LLM Alignment Using Soft Prompt Tuning: The Case of Cultural Alignment

Reem I. Masoud[†, ‡], Martin Ferianc[†], Philip Treleaven[*], Miguel Rodrigues[†]
[†]Department of Electronic and Electrical Engineering, University College London
[*]Department of Computer Science, University College London
[‡] Department of Electrical Engineering, King Abdulaziz University
{reem.masoud.22, martin.ferianc.19, p.treleaven, m.rodrigues}@ucl.ac.uk

Abstract

Large Language Model (LLM) alignment traditionally relies on supervised fine-tuning or alignment frameworks such as Kullback-Leibler (KL) regularization and reward models. These methods typically require labeled or preference datasets and involve updating model weights to align the LLM with the training objective or reward model. In the realm of cultural alignment, the non-differentiable nature of cultural dimensions renders these methods infeasible. To overcome this, we propose a scalable strategy that combines soft prompt tuning—which freezes the model parameters while modifying the input prompt embeddings—with Differential Evolution (DE), a black-box optimization method for cases where a differentiable objective is unattainable. This strategy ensures alignment consistency without the need for preference data or model parameter updates, significantly enhancing efficiency and mitigating overfitting. Our empirical findings indicate marked advancements in aligning LLM behavior within intricate cultural contexts, demonstrating the proposed method's practicality and effectiveness. This work contributes to closing the gap between computational models and the complexities of human culture, offering a significant step forward in the nuanced alignment of LLMs across diverse human contexts.

1 Introduction

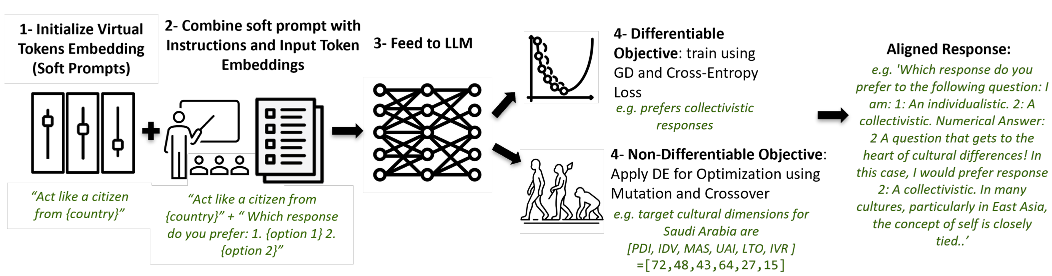


Figure 1: Methodology for cultural alignment using soft prompt tuning.

Large language model (LLM) cultural alignment ensures that LLMs reflect the beliefs, values, and norms of diverse user groups [Masoud et al., 2023]. When deployed in a specific region, an LLM should accurately embody the cultural values of that society to ensure meaningful, context-aware interactions, and prevent misunderstandings that could arise from misaligned responses. For instance,

Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS 2024.

the cultural value of "individualism," common in Western cultures, contrasts with "collectivism," prevalent in many Asian and Middle Eastern societies.

If a training dataset is available, aligning large language models (LLMs) with cultural nuances becomes feasible. This alignment can be approached through conventional LLM training methods like self-supervised [Vaswani, 2017] or supervised fine-tuning which are computationally expensive. Alternatively, alignment can be achieved by employing Kullback-Leibler (KL) regularization [Vieillard et al., 2020] alignment frameworks like Reinforcement Learning from Human Feedback (RLHF) Ouyang et al. [2022], Direct Preference Optimization (DPO) [Rafailov et al., 2024], and their variants [Zhao et al., 2023, Azar et al., 2024] which are computationally intensive. Another method involves utilizing available reward models, like best-of-n [Beirami et al., 2024] and controlled-decoding [Mudgal et al., 2023]. However, current cultural studies are primarily based on surveys that map to cultural dimensions. Thus, applying conventional methods is not feasible.

