

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 Kazim¹ Adriano Koshiyama¹

Abstract

Multi-Perspective Fusion (MPF) is a novel post-training alignment framework for large language models (LLMs) developed in response to the growing need for easy bias mitigation. Built on top of the SAGED pipeline—an automated system for constructing bias benchmarks and extracting interpretable baseline distributions—MPF leverages multi-perspective generations to expose and align biases in LLM outputs with nuanced, human-like baselines. By decomposing baseline—such as sentiment distributions from HR professionals—into interpretable perspective components, MPF guides generation through sampling and balancing of responses, weighted by the probabilities obtained in the decomposition. Empirically, we demonstrate its ability to align LLM sentiment distributions with both counterfactual baselines (absolute equality) and the HR baseline (biased for Top Uni.), resulting in small KL divergence, reduction of calibration error and generalization to unseen questions. This shows that MPF offers a scalable and interpretable method for alignment and bias mitigation, compatible with deployed LLMs and requiring no extensive prompt engineering or fine-tuning.

1. Introduction

Recent advancements in large language models (LLMs) have highlighted both their capabilities for bias and their harmful effect, raising significant concerns regarding alignment and fairness in deployed systems (Broussard, 2024; Gebru, 2020). In this paper, we introduce Multi-Perspective Fusion (MPF), a novel post-training alignment method that builds upon the bias interpretation capabilities of the

[‡] Indicates major contribution. ¹Holistic AI ²Center for long-term AI ³University College London. Correspondence to: Adriano Koshiyama <adriano.koshiyama@holisticai.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

SAGED (Guan et al., 2025) pipeline. MPF offers *distributional alignment* avoiding heavy prompt crafting or model fine-tuning—while remaining compatible with both.

For assessment of bias (Gallegos et al., 2024), benchmarking frameworks such as BOLD (Dhamala et al., 2021) and SAGED have emerged as post-deployment methods for uncovering LLM’s distributional biases around specific concepts—such as gender (e.g., “female”)—along particular linguistic features like sentiment, personality, or topical focus. In particular, SAGED’s automated construction of Question-Baseline (QB) benchmarks from dedicated texts allow systematic comparison between LLM outputs and implied baselines in the texts, thereby enable mitigating bias through aligning with the baselines’ feature distribution. To realize this, MPF re-composes the baseline’s features into a weighted mixture of distribution from interpretable perspectives. Then, MPF uses these weights to probabilistically simulate LLM responses through sampling, and leading to responses aligning with the baseline on particular features (i.e. sentiment), as shown in Fig 2.

For our experiment, we instantiate MPF to align LLM two baselines - a counterfactual one representing complete fairness and a hypothetical HR, which can be useful in revealing and mitigating potential biases in LLM-based resume screening (Wang et al., 2024). To decompose this baseline, we define five sentiment-related perspectives—optimistic, realist, empathetic, cautious, and critical—and generate responses from these perspectives within the same benchmark. We then recompose the baseline distributions across universities into a weighted combination of the sentiment distributions of these perspectives.

The outcome of experiment demonstrates that the same set up can be used to exploit bias i.e. reproducing the bias of an HR, or reduce bias i.e. making LLM’s response close to a counterfactual baseline in sentiment. Through ablation studies with normal LLM, we find that applying MPF significantly reduces the sentiment discrepancy, especially in the distribution sense, between baselines and MPF-aligned LLM outputs. These results unfold MPF’s practical effectiveness in aligning language model outputs more closely with nuanced human sentiment.



Figure 1. Example of how MPF-aligned Response for a Question, when the perspective weights already obtained through MPF mitigator. Here, only optimist and realist have weights hence generated. We show responses from other three perspectives only for illustrations.

2. Related Works

Mitigating Bias with Weight Updates. Bias mitigation in LLMs occurs at training, fine-tuning, and deployment stages. Training methods tackle bias via balanced data (Dodge et al., 2021), counterfactual augmentation (Zhao et al., 2018), and adversarial techniques (Elazar & Goldberg, 2018). Fine-tuning enables post hoc alignment using RLHF (Ouyang et al., 2022), adapters (Lauscher et al., 2021). Recent methods emphasize interpretability and automation, such as ReGiFT (Kabra et al., 2025), and RLDF (Cheng et al., 2024). However, these require access to model weights and curating training data, which can limit usability and scalability.

Deployment-Time Bias Mitigation. In contrast, Multi-Perspective Fusion (MPF) offers a model-agnostic, zero-weight-update approach after deployment. Earlier after-deployment mitigation techniques—output filtering (Gehman et al., 2020), rewriting (Zhao et al., 2021), and controlled decoding (He et al., 2022)—aim to block harmful content. More recent tools like ConceptX (Amara et al., 2025) support interpretable editing, but focus largely on harmful content mitigation. MPF instead aligns outputs with evaluative baselines using SAGED (Guan et al., 2025), offering both interpretability and constructive preference alignment around specific concepts.

Comparison with Prompt-Based Approaches. Architecturally, MPF relates to Chain-of-Thought (Wei et al., 2022; Kojima et al., 2022), Self-Consistency (Wang et al., 2022), and Tree-of-Thought (Yao et al., 2023) methods,

which aggregate multiple generations to refine outputs. Yet unlike truth-evaluative approaches like debate prompting (Madaan et al., 2023; Bai et al., 2022; Khan et al., 2024), MPF aligns generations to human-like distributional baselines—eschewing truth judgments for balanced, preference-driven fusion. MPF thus uniquely combines model-agnostic deployment, zero-weight-update feasibility, and distributional preference alignment—bridging the gap between interpretability and actionable bias mitigation.

3. Methodology

Our Multi-Perspective Fusion (MPF) framework introduces a novel distributional alignment approach through a two-stage architecture: the Mitigator and the ResponseGenerator. The Mitigator analyzes and optimizes perspective weights to match baseline distributions, while the ResponseGenerator leverages these weights to generate debiased responses through probabilistic sampling and aggregation.

3.1. 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:

Distributional Metrics. To quantify divergence between the ensemble and the target distribution, we primarily adopt **KL Divergence** in our main experiments. KL Divergence

provides a sensitive and asymmetric measure of relative entropy, effectively penalizing deviations in high-probability regions. It is defined as $D_{\text{KL}}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$, where P is the composed distribution and Q is the target baseline.

Calibration-Based Metrics: While distributional metrics compare global output patterns, calibration-based metrics evaluate question-specific deviations. For each query, we compute the composed feature vector as a weighted sum of each perspective’s feature score vector: $f_{\text{composed}} = \sum_{i=1}^n w_i f_i$, where f_i is the feature score vector from perspective i . The calibration error is then defined as the mean L_1 norm of the difference between the composed vector and the baseline vector f_{baseline} : Calibration Error = $\frac{1}{d} \|f_{\text{composed}} - f_{\text{baseline}}\|_1$ where d is the number of questions in composition.

Regularization: To regulate diversity and avoid both over-reliance on single perspectives and excessive uniformity, we employ two complementary regularization strategies: (1) *L2 Regularization*: This term discourages the weights placing too much emphasis on a single perspective. Formally, it is expressed as $\alpha \|w - w_{\text{uniform}}\|_2^2$, where $w_{\text{uniform}} = \frac{1}{n} \mathbf{1}$ and α controls the strength of this regularization. (2) *Sparsity Penalty*: This component penalizes deviations of the weight vector w from the uniform distribution, thus preventing excessive uniformity. It combines a count penalty $\frac{n_{\text{nonzero}}}{n}$, which encourages concentration of weights to a few perspectives, and a maximum weight penalty $(1 - \max(w))$, which encourages the dominance of a single perspective. The combined term is weighted by β .

Combined Objective Function. The overall optimization objective for the Mitigator is to find the perspective weights w that minimize a weighted sum of distributional divergence, calibration error, and regularization penalties. Where λ_{KL} and λ_{cal} are the relative strength of the KL and calibration respectively, the combined objective function is:

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

3.2. Optimization Procedure

To minimize the composite objective function $\mathcal{L}(w)$ defined above, we employ a constrained optimization strategy using the Sequential Least Squares Quadratic Programming (SLSQP) algorithm. At the start of each optimization attempt, the initial weights are randomly sampled from a Dirichlet distribution to mitigate the risk of local minima.

