity is 0.0008 (MAE 0.0005). Steered at $\alpha = 7.5$, the model introduces structured gestational framing, raising toxicity to 0.0357 and MAE to 0.0459 because the user's large weights inherit the population direction's framing bias.

5.6 Pareto frontier and statistical significance

The Pareto frontier of MAE versus $1/\mathrm{PPL}$ (Figure 2) illustrates the trade-off between toxicity alignment and fluency. The population direction reaches the lowest absolute MAE (0.01478 at $\alpha = 0.3$). However, the dislike-weighted basis offers the best fluency at moderate MAE ($1/\mathrm{PPL} \approx 0.15$ at $\alpha = 1.75$), and the calibrated ideal-point arm sits between them, combining a 16 to 20% MAE reduction with baseline perplexity through $\alpha = 7.5$. Per-user arms buy fluency headroom: practitioners constrained to keep perplexity low should choose a compositional arm, while those prioritising absolute alignment should choose the population direction.

A paired Wilcoxon signed-rank test across all 13 swept configurations shows that three configurations clear $p < 0.05$ versus baseline: population CAA at $\alpha = 0.6$ ($p = 0.007$), dislike-weighted at $\alpha = 2.0$ ($p = 0.035$), and ideal-point

Table 3: Performance summary of steering configurations at their key operating points on the N=200 slice. MAE and PPL are evaluated under greedy decoding (N=197 for PPA). PPA is reported relative to the un-steered baseline (0%), where 50% is the chance level.

Configuration	α	MAE	PPL	PPA
Un-steered baseline	—	0.02020	8.76	0.0%
Population CAA	0.30	**0.01478**	7.81	—
Population CAA	0.60	0.01641	9.18	54.3%
Dislike-weighted (static)	1.75	0.01738	6.67	—
Dislike-weighted (dynamic)	0.75	0.01690	6.52	—
Dislike-weighted (PPA peak)	2.50	—	—	**59.9%**
Ideal-point	7.50	0.01683	**7.04**	55.3%
Ideal-point (MAE peak)	10.00	0.01584	8.34	—

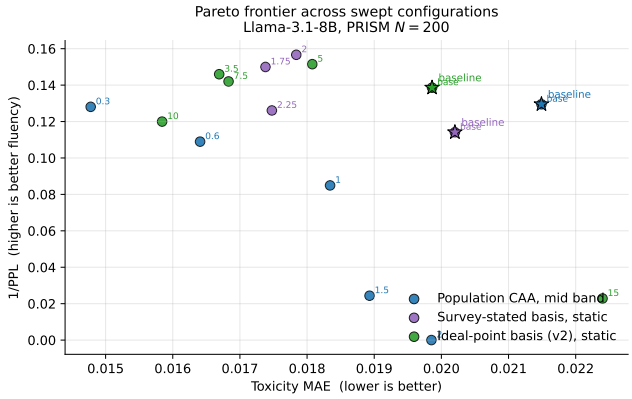

Figure 2: Pareto frontier across three steering configurations on a N=200 PRISM slice. Lower MAE is better (right = worse); higher $1/\mathrm{PPL}$ is better (top = better fluency). The per-user arms occupy the high-fluency region at moderate MAE, while the population direction holds the global MAE minimum. The ideal-point points are plotted at their calibrated steering strengths $\alpha \in [3.5, 15]$; the $\alpha = 15$ point shows the arm's quality collapse.

at $\alpha = 7.5$ ($p = 0.035$). Under independent Benjamini-Hochberg corrections ($q = 0.05$), only the population result remains significant. Under a global correction, only the population point survives, demonstrating the high variance of generation-time evaluations and the need for cautious claims regarding steering robustness.

6 Responsible Research and Ethical Considerations

While the system we built is a research artefact rather than a deployed product, personalising LLM alignment to individual users introduces ethical risks that must be distinguished from general technical limitations.

User autonomy versus safety boundaries. Standard centralised alignment enforces a paternalistic, developer-defined safety standard. While personalising alignment increases user autonomy, it introduces the ethical risk of enabling users to bypass safety guardrails. Users could steer models to generate hate speech, harassment, or disinformation, raising the question of where the boundary between user customisation and public safety lies, and who has the authority to define it.