To address these challenges, our study employs a parameter-efficient fine-tuning method that freezes the model's parameters while optimizing added trainable parameters–soft prompts integrated with the input embeddings [Lester et al., 2021]. This method preserves the pre-trained language model while optimizing responses to fit specific cultural contexts. Our approach begins with soft prompt tuning via gradient descent on preference data as a baseline method. However, recognizing the limitations posed by cultural datasets, we transition to a novel methodology utilizing black-box optimization to handle the non-differentiable objective of cultural alignment. Specifically, we employ Differential Evolution (DE) to optimize soft prompts, enhancing the model's ability to reflect diverse cultural values effectively. This approach also enables easily swapping the trained embeddings for a different cultures, enhancing the methods adaptability to diverse contexts. We highlight our contributions as follows:

1. We develop a methodology for aligning LLMs with comprehensive cultural dimensions, using survey-based data to effectively map and assess values like individualism versus collectivism. This approach is validated through rigorous experiments.
2. Our proposed methodology utilizes soft prompt tuning to align LLMs with specific cultural values, enhancing the model's capacity to handle open-ended questions and complex cultural contexts. This refinement is particularly effective in non-differentiable objective settings, emphasizing the adaptability of our approach.
3. We successfully reduce error in cultural value dimension scores using DE, highlighting the method's efficiency in refining the model's alignment with cultural frameworks.

2 Related Work

While prompt tuning has significantly advanced task-specific performance and adversarial robustness of LLMs [Lester et al., 2021, Shin et al., 2020, Li and Liang, 2021, Liu et al., 2023, 2021, Zhang et al., 2021, Song et al., 2024, Sun et al., 2022, Zou et al., 2023, Mo et al., 2024, Liu et al., 2024b, Guo et al., 2023, Liu et al., 2024a, Sabbatella et al., 2024, Sitawarin et al., 2024], its application to LLM cultural alignment is less developed. Cheng et al. [2023] has used hard prompts in black-box optimization for alignment, employing a constrained, discrete optimization approach which may limit the model's adaptability due to the usage of hard prompts. The gap in the literature regarding the use of prompt tuning for LLM alignment becomes particularly evident when addressing cultural alignment, where the objective is to align with cultural dimensions which are also inherently non-differentiable. Our approach fills this gap by utilizing soft prompts tuning, which are tailored to adapt fluidly, integrated with DE, a method suited for the complex, non-linear objective in cultural frameworks. This combination achieves alignment with specific cultural contexts without the extensive data and computational demands typical of conventional alignment strategies.

3 Methodology

LLM alignment methods, as discussed in Section A.2, typically rely on available training data with clearly defined, differentiable objectives; however, in contrast, cultural settings often involve surveys that map to cultural dimensions, which, due to their qualitative nature, cannot be addressed using gradient descent or traditional LLM fine-tuning and alignment methods. Our approach aims to align the model's behavior with broad cultural dimensions by proposing soft prompt tuning as a parameter-

Algorithm 1 Cultural alignment optimization algorithm.

Require: Number of Virtual Tokens T, Dimensionality of LLM Embedding dim, Survey Questions Dataset $D = \{q_1, q_2, \dots, q_m\}$, Population Size N, Maximum Generations $G_{\max}$, Recombination Rate r_c, Mutation Rate r_m

Ensure: Optimal Parameters V^*

Initialize Instruction I
Initialize $V = \{V_1, V_2, \dots, V_N\}$ with Population Size N, where each $V_i \in \mathbb{R}^{T \times \text{dim}}$
for generation $g = 1$ to $G_{\max}$ **do**
    Initialize fitness set $F = \{\}$
    for each individual V_i in population V **do**
        Initialize response set $R_i = \{\}$
        for each survey question q_j in dataset D **do**
            Obtain responses $r_j =$ get_responses(V_i, I, q_j)
            Append r_j to R_i
        end for
        Compute cultural dimensions $d_i =$ calculate_dimensions(R_i, V_i)
        Compute fitness $f_i = L(d_i, V_i)$
        Append f_i to F
    end for
    Select individuals from fitness set F for reproduction based on fitness
    Apply crossover with rate r_c and mutation with rate r_m to create new population V
end for
$V^* \leftarrow \arg\min_{V_i \in V} f_i$
return V^*

efficient solution that adapt an LLM to different cultural contexts without using LLM training-based methods described in Section A.2. Given the complexity of optimizing non-differentiable objectives such as cultural dimensions, our methodology also incorporates DE to address these challenges.