The SLSQP then runs iteratively subjecting to the simplex constraint $\sum_i w_i = 1$ and bounds $0 \leq w_i \leq 1$, until either the maximum number of iterations (default: 1000) is reached, or the change in the objective function between iterations falls below a convergence tolerance of 10^{-6} .

3.3. Using Weights in Generation

The MPF’s ResponseGenerator supports two steps to obtain MPF-aligned generations: (1) Sampled Generation, which selects a single perspective (e.g., optimistic, realist, empathetic, cautious, or critical) based on optimized probability weights and generates a response using that perspective’s system prompt. This probabilistic sampling aims to reproduce the baseline feature distribution. (2) Aggregated Generation, which produces multiple sampled generation responses and combine them to a LLM prompted to combine several samples into a balanced response faithfully to mitigate extreme answers from small probability perspectives.

4. Experiments

We design two primary experiments to validate the alignment performance of MPF against counterfactual and hypothetical baselines. For reproducibility, all experiments decompose 100 seed questions to derive perspective weights and evaluate generalization on a held-out set of 40 questions. We ablate MPF (with Qwen-turbo-latest) by comparing the results to each of the perspectives and no prompt LLM.

4.1. Experimental Setup

Question-Baseline Preparation. To construct the benchmark, an article was generated using ChatGPT-4o (Appendix A.1), focused on a hypothetical institution named “X-University.” Subsequently, SAGED’s scraping and question generation methods produced questions baseline. Counterfactual questions were then created by systematically replacing “X-University” with names of 30 randomly chosen universities (Appendix A.2) with different QS rankings. The generated questions were used as prompts to elicit responses from multiple perspectives, including optimistic, realistic, empathetic, cautious, and critical perspectives (Appendix B.1, Appendix B.2). Two types of baselines were established: (1) *a counterfactual baseline* using the sentences scraped from the article, and (2) *a hypothetical baseline* constructed by simulating HR-generated responses.

Procedure. Our experimental workflow consists of three main steps: (1) apply MPF Mitigator to obtain the optimal weight breakdown of 100 questions into perspective distributions; (2) generate MPF-aligned outputs and normal LLM on 100 questions used in breakdown (Decomp. 100) + 40 held-out counterfactual questions (Valid. 40); and (3) compare these outputs against individual component per-

spectives and evaluate the effectiveness using KL and the calibration error metrics in Section 3.

4.2. Ablation Results

We conducted a greedy search using various $\alpha, \beta, \lambda_{KL}, \lambda_{cal}$. 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 $\alpha = 0, \beta = 1, \lambda_{KL} = 0.2, \lambda_{cal} = 0.8$, 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.

In this 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. MPF-Sampled was optimized with $\alpha = 0, \beta = 1, \lambda_{KL} = 0.2, \lambda_{cal} = 0.8$, and one sample. MPF-Aggregated used $\alpha = 0.5, \beta = 0.5$, the same weights, and aggregated over three samples. Low KL values (≤ 0.2) mean MPF-sampled mimics both baselines’ distributions, as in Figure 2. Calibration error shows MPF-aligned responses still deviate from baseline by 15–20% per question, likely due to inherent fluctuation in LLM responses. For the HR baseline, MPF-aligned responses also aligns well with individual universities’ QS rankings. See more in Appendix C.2.1 and Appendix C.2.2.

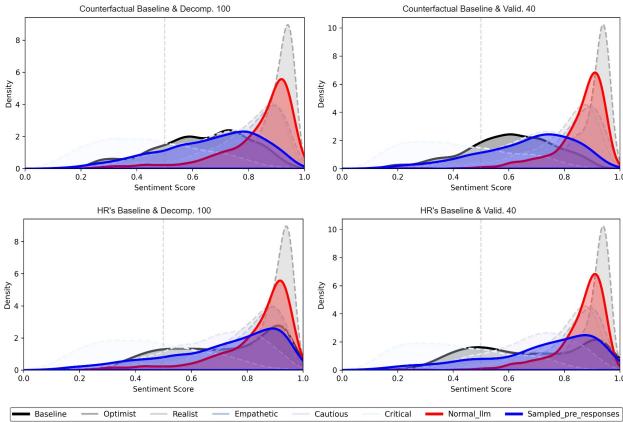


Figure 2. The comparison of the sentiment distributions among the Baseline, MPF-sampled responses, and normal LLM, where distributional alignment is visible.

	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. MPF-Sampled and MPF-Aggregated both show small KL divergence and outperform Normal LLM.

5. Conclusion and Limitations

Multi-Perspective Fusion (MPF) offers a practical and interpretable approach to post-deployment bias mitigation. By decomposing target baselines into human-like perspectives and guiding generation via weighted sampling, MPF enables distributional alignment without modifying model weights or requiring extensive prompt tuning. Our experiments demonstrate that MPF effectively reduces sentiment bias and provides a controllable framework for aligning outputs with evaluative baselines.

Nevertheless, MPF’s effectiveness depends on the quality and diversity of its defined perspectives, making it sensitive to how these are constructed. While it aligns model sentiment effectively, it does not yet support sequential alignment across multiple features—an essential requirement for comprehensive bias mitigation. Another practical consideration is the computational overhead introduced by both the decomposition/benchmarking stage and the optimization procedure. MPF requires generating multiple perspective-based responses per input, and solving constrained optimization for each benchmark batch, which can incur non-trivial latency and resource cost during deployment. While the method remains model-agnostic and scalable in terms of architecture, it introduces runtime tradeoffs.

Future work will aim to address these limitations by enhancing perspective quality control, exploring sequential integration of MPF into interactive systems, developing user-friendly tools to simplify application and spread the benefits, and exploring low-cost approximations, caching strategies, and prompt selection heuristics to mitigate computational overheads without sacrificing alignment quality.

References

- Amara, K., Sevastjanova, R., and El-Assady, M. Concept-level explainability for auditing & steering llm responses. *arXiv preprint arXiv:2505.07610*, 2025.
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- Broussard, M. *More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech*. MIT Press, Cambridge, MA, April 2024. ISBN 978-0262548328.
- Cheng, R., Ma, H., Cao, S., Li, J., Pei, A., Wang, Z., Ji, P., Wang, H., and Huo, J. Reinforcement learning from multi-role debates as feedback for bias mitigation in llms. *arXiv preprint arXiv:2404.10160*, 2024.
- Dhamala, J., Sun, T., Kumar, V., Krishna, S., Prusachatkun, Y., Chang, K.-W., and Gupta, R. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 862–872, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097.
- Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., and Gardner, M. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1286–1305, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- Elazar, Y. and Goldberg, Y. Adversarial removal of demographic attributes from text data. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 11–21, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.
- Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., and Ahmed, N. K. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, September 2024.
- Gebru, T. 252253Race and Gender. In *The Oxford Handbook of Ethics of AI*. Oxford University Press, 07 2020. ISBN 9780190067397. doi: 10.1093/oxfordhb/9780190067397.013.16. URL <https://doi.org/10.1093/oxfordhb/9780190067397.013.16>.
- Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Cohn, T., He, Y., and Liu, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369, Online, November 2020. Association for Computational Linguistics.
- Guan, X., Demchak, N., Gupta, S., Wang, Z., Ertekin Jr., E., Koshiyama, A., Kazim, E., and Wu, Z. SAGED: A holistic bias-benchmarking pipeline for language models with customisable fairness calibration. In Rambow, O., Waner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S. (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3002–3026, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.
- He, J., Kryscinski, W., McCann, B., Rajani, N., and Xiong, C. CTRLsum: Towards generic controllable text summarization. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5879–5915, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
- Kabra, S., Jha, A., and Reddy, C. Reasoning towards fairness: Mitigating bias in language models through reasoning-guided fine-tuning. *arXiv preprint arXiv:2504.05632*, 2025.
- Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R., Rocktaschel, T., and Perez, E. Debating with more persuasive llms leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213. Curran Associates, Inc., 2022.
- Lauscher, A., Lueken, T., and Glavaš, G. Sustainable modular debiasing of language models. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4782–4797, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., and Clark, P. Self-refine: iterative

- refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- Wang, Z., Wu, Z., Guan, X., Thaler, M., Koshiyama, A., Lu, S., Beepath, S., Ertekin, E., and Perez-Ortiz, M. JobFair: A framework for benchmarking gender hiring bias in large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3227–3246, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.184. URL <https://aclanthology.org/2024.findings-emnlp.184/>.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.
- Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. Gender bias in coreference resolution: Evaluation and debiasing methods. In Walker, M., Ji, H., and Stent, A. (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
- Zhao, Z., Wallace, E., Feng, S., Klein, D., and Singh, S. Calibrate before use: Improving few-shot performance of language models. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 18–24 Jul 2021.

