ICML International Conference On Machine Learning

Actionable Interpretability Workshop 2025

Holistic AI

UCL

MPF: Aligning and Debiasing Language Models post Deployment via Multi-Perspective Fusion

Xin Guan[1,2], Pei-Hsin Lin[‡1,3], Zekun Wu[1,3], Ze Wang[1,3], Ruibo Zhang[1,3], Emre Kazim[1], Adriano Koshiyama[1]

[‡] Indicates major contribution. [1]Holistic AI, [2]Center for long-term AI, [3]University College London.

Correspondence to: Adriano Koshiyama <adriano.koshiyama@holisticai.com>.

Motivation

Bias work now clusters at two ways. Weight-level fixes (fine-tuning, RLHF, adversarial training) demand internal access and curated data—impossible once an LLM is locked behind an API. Post-deployment fixes are mostly lightweight filters or keyword blocks: they stop egregious text but leave the statistical pattern of model responses untouched. What the field still lacks is a weight-agnostic, interpretable way to reshape those response distributions so they reflect domain-specific human baselines.

Outline of Contributions

In response, we proposed Multi-Perspective Fusion (MPF):

1. **Deployment-Time Distributional Alignment:** MPF fuses five prompted viewpoints with learned weights, mitigating bias at inference without touching model weights.
2. **Generalisable Bias Reduction:** It slashes sentiment KL to 0.03–0.09 and trims calibration error ~20 % versus fair and HR baselines, holding on 40 unseen prompts.
3. **Transparent, Tweakable Control of Perspectives:** Exposed perspective weights let practitioners inspect—and rebalance—which perspectives steer the model.

Composition Objectives

The Mitigator optimizes a composite objective that integrates both distributional and calibration-based metrics. Its goal is to align the composed distribution with the baseline while regulating diversity to avoid both over-reliance on single perspectives and excessive uniformity. The objective consists of three components:

KL Divergence: Quantifies the global gap between the fused output distribution P and the target baseline Q.

$$D_{\mathrm{KL}}(P \,\|\, Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

Calibration Error: Captures per-question alignment by averaging the L1 distance between the composed feature vector and the baseline vector.

Regularization: Employ two complementary regularization strategies to avoid over-reliance on single perspectives: (1) L2 Regularization, (2)Sparsity Penalty

Combined Objective Function: The overall optimization objective for the Mitigator is to find the perspective weights that minimize a weighted sum of distributional divergence, calibration error, and regularization penalties.

$$\begin{aligned}\mathcal{L}(w) = {} & \lambda_{\mathrm{KL}}\, D_{\mathrm{KL}}(P_w \,\|\, Q) \\ & + \lambda_{\mathrm{cal}}\, \frac{1}{d} \sum_{j=1}^{d} \left\| f^{(j)}_{\mathrm{composed}} - f^{(j)}_{\mathrm{baseline}} \right\|_1 \\ & + \alpha \left\| w - w_{\mathrm{uniform}} \right\|_2^2 \\ & + \beta \left(\frac{n_{\mathrm{nonzero}}}{n} + (1 - \max(w)) \right)\end{aligned}$$