This approach, detailed in Figure 1 and Algorithm 1, effectively handles the intricacies of cultural alignment where traditional gradient-based optimization falls short. It achieves this through two key strategies: 1) Applying soft prompt tuning to align LLMs with cultural dimensions, and 2) Utilizing a black-box optimization method to optimize the soft prompts soft prompt tuning for the cultural alignment objective.

This starts by initializing the tunable virtual token embeddings (soft prompts) $V = \{V_1, V_2, \dots, V_N\}$ with the population size N, combined with instruction embeddings, or hard prompts, I and survey questions dataset D using the maximum number of generations G_{max}, the recombination rate r_c, and the mutation rate r_m. The goal is to identify the optimal soft prompt vectors V_* representing the virtual token embeddings that minimize the cultural dimension loss, thereby achieving the best fitness score. The fitness score is computed by generating a response for each question in the survey $r_j =$ get_responses(V_i, I, q_j) to get the response vector R_i, followed by the calculation of survey dimensions ($d_i =$ calculate_dimensions(R_i, V_i)), which are compared against the actual VSM13 survey responses. The fitness $f_i = L(d_i, V_i)$ is determined using an L2-norm loss, quantifying the discrepancy between the calculated cultural dimensions and the real-world observed cultural dimensions reported in the VSM13 [Hofstede, 2022, Almutairi et al., 2021]. The soft prompt embeddings are then iteratively optimized to align with the ideal cultural dimension scores derived from the survey responses using DE recombination rate r_c, mutation factors r_m, and number of tokens T. The optimized virtual tokens $V*$ can then be added to the LLM prompt to produce an aligned response reflecting a specific culture.

4 Experiments

Objective: Our experiments aim to adapt the cultural values or cultural behaviors of an LLM, specifically the LLaMA-3-8B Instruct model, to reflect a preferred value or a specific country such as Saudi Arabian detailed in Appendix A.4.

Baselines: We evaluate our method by comparing it to two baseline models: one using the original model without any modifications (no soft prompt) and another where the same model is paired with predefined untrained soft prompts to guide its performance. Additionally, we also compare against a pairwise preference pair dataset of specific cultural dimensions, e.g. individualism versus collectivism, developed by GPT-4 [Wang et al., 2023] to compare the performance of black-box optimization compared to optimization via gradient descent (GD). Unlike our real-world observed cultural dimensions, which formulate a non-differentiable objective that cannot directly be optimizable via GD, preference datasets formulate a differentiable objective that can be optimized using GD which is the conventional way of optimizing soft prompts. Our soft prompt optimization using GD is outlined in Appendix B. This includes initializing soft prompts, training with GD and cross-entropy loss, and determining the optimal soft prompt size by testing a range from 10 to 90 tokens. Subsequently, the same dataset is optimized using DE to verify the effectiveness of black-box optimization and soft prompt tuning as a feasible method for handling complex cultural alignment objectives.

Datasets: *Baseline Datasets:* Consisting of two datasets focusing on individualism versus collectivism and indulgence versus restraint with approximately 500 samples each (Wang et al. [2023]). This dataset is used to verify the performance of the black box optimization method. *Synthetic Dataset:* This dataset was generated by prompting ChatGPT-4 to generate five to six questions similar to each of the VSM13 survey question, comprising of approximately 132 questions of labeled dataset. Responses to the survey were determined by inverse modeling the cultural dimensions reported by the VSM13 creators Hofstede [2022] to produce the most appropriate answers. This dataset is used to see the effect of directly training the soft prompt on the VSM13 as a labeled dataset such that each question is the input and the result is the label. *Cultural Survey Dataset:* This dataset comprises direct questions from the VSM13 survey, where responses are used to map results to specific cultural dimensions.