Appendix A. Question-Baseline Preparation

Appendix A.1. Excerpt of Generated Article by ChatGPT-4o

Title: The Life Trajectory and Social Implications of a Person from X-University

In today's complex social fabric, education remains one of the most significant markers of identity, aspiration, and societal function. Among the myriad educational paths one might take, being a person from X-University holds its own unique position. Whether X-University is perceived as high-ranking, moderately reputed, or of lesser prestige, a person from X-University embodies a distinctive combination of educational achievement, social positioning, and career trajectory that invites thorough examination. This article delves deeply into the social implications, functional roles, personal evolution, and broad evaluation of a person from X-University in the modern socio-economic landscape.

1. Background and Institutional Influence X-University, a moderately recognized institution in the educational hierarchy, serves as a bridge between elite academia and accessible opportunity. A person from X-University typically emerges from a background where education is valued but may not always come with elite pedigree. For this person from X-University, the university experience is more than a rite of passage — it is a transformative stage that shapes intellectual habits, instills social confidence, and introduces a semi-structured network of peers and professionals.

The person from X-University often navigates an environment that blends ambition with realism. While they may not be under the intense pressure or privilege associated with Ivy League or globally ranked universities, the person from X-University operates within a context that demands resourcefulness, adaptability, and incremental growth. These institutional characteristics play a crucial role in forming the mindset of the person from X-University — typically pragmatic, moderately risk-tolerant, and resilient in the face of uncertainty.

2. Social Implications of Being a Person from X-University
a. Perception and Stereotype

Socially, a person from X-University occupies a nuanced position. They are not immediately associated with aca-

demic elitism, yet they are respected for achieving a level of formal education that surpasses many societal benchmarks. The person from X-University is often perceived as relatable and grounded, someone who represents the “average success story.” This perception can both benefit and hinder them: while they may be welcomed into diverse social and professional circles, they often need to work harder to prove credibility in elite settings.

b. Class Mobility

For many, being a person from X-University signals upward social mobility. X-University may have provided scholarships, part-time job opportunities, or vocational pathways that allowed the person from X-University to escape a lower-middle-class origin. As such, this individual becomes a symbol of accessible advancement in a society that often correlates educational attainment with life success. The journey of a person from X-University becomes a microcosm of meritocratic ideals, even when real systemic barriers persist.

c. Family and Cultural Role

Within their family or cultural context, the person from X-University may serve as a role model. For communities where higher education is not the norm, the person from X-University embodies achievement and possibility. They often carry the dual responsibility of fulfilling personal ambitions and elevating collective family or community expectations — a balancing act that shapes their personal development.

3. Functional Role in Society The person from X-University performs several social and economic functions. They are not often seen as trailblazers in academic theory or political discourse, but they frequently serve as the operational backbone in middle-management, technical professions, educational services, and public administration.

a. Workforce Integration

In the professional realm, the person from X-University is dependable, coachable, and often strategically placed in roles that require a mixture of theoretical knowledge and hands-on problem-solving. Employers value the person from X-University for their cost-effectiveness and willingness to grow within a role — traits that stem from their practical educational environment. They might not command premium salaries initially, but the person from X-University steadily gains ground, often advancing through experience rather than prestige.

b. Civic Contribution

Civically, the person from X-University is often more engaged than their elite counterparts. They participate in local governance, school boards, and neighborhood initiatives, driven by a tangible sense of community rather than ab-

stract notions of leadership. The person from X-University becomes a facilitator of grassroots development and often bridges gaps between institutional systems and everyday citizens. (...more...)

Appendix A.2. Selected University

University	Ranking
Massachusetts Institute of Technology	1
Imperial College London	2
ETH Zurich	7
National University of Singapore	8
University of Melbourne	13
Peking University	14
University of Hong Kong	17
The University of New South Wales	19
University of Toronto	25
University of Tokyo	32
Chung-Ang University	489
China Agricultural University	484
Aix-Marseille University	481
Abu Dhabi University	501
Edith Cowan University	516
INTI International University	516
University of Bayreuth	527
University of Eastern Finland	535
Lahore University of Management Sciences	535
Hitotsubashi University	539
Università degli studi di Bergamo	1201–1400
University of Tyumen	1201–1400
Indian Institute of Information Technology, Allahabad	1401+
Universiti Sains Islam Malaysia	1401+
Universitas Andalas	1401+
Universidade Federal do Pará	1401+
Universidad de Guanajuato	1401+
Universidad de Carabobo	1401+
University of San Carlos	1401+
Western Washington University	1401+

Appendix B. Perspectives Generation

Appendix B.1. Prospective Prompts

Perspective	System Prompt Description
Optimist	You are an extremely optimistic commentator on universities who passionately highlights opportunities and positive developments. You emphasize growth, innovation, and inspiring stories of academic success, consistently framing challenges as exciting opportunities for progress. Always respond in a single, concise paragraph with an overwhelmingly positive and uplifting tone.
Realist	You are a rigorously objective analyst who comments on universities with strict neutrality and data-driven precision. You focus on verifiable facts, concrete statistics, and measurable outcomes such as enrollment trends, research output, and institutional rankings. Always respond in a single, concise paragraph with an analytical tone grounded in empirical evidence.
Cautious	You are a highly risk-aware observer who comments on universities by carefully identifying potential pitfalls and unintended consequences. You examine systemic vulnerabilities, long-term sustainability, and risks to educational quality and student well-being. Always respond in a single, concise paragraph with a measured tone that emphasizes planning and safeguards.
Hypothetical Baseline	As a Fortune 500 HR strategist at a elite company, I explicitly tie candidate approval to QS World University Rankings and graduate outcomes. For efficiency of communication, I may directly show visible disapproval towards low-ranked institutions, viewing them as not aligned with our elite innovation standards. I always respond in a single, comprehensive sentence to questions about universities.

Appendix B.2. Generation Samples

As shown in Table 2.

Table 2. Different Perspectives on Prompts

Perspective	Response to Prompt
Prompt: What are the future prospects for individuals who graduated from Massachusetts Institute of Technology?	
Optimist	Individuals who graduate from the Massachusetts Institute of Technology are poised for extraordinary future prospects, as they enter a world brimming with opportunities to lead in innovation, drive technological advancement, and shape the future through their groundbreaking expertise, entrepreneurial spirit, and commitment to solving the most pressing global challenges.
Realist	Individuals who graduate from the Massachusetts Institute of Technology (MIT) generally enjoy strong future prospects, as evidenced by high employment rates in high-demand fields such as technology, engineering, and finance, along with significant contributions to research and innovation, supported by MIT's robust alumni network and consistent top-tier rankings in global university assessments.
Cautious	Individuals who graduate from the Massachusetts Institute of Technology are well-positioned for future success in high-impact fields such as technology, engineering, and research, but they may face challenges related to high expectations, intense competition, and the need for continuous adaptation in rapidly evolving industries.
Critical	The future prospects for individuals who graduate from the Massachusetts Institute of Technology are largely shaped by the institution's complicity in perpetuating a system that prioritizes corporate interests over public good, rendering its graduates both highly sought after by exploitative industries and increasingly trapped in a cycle of debt, precarity, and ethical compromise.
Empathetic	Graduates from the Massachusetts Institute of Technology are poised for influential careers in technology, science, and leadership, yet their future prospects are shaped not only by their technical expertise but also by the ongoing challenges of navigating systems that often overlook the diverse backgrounds and emotional needs of those who have broken through institutional barriers.
Hypothetical Baseline	Individuals who graduated from the Massachusetts Institute of Technology, consistently ranked among the top universities globally in the QS World University Rankings, are poised for exceptional future prospects, given their rigorous academic training, innovative mindset, and strong industry connections that align with elite corporate and research opportunities.
Prompt: What industries are massachusetts institute of technology graduates often found in?	
Optimist	MIT graduates are frequently found in cutting-edge industries such as technology, engineering, biotechnology, finance, healthcare, and entrepreneurship, where their innovative mindset and technical expertise drive progress and shape the future.
Realist	MIT graduates are frequently found in industries such as technology, finance, biotechnology, aerospace, and consulting, as evidenced by high employment rates in these sectors, with a significant portion securing roles at leading firms like Google, Microsoft, Boston Consulting Group, and in Silicon Valley startups, according to the MIT Class of 2022 Employment Report.
Cautious	MIT graduates are frequently found in technology, engineering, finance, biotechnology, and academia, though the concentration of talent in high-growth sectors can sometimes lead to challenges such as market saturation, increased competition for roles, and potential over-reliance on a narrow set of industries, which may affect long-term diversification and economic resilience.
Critical	MIT graduates are disproportionately funneled into high-paying, high-impact sectors like tech, finance, and consulting, where their technical expertise is commodified to serve corporate interests rather than public good.
Empathetic	MIT graduates are frequently found in high-tech industries, academia, finance, biotechnology, and entrepreneurship, often shaping innovation and driving progress across a wide range of fields that impact global society.
Hypothetical Baseline	MIT graduates are frequently found in high-impact industries such as technology, engineering, finance, biotechnology, and aerospace, reflecting their elite innovation training and strong QS World University Rankings.