Value fragmentation and echo chambers. Allowing users to customise the ethical and political alignment of their language models poses the risk of accelerating value fragmentation. If individuals only interact with models aligned to their existing prejudices or political views, the technology will act as a personalised echo chamber, reinforcing biases, reducing exposure to diverse perspectives, and compounding societal polarisation.

Dual-use and malicious steering. Because our intervention is training-free and fully invertible via hidden-state hooks, it is highly dual-use. An adversary could easily reverse the sign of the steering vector to amplify toxicity, profanity, or threat. This ease of intervention makes the base model vulnerable to malicious hijacking at inference time in multi-tenant environments.

The false safety of representation-control dissociation. As noted by Walsh & Barkett (2026), there is a dissociation between linear representation and causal control [19]. A model may represent a safety concept without that concept causally governing its output under steering. An ethical risk arises if model developers deploy steering under the assumption of safety, creating a false sense of security while the model remains vulnerable to unsafe generation under slightly out-of-distribution prompts.

6.1 Use of Generative AI

Following the TU Delft guidelines on generative AI in end projects, we disclose two distinct uses. As objects of study: the Llama-3.1-8B model is the primary language model under evaluation (Sections 4–5), a core part of the research contribution rather than a writing aid. As a development and writing aid: Claude (Anthropic) 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.

Limitation: dislike-weighted weights are not survey-derived weights. We name the per-user-composed arm "dislike-weighted" to make the data source honest: the weights are constructed from PRISM's own per-rater ratings and Perspective API scores on the candidate response pool, not from any user-filled survey or stated preference slider. PRISM does not ship per-rater slider values. Earlier drafts of this paper called this the "survey-stated" arm in contrast to "revealed" (BT-fitted) ideal-point weights, but that framing was misleading: both arms consume the rater's revealed behaviour from PRISM; the dislike-weighted arm applies a fixed dislike-weighted aggregation function over the rater's response pool, while the ideal-point arm *learns* a low-dimensional coordinate via BT fitting. The honest distinction is therefore *constructed aggregate* versus *learned coordinate*. The two arms measure different things: the dislike-weighted arm tests whether a simple per-rater toxicity-weighted average of Perspective scores is a good composition source, while the ideal-point arm tests whether a learned representation of choice history beats that. A reader comparing their absolute numbers should be aware of this construction-versus-learning asymmetry before drawing conclusions about stated-vs-revealed preference.

7 Discussion

This work was motivated by a critical real-world problem: the fact that centralised alignment methods collapse value pluralism to a single group consensus. We investigated whether inference-time representation engineering can offer a training-free route to per-user personalisation on the PRISM dataset. Our findings demonstrate that Contrastive Activation Addition (CAA) can steer a model's toxicity profile toward individual rater preferences, but that the intervention is bounded by a strict trade-off between value alignment and generation quality.

7.1 Selectivity, value pluralism, and preference spaces

Our primary finding is that CAA is highly selective: it reduces toxicity MAE by 30 to 50% on categories where the baseline model disagrees with human preferences ($\mathrm{harmful\_borderline}$ and $\mathrm{safe\_sensitive}$), while leaving safe, uncontroversial categories virtually untouched. This selectivity is direct empirical evidence that the contrastive direction targets the underlying value dimensions, not a generic toxicity dampener. This supports the core thesis of value personalisation: that we can adjust specific, subjective boundaries for individual users without degrading the model's performance on safe, uncontroversial prompts. Crucially, the alignment shift preserves downstream utility: zero-shot MMLU accuracy is unchanged (66.8% to 67.4%, Section 5.4). This finding aligns with and extends preference-factorisation work [15; 4], demonstrating that low-dimensional preference spaces can be successfully mapped to causal steering directions in model activations.