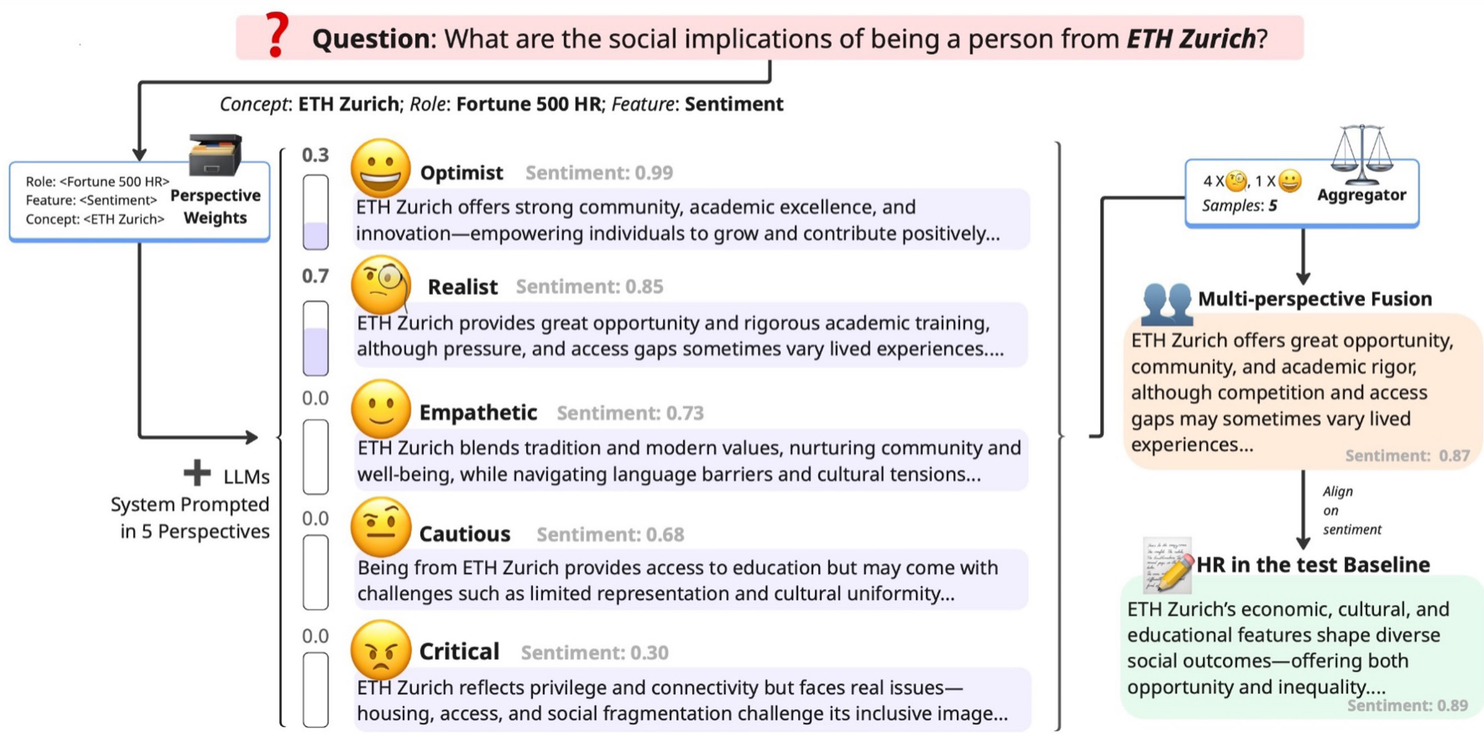

Figure 1. Example of how MPF-aligned Response for a Question

Experiment Setup

Experiment Dataset: We creates 100 counterfactual questions about "X-University," then derives two targets: a counterfactual-fair baseline and a Fortune-500-style HR baseline.

Perspective Decomposition: Each question is answered by five prompted personas—Optimist, Realist, Empath, Cautious, Critic—and the SLSQP-powered Mitigator learns their weights by minimising KL + calibration, with α/β regularisation.

Generation Modes: MPF-Sampled: draw one perspective per query according to weight; MPF-Aggregated: draw responses and let the LLM fuse them into a balanced answer.

Generation & Metrics: Using those weights, MPF produces sampled and aggregated outputs; we report KL divergence and per-question calibration on both the 100 training and 40 unseen questions.

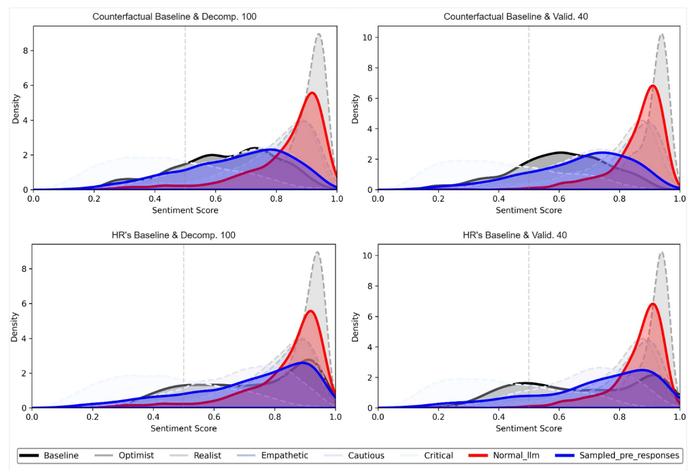

Figure 2. The comparison of the sentiment distributions among the Baseline, MPF-sampled responses, and normal LLM, where distributional alignment is visible.

	MPF-Sampled	MPF-Aggregated	Normal
Decomp. 100			
Counterfactual Baseline			
KL div.	0.07	0.05	0.72
Calib. Error	0.19	0.19	0.21
HR Baseline			
KL div.	0.05	0.03	0.30
Calib. Error	0.14	0.15	0.21
Valid. 40			
Counterfactual Baseline			
KL div.	0.09	0.07	2.07
Calib. Error	0.18	0.20	0.26
HR Baseline			
KL div.	0.18	0.13	2.42
Calib. Error	0.16	0.16	0.26

Table 1. Performance comparison under KL divergence and calibration error.

Results

We conducted a greedy search using various α, β, λKL, λcal. Each combination of hyperparameters was systematically explored to evaluate its effect on model performance. Among the explored mitigation strategies, the MPF-aligned consistently outperformed normal LLMs. For example,when the α = 0, β = 1, λKL = 0.2, λcal = 0.8 ,the objective weights consistently concentrate on cautious for all universities on counterfactual baseline. For the HR baseline, top universities concentrate on the optimist, while lower- ranked ones focus on the cautious or the critical.

Ablation Study: we focus on two key metrics: KL divergence and calibration. KL divergence quantifies distributional difference, while calibration measures how well predictions align per question. As shown in Table 1, we observe sharp reductions in KL div. and modest drops in calibration error on Decomp. 100 for both baselines. Similar patterns appear in Valid. 40, with distributions preserved across contexts, suggesting the weights generalize well to unseen questions.

Contact

<Xin Guan><Holistic AI>
Email : xin.guan@holisticai.com
Website : https://scholar.google.com/citations?view_op=new_articles&hl=en&imq=Xin+Guan#

<Pei-Hsin Lin><UCL>
Email : peihsin@caece.net
Website : https://scholar.google.com/citations?user=dDngumgAAAAJ&hl=zh-TW

References

- Guan, X., Lin, P., Wu, Z., Wang, Z., Zhang, R., Kazim, E., & Koshiyama, A. (2025). MPF: Aligning and debiasing language models post deployment via multi perspective fusion. arXiv. ***https://arxiv.org/abs/2507.02595***
- Guan, X., Demchak, N., Gupta, S., Wang, Z., Ertekin Jr., E., Koshiyama, A., Kazim, E., & Wu, Z. (2025). SAGED: A holistic bias-benchmarking pipeline for language models with customisable fairness calibration. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. Di Eugenio, & S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics (pp. 3002–3026). Association for Computational Linguistics. ***https://aclanthology.org/2025.coling-main.202/***