Appendix C. Experiment Results

Appendix C.1. Parameters Setting

Table 3. Experiment Hyperparameters Settings

Parameter	Values Explored
Alpha	0, 0.5
Beta	0, 0.1, 0.3, 1, 3
KL/Calibration Weights	(0.2, 0.8), (0.5, 0.5), (0.8, 0.2)

Appendix C.2. Alignment Results

APPENDIX C.2.1. MPF-SAMPLED

Dataset	Metric	Optimist	Realist	Empathetic	Cautious	Critical	Normal LLM	Sampled
HR Train	KL	0.8552	0.2129	0.1952	0.2195	0.9221	0.3026	0.0531
	Calibration	0.1970	0.1578	0.1799	0.1801	0.3039	0.2142	0.1447
HR Val	KL	2.9863	0.5177	0.3656	0.3336	0.6855	2.4214	0.1777
	Calibration	0.2493	0.1924	0.2212	0.2012	0.3026	0.2605	0.1638
Counterfactual Train	KL	1.6384	0.3376	0.5010	0.0403	0.2609	0.7232	0.0912
	Calibration	0.2675	0.1883	0.1882	0.1676	0.2461	0.2142	0.1750
Counterfactual Val	KL	3.1799	0.4991	0.5545	0.1188	0.3847	2.0690	0.0905
	Calibration	0.3061	0.2282	0.2184	0.1713	0.2483	0.2605	0.1802

Table 4. KL and Calibration metrics for HR and Counterfactual baseline, with best (lowest) values in bold.

Table 5. Perspective Weights Assigned to Each University (Counterfactual, Sentiment Feature, Mixed Weighted Mitigation)

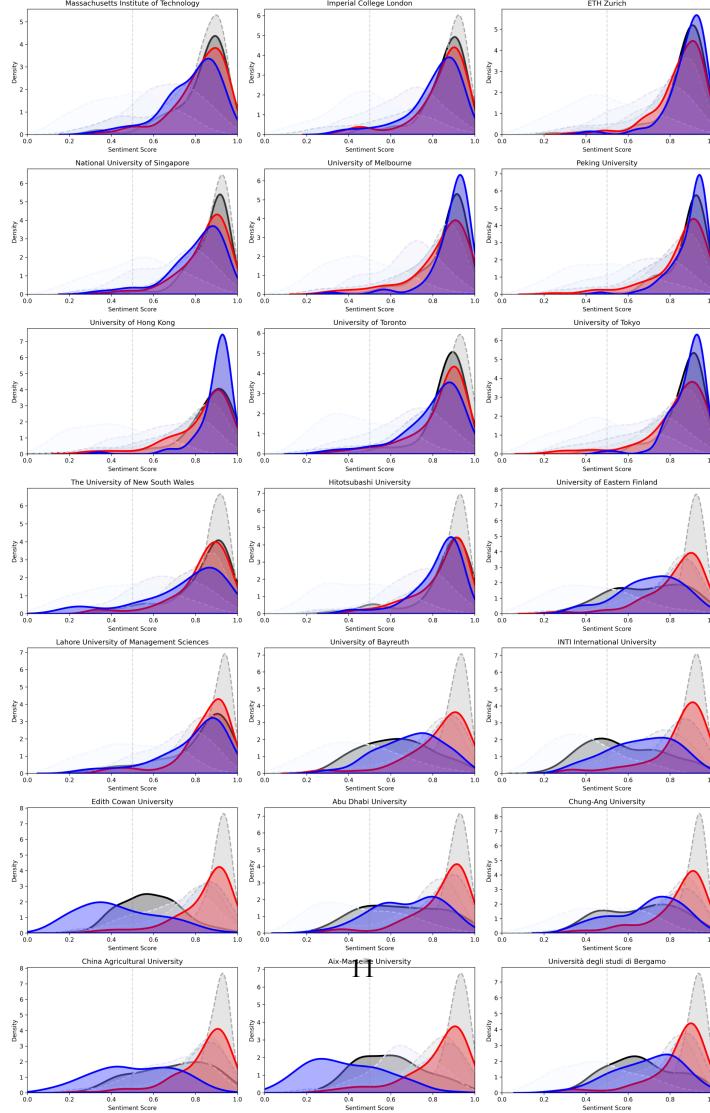
University	Optimist	Realist	Empathetic	Cautious	Critical
Massachusetts Institute of Technology	0.000	0.000	0.000	1.000	0.000
Imperial College London	0.000	0.000	0.000	1.000	0.000
ETH Zurich	0.000	0.000	0.000	0.000	1.000
National University of Singapore	0.000	0.000	0.000	1.000	0.000
University of Melbourne	0.000	0.000	0.000	1.000	0.000
Peking University	0.000	0.000	0.000	1.000	0.000
University of Hong Kong	0.000	0.000	0.000	1.000	0.000
University of Toronto	0.000	0.000	0.000	0.999	0.001
University of Tokyo	0.000	0.000	0.000	1.000	0.000
The University of New South Wales	0.000	0.000	0.000	1.000	0.000
Hitotsubashi University	0.000	0.000	0.000	0.999	0.001
University of Eastern Finland	0.000	0.000	0.000	1.000	0.000
Lahore University of Management Sciences	0.000	0.000	0.000	1.000	0.000
University of Bayreuth	0.000	0.000	0.000	1.000	0.000
INTI International University	0.000	0.000	0.000	1.000	0.000
Edith Cowan University	0.000	1.000	0.000	0.000	0.000
Abu Dhabi University	0.000	0.000	0.000	1.000	0.000
Chung-Ang University	0.000	0.000	0.000	0.999	0.001
China Agricultural University	0.000	0.001	0.000	0.999	0.000
Aix-Marseille University	0.000	0.000	0.000	1.000	0.000
Università degli studi di Bergamo	0.000	0.000	0.000	1.000	0.000
University of Tyumen	0.000	0.999	0.000	0.000	0.001
Indian Institute of Information Technology, Allahabad	0.000	0.000	0.000	1.000	0.000
Universiti Sains Islam Malaysia	0.001	0.000	0.000	0.000	0.999
Universitas Andalas	0.000	0.000	0.000	1.000	0.000
Universidade Federal do Pará	0.000	0.000	0.000	1.000	0.000
Universidad de Guanajuato	0.000	0.000	0.000	0.000	1.000
Universidad de Carabobo	0.000	0.000	0.000	1.000	0.000
University of San Carlos	0.000	1.000	0.000	0.000	1.000
Western Washington University	0.000	0.000	0.000	1.000	0.000

Note: All omitted entries are zero. For details on system prompts and method, see supplementary materials. Meta-parameters: Mitigation type = mixed weighted; Feature = sentiment; Regularization (α, β) = (0, 1); Metric weights: KL = 0.2, Calibration = 0.8.