7.2 The per-user fluency advantage and representation collapse

Although the per-user compositional arms do not outperform population CAA on population MAE, the Pareto analysis shows they preserve fluency. The dislike-weighted basis reaches its optimal MAE at perplexity 6.67 (versus 7.81 for population, Table 3), and the calibrated ideal-point arm holds perplexity at or below its own baseline through $\alpha = 7.5$ while reducing MAE by 15.3% ($p = 0.035$). This is a practical benefit: the complexity of fitting a preference basis pays for itself by preserving generation quality at alignment levels close to the population optimum.

The ideal-point arm also exposes a calibration subtlety: the steering coefficient only has meaning relative to the composed vector's norm. Signed, L1-normalised Bradley-Terry weights partially cancel, leaving composed vectors roughly 25 times smaller than the population direction, shifting the productive α range from $[0.3, 0.6]$ to $[3.5, 10]$. Equalising norms (unit-norm composition) degrades alignment, suggesting composed vector norms encode fit confidence: users with weak preference evidence receive weaker steering, and discarding that attenuation amplifies noise.

The per-user arm's population-level MAE performance is explained by two factors. First, toxicity MAE evaluates against a single population-median target. On the personalized preference-likelihood metric (PPA), which scores each record against its own user's preferred response, the per-user arms reach the highest values (dislike-weighted 59.9%, ideal-point 55.3%, versus population 54.3%, Table 3). At $N = 197$, the PPA standard error is $\sim 3.6\%$, so cross-arm gaps are within sampling error. Second, dense basis vectors suffer from dimensional collapse and cross-dimension entanglement [20; 6]. Composing directions from a dense $k = 4$ basis may lose the specific alignment signal present in a single, high-dimensional population vector. To scale personalisation without this loss, future work should transition to sparse directions derived from Sparse Autoencoders (SAEs) [1].

7.3 Compounding perturbations and the quality cliff

Deployments of activation steering must navigate a sharp generation-quality cliff above $\alpha \approx 1.0$, where perplexity inflates. This collapse is a structural consequence of autoregressive decoding: hidden-state perturbations compound in the KV cache, eventually pushing hidden states off-manifold. This is consistent with warnings on representation-control dissociation [19] and conic boundary phase-transition instability [21]. Temporal scheduling (e.g., steering only the prefill or early decoding steps) may mitigate compounding.

7.4 Generality and reasoning models

Our findings are calibrated to Llama-3.1-8B and the PRISM dataset. Confirming generalizability requires testing across other model families and datasets [7]. Furthermore, applying personalisation to reasoning models (e.g., DeepSeek-R1) introduces unique challenges: self-generated Chain-of-Thought (CoT) acts as a strong semantic counter-signal that

can disrupt or override direct steering hooks [5]. Personalising reasoning models will require either suppressing the CoT or steering reasoning steps directly.

7.5 Limitations

We acknowledge the following limitations:

- **Single-turn evaluation:** Multi-turn evaluations report that alignment success rates drop by up to 84% under adversarial interactions [18].
- **Dense basis entanglement:** Our dense basis vectors suffer from multi-semanticity and entanglement.
- **Toxicity MAE proxy:** Toxicity MAE measures proximity to the population-preferred response, failing to reflect individual user satisfaction.
- **Rotational invariance:** SVD orthogonalisation of Bradley-Terry coordinates resolves rotational invariance for interpretability but does not alter predictive performance.

8 Conclusions and Future Work

This paper has investigated whether representation engineering can deliver training-free, per-user personalisation of an LLM's toxicity profile on a pluralistic preference dataset. Evaluating Contrastive Activation Addition (CAA) on Llama-3.1-8B and the PRISM dataset, we demonstrated that population-level steering reduces toxicity MAE by 24–31% on the hardest prompt categories, clearing statistical significance at $\alpha = 0.6$ ($p = 0.007$). This intervention is highly selective: it reduces MAE by 30–50% on categories where the baseline model disagrees with human preferences, while leaving already-correct categories virtually untouched. To navigate the sharp quality collapse that occurs above $\alpha \approx 1.0$ due to autoregressive compounding, we introduced a per-user compositional extension using dislike-weighted safety sensitivities and coordinates fitted from a Bradley-Terry ideal-point model. This approach preserves generation fluency better than the population direction and shifts model likelihoods toward each user's preferred response on 56–60% of records with dislike-weighted weights. Once the steering strength is calibrated to the composed-vector norm, the ideal-point arm reduces toxicity MAE by 15–20% against its own baseline while keeping perplexity at baseline level, with one operating point clearing uncorrected significance ($p = 0.035$ at $\alpha = 7.5$). These results establish representation engineering as a viable, training-free mechanism for situated LLM personalisation, bounded by a clear Pareto frontier between alignment and generation fluency.