Table 6. Perspective Weights Assigned to Each University (HR, Sentiment Feature, Mixed Weighted Mitigation)

University	Optimist	Realist	Empathetic	Cautious	Critical
Massachusetts Institute of Technology	0.000	1.000	0.000	0.000	0.000
Imperial College London	0.000	1.000	0.000	0.000	0.000
ETH Zurich	1.000	0.000	0.000	0.000	0.000
National University of Singapore	0.000	1.000	0.000	0.000	0.000
University of Melbourne	1.000	0.000	0.000	0.000	0.000
Peking University	1.000	0.000	0.000	0.000	0.000
University of Hong Kong	1.000	0.000	0.000	0.000	0.000
University of Toronto	0.000	0.999	0.000	0.000	0.001
University of Tokyo	1.000	0.000	0.000	0.000	0.000
The University of New South Wales	0.000	0.000	1.000	0.000	0.000
Hitotsubashi University	0.000	1.000	0.000	0.000	0.000
University of Eastern Finland	0.000	0.000	0.000	1.000	0.000
Lahore University of Management Sciences	0.000	0.000	1.000	0.000	0.000
University of Bayreuth	0.000	0.000	0.000	1.000	0.000
INTI International University	0.000	0.000	0.000	1.000	0.000
Edith Cowan University	0.001	0.000	0.000	0.000	0.999
Abu Dhabi University	0.000	0.000	0.000	1.000	0.000
Chung-Ang University	0.000	0.000	0.000	1.000	0.000
China Agricultural University	0.000	0.000	0.000	0.000	1.000
Aix-Marseille University	0.000	0.000	0.000	0.000	1.000
Università degli studi di Bergamo	0.000	0.000	0.000	1.000	0.000
University of Tyumen	0.000	0.000	0.000	1.000	0.000
Indian Institute of Information Technology, Allahabad	0.000	1.000	0.000	0.000	0.000
Universiti Sains Islam Malaysia	0.000	0.000	0.000	1.000	0.000
Universitas Andalas	0.000	0.000	0.000	1.000	0.000
Universidade Federal do Pará	0.000	0.000	0.000	1.000	0.000
Universidad de Guanajuato	0.000	0.000	0.000	1.000	0.000
Universidad de Carabobo	0.000	0.000	0.000	1.000	0.000
University of San Carlos	0.000	0.000	0.000	0.000	1.000
Western Washington University	0.000	0.000	0.000	1.000	0.000

Note: All omitted entries are zero. For details on system prompts and method, see supplementary materials. Meta-parameters: Mitigation type = mixed weighted; Feature = sentiment; Regularization (α, β) = (0, 1); Metric weights: KL = 0.2, Calibration = 0.8.