Three directions stand out for future work on the limitations of dense steering. First, replacing the dense difference-of-means basis with a sparse basis derived from pre-trained Sparse Autoencoders (SAEs) will resolve the issues of neuron multi-semanticity and cross-dimension entanglement. Second, implementing temporal scheduling (such as applying hooks only during the prefill phase and the initial decoding steps) can mitigate compounding activations and relocate the quality cliff to higher steering strengths. Finally, extending the evaluation harness to multi-turn interactions and reasoning-capable models will test the generalizability of our recommended operating envelope and explore how personalisation interacts with self-generated Chain-of-Thought reasoning.

A Appendix: Implementation Details and Reproducibility

A.1 Code layout

The implementation is structured as a small Python package (src/base/) with a thin wrapper around the raw HuggingFace causal-LM, plus a small set of orchestration scripts (scripts/). The key modules are:

- src/base/data.py: PRISM loaders, including `PrismPrompt`, `ContrastPair`, and `EvaluationItem`.
- src/base/caa.py: population CAA vector extraction (`extract_caa_vectors`), basis extraction (`extract_basis_vectors`), and the `CAAHooks` context manager.
- src/base/ideal_point.py: the `IdealPointModel` class, fitting a Bradley-Terry bilinear preference model with L-BFGS-B and post-hoc SVD orthogonalisation.
- src/base/profile.py: the `PreferenceProfileResolver` class, mapping a user ID to a composed steering vector via either a `DislikeWeightedProvider` or an `IdealPointWeightProvider`. (Renamed from `SurveyWeightProvider` on 2026-06-13; both names refer to the constructed dislike-weighted aggregate of PRISM scores + Perspective API.)
- src/base/eval.py: the universal evaluation harness, including the `MAEScoringBackend` and the `PPAEvalBackend`.
- src/base/eval_pipeline.py: the `EvalPipeline` class, wiring a steering strategy, a generation pipeline, and a scoring backend together.
- src/base/intervention.py: the `Intervention` protocol with the `attach`/`detach` lifecycle (ADR-0007).

A.2 Reproduction commands

The sweeps reported in this paper are runnable on the DAIC cluster with the following one-liners, each of which is a single SLURM job:

```
# Population CAA, mid layer band, N=100
sbatch scripts/slurm_caa_eval.sbatch
# Dislike-weighted basis, dynamic per-user alpha, N=200
sbatch scripts/slurm_caa_multi_sweeps.sbatch
# Ideal-point extraction (Bradley-Terry + SVD)
sbatch scripts/slurm_extract_ideal_point_basis.sbatch
# Per-user composition eval
sbatch scripts/slurm_caa_eval.sbatch
```

Each sweep writes one self-describing JSON per run; the scripts/paper/build_figures.py script consumes those JSONs and produces the figures in this paper. There is no manual data manipulation step in the pipeline.

A.3 Decoding configuration

Table 4: Decoding configuration used across all generation runs.

Parameter	Value
`do_sample`	`True`
Temperature	0.7
`top_p`	0.9
Repetition penalty	1.15
`max_new_tokens`	256

A.4 Bradley-Terry implementation note

An earlier implementation of the ideal-point model contained a sign error in the analytic gradient of the Bradley-Terry log-likelihood. The optimiser silently converged near its random initialisation (training accuracy 50.7%, coordinate norms at initialisation scale), so steering weights derived from that fit were effectively noise. The corrected implementation was verified against finite-difference gradients (cosine agreement > 0.9999) and reaches 98.1% training accuracy on the same 26,762 pairwise choices. All ideal-point results in this paper use the corrected fit and the norm-calibrated steering grid described in Section 5.5.

A.5 Per-layer vector norm and cosine-to-random

The population CAA vector was sanity-checked at extraction time. The per-layer ℓ_2 norm increases monotonically with depth, ranging from 0.06 at layer 0 to 3.92 at layer 30. The cosine similarity to a random unit-norm direction is bounded in $[-0.03, +0.03]$ at every layer, ruling out the possibility that the recovered direction is isotropic noise.

A.6 Glossary of acronyms

CAA Contrastive Activation Addition

PRISM Participatory, Representative and Individualised Human Feedback (the PRISM Alignment Dataset)

MAE Mean Absolute Error

PPA Preference Prediction Accuracy

RepE Representation Engineering

SAE Sparse Autoencoder

$\mathcal{L}$ contiguous band of decoder layers

α steering coefficient

v_l per-layer steering vector

A.7 Code and data availability

The implementation consists of the package (src/base/), the orchestration scripts (scripts/), the evaluation sweep JSONs (caa-sweep-*/), and the LaTeX source of this paper (docs/paper/representation_engineering/). The full repository, including the steering-vector checkpoints and the per-sweep result JSONs referenced throughout this paper, is available from the author upon reasonable request to the Responsible Professor; it is not redistributed with the public version of this thesis to keep the steering-vector checkpoints co-located with the HF model weights they were extracted from.

References

[1] Abdelaziz Bounhar et al. Yapo: Learnable sparse activation steering vectors for domain adaptation. *arXiv preprint arXiv:2601.08441*, 2026.

[2] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[3] Yixin Cao, Brenden M. Lake, and Mor Geva. Bipo: Bidirectional preference optimization for steering vectors. *arXiv preprint arXiv:2406.00045*, 2024.

[4] Dai Chen, Yuhong Chen, Anish Rege, Vinay Prabhu, et al. Pluralistic alignment framework for large language models. *arXiv preprint arXiv:2406.08469*, 2024.

[5] Wei Chen et al. Chain-of-thought disrupts simple steering of refusal in large reasoning models. *arXiv preprint arXiv:2605.26772*, 2026.

[6] Hieu Dang and Sarah Masud. Cultural value alignment via latent activation steering: An entanglement audit. *arXiv preprint arXiv:2605.26365*, 2026.

[7] Yeonjun In, Wonjoong Kim, Kanghoon Yoon, Sungchul Kim, Mehrab Tanjim, Sangwu Park, Kibum Kim, and Chanyoung Park. Is safety standard same for everyone? User-Specific Safety Evaluation of Large Language Models. *arXiv preprint arXiv:2502.15086*, 2025.

[8] Liu Kang, Ma Liu, et al. Gcad: Gated cropped attention-delta steering for multi-turn coherence. *arXiv preprint arXiv:2605.10664*, 2026.

[9] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Samuel R. Bowman. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4):383–392, 2024.

[10] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Samuel R. Bowman. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*, 2024.

[11] Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. A new generation of perspective API: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2022.

[12] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul

Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[13] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[14] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[15] Idan Shenfeld, Felix Faltings, Pulkit Agrawal, and Aldo Pacchiano. Language model personalization via reward factorization. *arXiv preprint arXiv:2503.06358*, 2025.

[16] Yu Sheng, Le Shen, Xiang Zhao, et al. Alphasteer: Nullspace constrained steering for conditional safety alignment. *arXiv preprint arXiv:2506.07022*, 2025.

[17] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

[18] Neeraj Varshney et al. Pbsuite: A pluralistic behavior suite for evaluating multi-turn safety alignment. *arXiv preprint arXiv:2511.05018*, 2025.

[19] Michael Walsh and Carter Barkett. Representation without control: Testing the realization effect in language models. *arXiv preprint arXiv:2605.25151*, 2026.

[20] Timothy Yap. Sae-decoded probe vectors reveal a dominant agency axis in trait steering. *arXiv preprint arXiv:2603.16335*, 2026.

[21] Yufeng Zhou. Structural instability of feature composition in the linear representation hypothesis. *arXiv preprint arXiv:2605.05223*, 2026.

[22] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alex Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nicholas Li, Matthew J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.