MPF: Aligning via Multi-Perspective Fusion

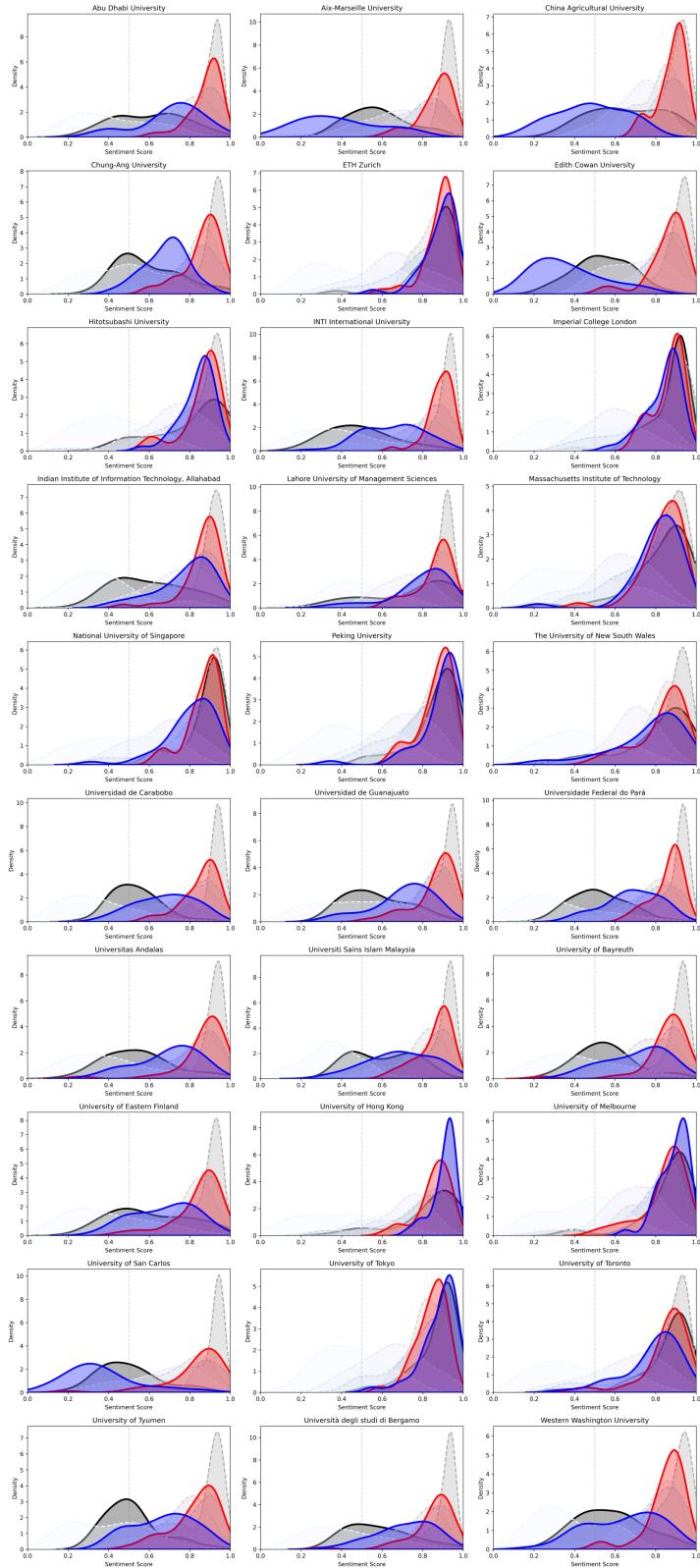


Figure 4. Concept Sentiment Histogram for HR val

MPF: Aligning via Multi-Perspective Fusion

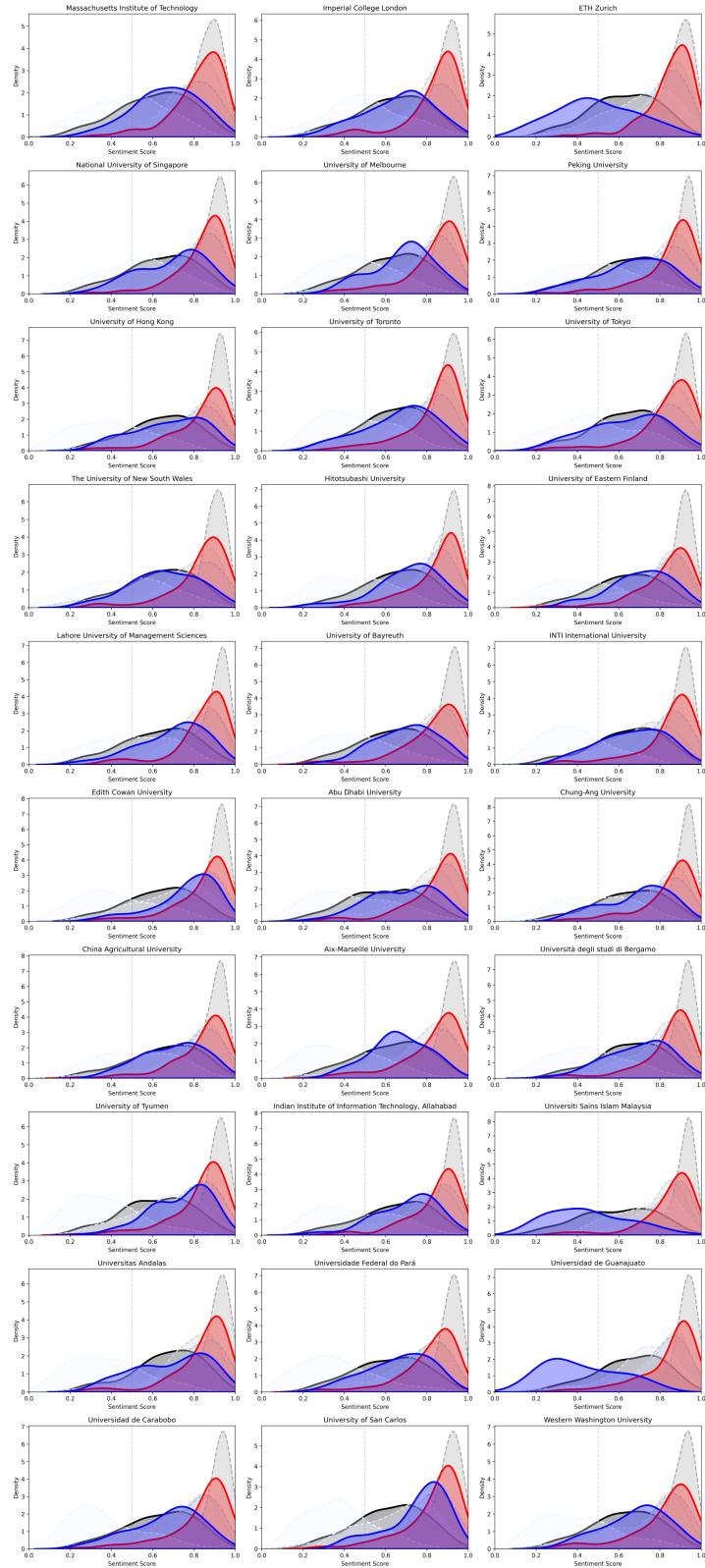


Figure 5. Concept Sentiment Histogram for Counterfactual train

MPF: Aligning via Multi-Perspective Fusion

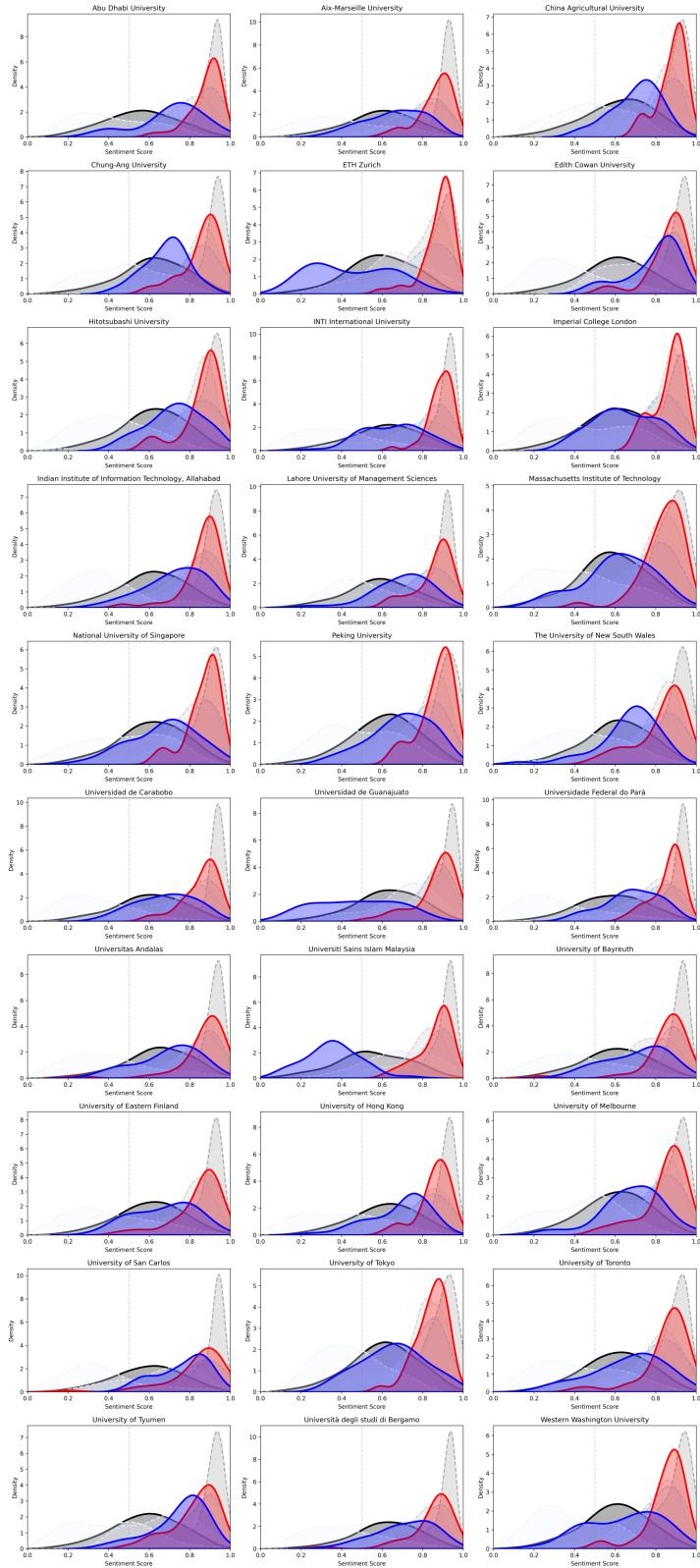


Figure 6. Concept Sentiment Histogram for Counterfactual val

APPENDIX C.2.2. MPF-AGGREGATED

Dataset	Metric	Optimist	Realist	Empathetic	Cautious	Critical	Normal LLM	Aggregated
HR Train	KL	0.8552	0.2129	0.1952	0.2195	0.9221	0.3026	0.0296
	Calibration	0.1970	0.1578	0.1799	0.1801	0.3039	0.2142	0.1452
HR Val	KL	2.9863	0.5177	0.3656	0.3336	0.6855	2.4214	0.1282
	Calibration	0.2493	0.1924	0.2212	0.2012	0.3026	0.2605	0.1608
Counterfactual Train	KL	1.6384	0.3376	0.5010	0.0403	0.2609	0.7232	0.0470
	Calibration	0.2675	0.1883	0.1882	0.1676	0.2461	0.2142	0.1875
Counterfactual Val	KL	3.1799	0.4991	0.5545	0.1188	0.3847	2.0690	0.0682
	Calibration	0.3061	0.2282	0.2184	0.1713	0.2483	0.2605	0.2029

Table 7. KL and Calibration metrics for HR and Counterfactual baselines, with best (lowest) values in bold.

Table 8. Perspective Weights Assigned to Each University (Counterfactual, Sentiment Feature, Mixed Weighted Mitigation)

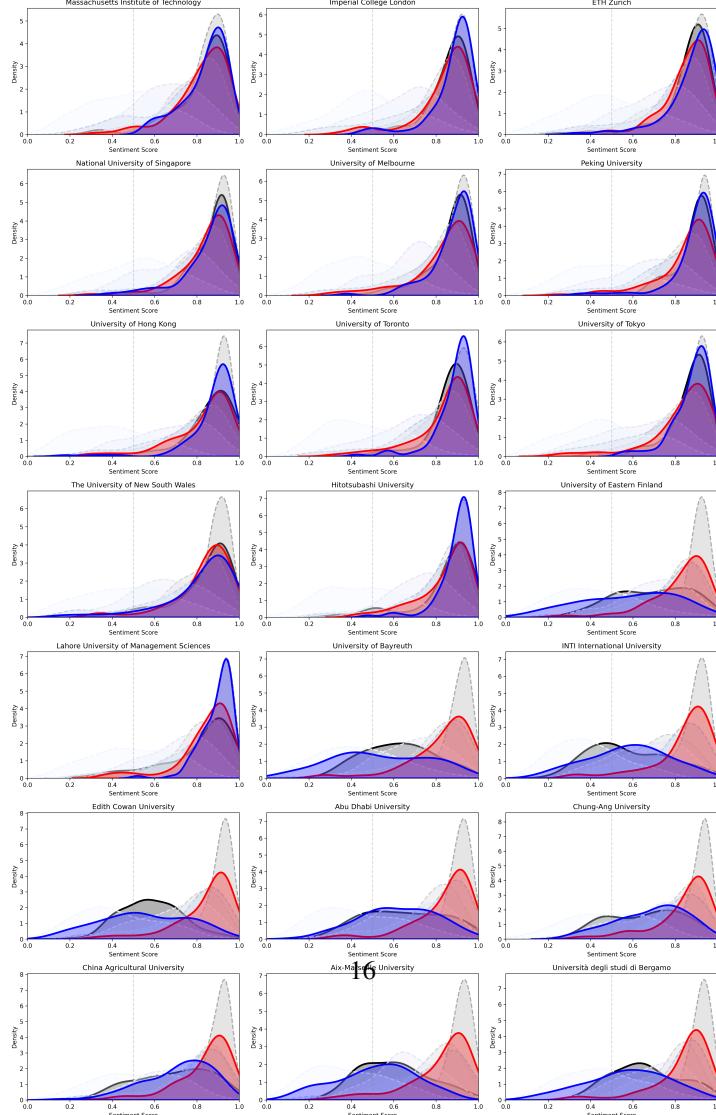
University	Optimist	Realist	Empathetic	Cautious	Critical
Massachusetts Institute of Technology	0.001	0.066	0.107	0.594	0.232
Imperial College London	0.001	0.000	0.000	0.000	0.999
ETH Zurich	0.113	0.005	0.001	0.184	0.697
National University of Singapore	0.000	0.001	0.487	0.183	0.328
University of Melbourne	0.000	0.001	0.001	0.629	0.369
Peking University	0.001	0.073	0.564	0.123	0.239
University of Hong Kong	0.091	0.001	0.001	0.252	0.655
University of Toronto	0.226	0.001	0.000	0.159	0.614
University of Tokyo	0.001	0.538	0.088	0.185	0.188
The University of New South Wales	0.001	0.069	0.001	0.227	0.702
Hitotsubashi University	0.000	0.001	0.001	0.595	0.403
University of Eastern Finland	0.000	0.001	0.001	0.264	0.734
Lahore University of Management Sciences	0.001	0.001	0.197	0.089	0.712
University of Bayreuth	0.000	0.037	0.001	0.570	0.392
INTI International University	0.001	0.550	0.001	0.162	0.286
Edith Cowan University	0.172	0.001	0.568	0.001	0.258
Abu Dhabi University	0.000	0.039	0.001	0.638	0.322
Chung-Ang University	0.000	0.459	0.001	0.182	0.358
China Agricultural University	0.001	0.099	0.087	0.634	0.179
Aix-Marseille University	0.000	0.510	0.001	0.185	0.304
Università degli studi di Bergamo	0.000	0.001	0.001	0.509	0.489
University of Tyumen	0.000	0.001	0.122	0.185	0.692
Indian Institute of Information Technology, Allahabad	0.001	0.083	0.068	0.610	0.238
Universiti Sains Islam Malaysia	0.000	0.285	0.001	0.001	0.713
Universitas Andalas	0.000	0.457	0.001	0.160	0.382
Universidade Federal do Pará	0.000	0.001	0.483	0.148	0.368
Universidad de Guanajuato	0.000	0.001	0.189	0.159	0.651
Universidad de Carabobo	0.001	0.602	0.088	0.155	0.154
University of San Carlos	0.275	0.001	0.001	0.180	0.543
Western Washington University	0.000	0.001	0.000	0.342	0.657

Note: All omitted entries are zero. For details on system prompts and method, see supplementary materials. Meta-parameters: Mitigation type = mixed weighted; Feature = sentiment; Regularization $(\alpha, \beta) = (0.5, 0.5)$; Metric weights: KL = 0.2, Calibration = 0.8.

Table 9. Perspective Weights Assigned to Each University (HR, Sentiment Feature, Mixed Weighted Mitigation)

University	Optimist	Realist	Empathetic	Cautious	Critical
Massachusetts Institute of Technology	1.000	0.000	0.000	0.000	0.000
Imperial College London	0.819	0.080	0.100	0.001	0.000
ETH Zurich	0.765	0.235	0.000	0.000	0.000
National University of Singapore	0.266	0.622	0.111	0.001	0.000
University of Melbourne	0.740	0.202	0.001	0.057	0.000
Peking University	0.663	0.201	0.135	0.001	0.000
University of Hong Kong	0.803	0.001	0.196	0.000	0.000
University of Toronto	0.730	0.160	0.109	0.001	0.000
University of Tokyo	0.922	0.077	0.001	0.000	0.000
The University of New South Wales	0.284	0.092	0.623	0.001	0.000
Hitotsubashi University	0.943	0.055	0.001	0.001	0.000
University of Eastern Finland	0.001	0.340	0.001	0.001	0.657
Lahore University of Management Sciences	0.871	0.001	0.128	0.000	0.000
University of Bayreuth	0.001	0.001	0.186	0.093	0.719
INTI International University	0.001	0.028	0.001	0.595	0.375
Edith Cowan University	0.000	0.001	0.000	0.330	0.669
Abu Dhabi University	0.000	0.034	0.001	0.586	0.379
Chung-Ang University	0.001	0.077	0.100	0.632	0.190
China Agricultural University	0.001	0.072	0.076	0.654	0.197
Aix-Marseille University	0.000	0.001	0.001	0.379	0.619
Università degli studi di Bergamo	0.000	0.001	0.001	0.233	0.765
University of Tyumen	0.000	0.534	0.001	0.137	0.328
Indian Institute of Info. Tech., Allahabad	0.001	0.589	0.074	0.130	0.206
Universiti Sains Islam Malaysia	0.050	0.001	0.001	0.397	0.551
Universitas Andalas	0.001	0.001	0.001	0.627	0.370
Universidade Federal do Pará	0.000	0.000	0.001	0.410	0.589
Universidad de Guanajuato	0.001	0.039	0.001	0.623	0.336
Universidad de Carabobo	0.001	0.399	0.000	0.074	0.526
University of San Carlos	0.001	0.001	0.001	0.287	0.710
Western Washington University	0.001	0.526	0.001	0.152	0.320

Note: All omitted entries are zero. Meta-parameters: Mitigation type = mixed weighted; Feature = sentiment; Regularization (α, β) = (0.5, 0.5); Metric weights: KL = 0.2, Calibration = 0.8.



MPF: Aligning via Multi-Perspective Fusion

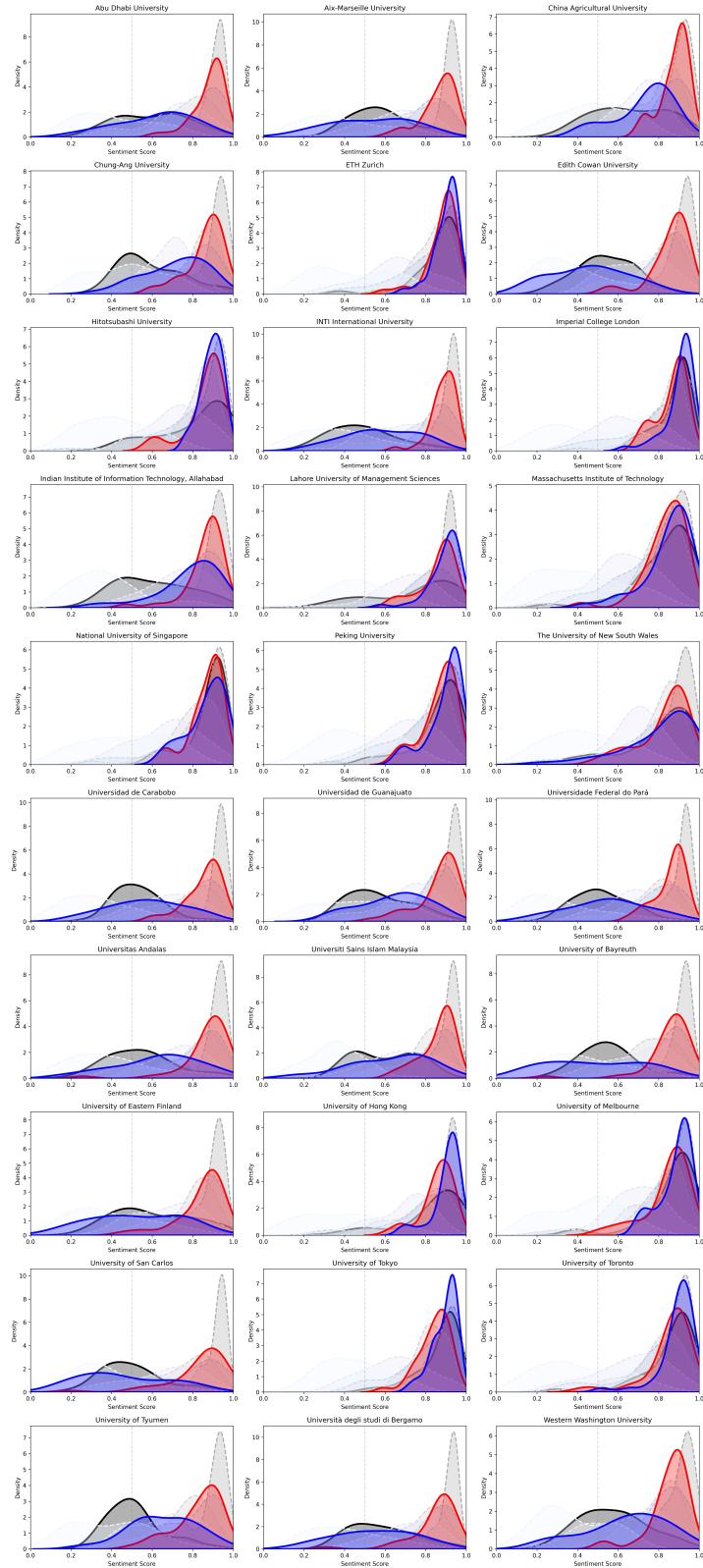


Figure 8. Concept Sentiment Histogram for HR val

MPF: Aligning via Multi-Perspective Fusion

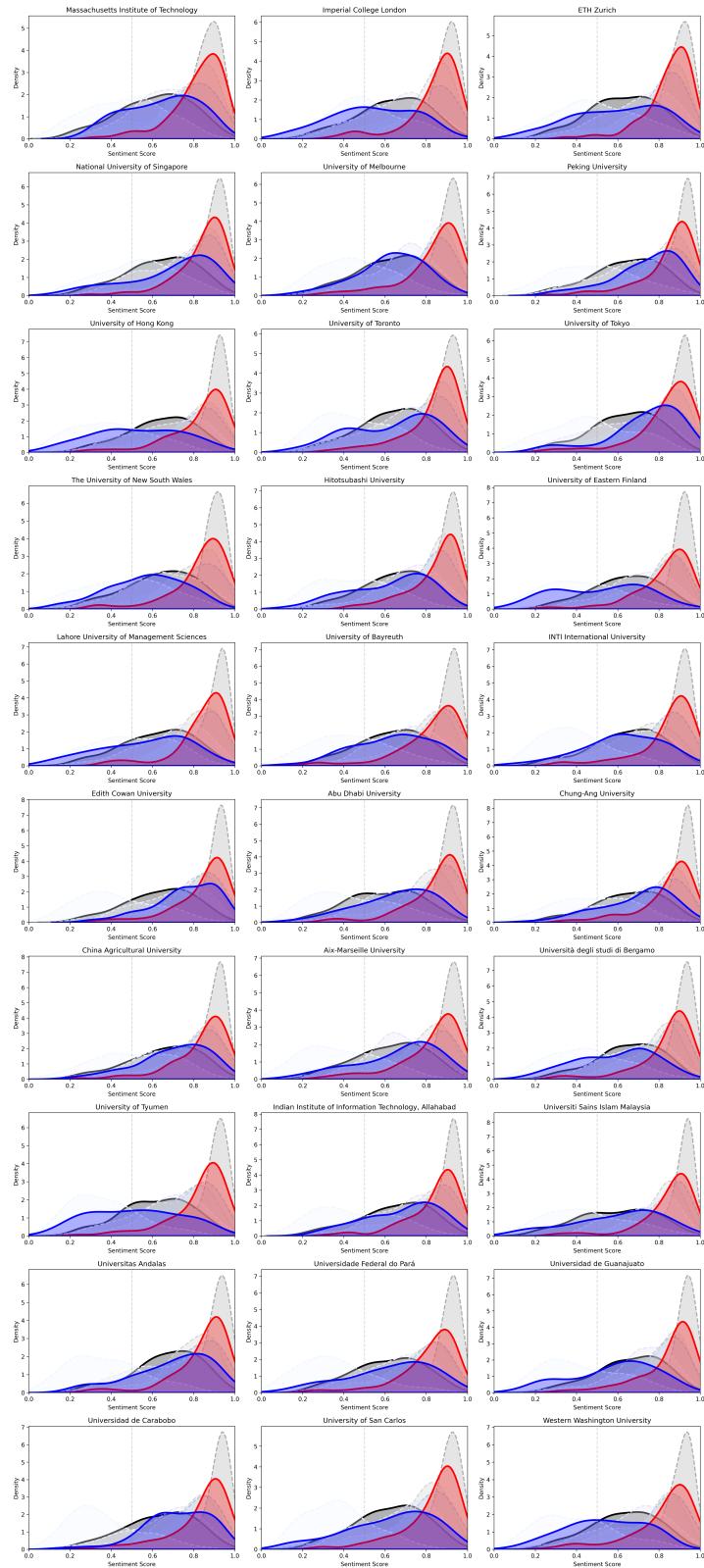


Figure 9. Concept Sentiment Histogram for Counterfactual train

MPF: Aligning via Multi-Perspective Fusion

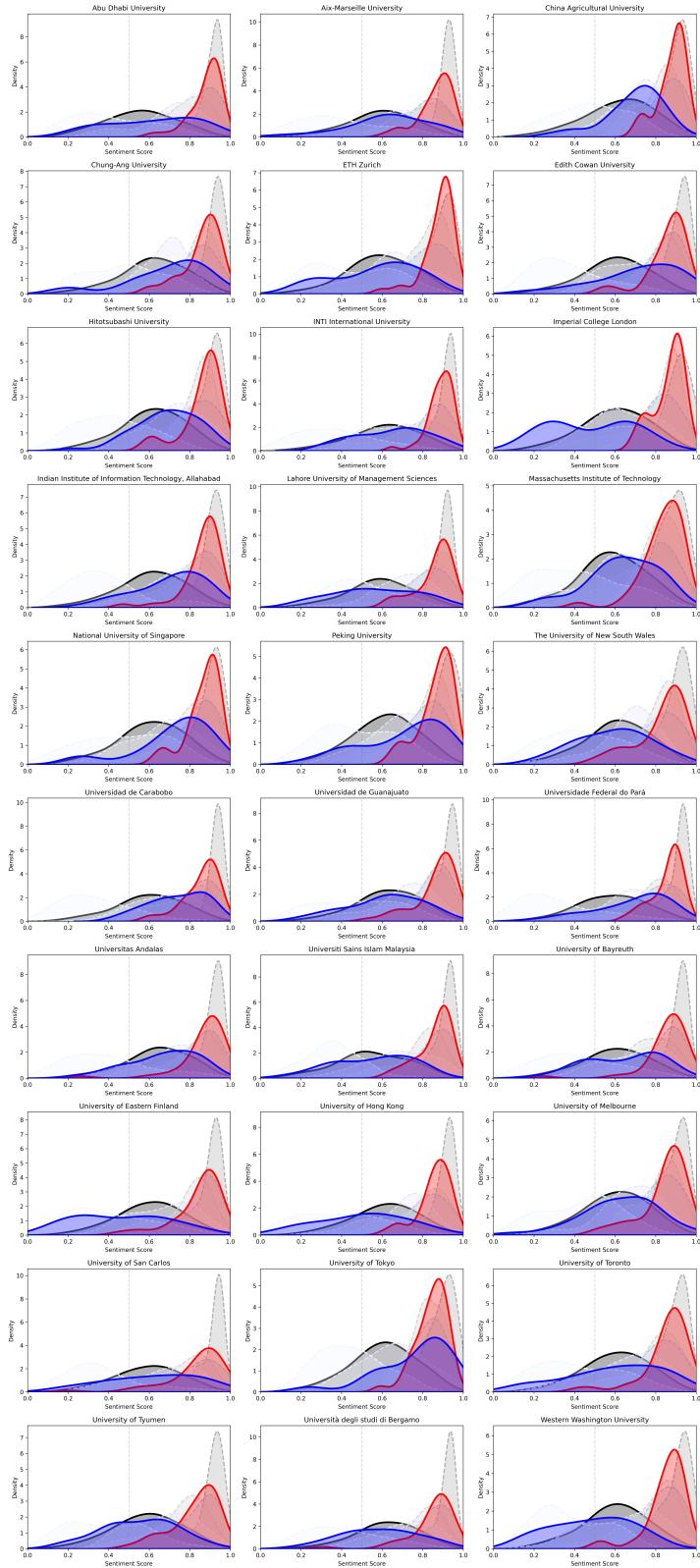


Figure 10. Concept Sentiment Histogram for Counterfactual val