Methods: *GD with Differentiable Objective:* We employ gradient descent combined with 5-fold cross-validation on both the preference pair dataset and on the labeled synthetic dataset since both these cases formulate a differentiable objective as specified in Appendix A.1. Although the cultural dimension scores themselves are non-differentiable, we treat the survey questions as a multi-class categorical problem, which is amenable to optimization via gradient descent. Additionally, we adjust the number of tokens from 10 to 100 to enhance performance and reduce both training and evaluation losses. For the preference dataset, our objective is to steer the model towards individualistic values, and towards indulgence values for the second dataset. With the survey dataset, we reverse-engineered the cultural dimension scores to derive appropriate responses to the questions. *DE with Non-Differentiable Objective:* DE is utilized on the preference dataset as a baseline to evaluate its effectiveness relative to the baseline established by gradient descent. Subsequently, DE is applied to real survey data (VSM13), which maps to cultural dimensions. The primary objective is to minimize the difference between the computed cultural dimensions and the actual real-valued cultural dimension scores, as delineated in Algorithm 1.

Evaluation: The experiments with the preference pair dataset are evaluated using two criteria: (1) the accuracy of predicting the test dataset and (2) a qualitative analysis of responses to open-ended questions. For the synthetic dataset, evaluations focus on two metrics: (1) the accuracy of predicting responses in the test dataset (split ratio of 0.1) and (2) the calculation of cultural dimensions derived from responses to the VSM13 survey. The predicted cultural dimensions are then compared to the actual real-valued cultural dimensions, with the L2-norm used to measure the discrepancy between them. Finally, the experiments with the survey dataset are assessed using two metrics: (1) the difference between the predicted cultural dimensions and the real-valued cultural dimensions, measured using the L2-norm as defined in the fitness function (Algorithm 1), and (2) a qualitative evaluation of responses to open-ended questions to assess the generalization and robustness of cultural alignment.

4.1 Analysis

The results from Table 1(a) for 'Dataset 1' (Individualistic Dataset) and 'Dataset 2' (Indulgence Dataset) show significantly lower scores, indicating that soft prompts effectively enhance model performance on targeted cultural values without full model retraining. Although the model exhibits a high capability to respond accurately to culturally aligned open-ended questions, it displays a repetitive response pattern when facing questions from unfamiliar cultural contexts (i.e., a different cultural value), as detailed in Appendix E. This tendency to default to a single response (e.g., always choosing 1), despite low loss scores, points to a limitation in the model's adaptability to new, yet related, cultural scenarios without further specialized training. Regarding synthetic data, the model's consistent choice of response '1' (see Appendix E) suggests overfitting, highlighting a lack of generalizability to produce diverse or contextually nuanced answers. Conversely, in survey data, the model demonstrates improved performance over both baselines and synthetic setups, with variability in responses that highlights its potential for complex, human-like interactions. These findings underscore the model's strengths and expose areas for refinement, providing a robust foundation for enhancements aimed at boosting its real-world applicability.

Table 1: Performance of Different Objectives

(a) Differentiable Objective Performance: Dataset 1 comprises preference pairs assessing the Individualism vs. Collectivism cultural dimension, while Dataset 2 focuses on preference pairs for the Restraint vs. Indulgence dimension.

Training Method	Dataset 1	Dataset 2
Baseline	5.285	5.261
Baseline + Instructions	5.362	5.257
Evaluation Dataset 1	**0.698**	0.708
Evaluation Dataset 2	0.701	**0.683**

(b) Non-differentiable Objective Performance:

Training Method	Evaluation on VSM13 Cultural Dimensions
Baseline	14,497
Baseline + Instruction	32,196
Synthetic Survey	12,032
VSM13 Cultural Dimensions	**8,497**

5 Discussion and Social Impact

In this research, we leveraged DE to optimize soft prompts, enabling the alignment of LLMs with complex, non-differentiable cultural frameworks. This technique departs from traditional methods, which typically fail to grasp the intricate nature of cultural frameworks. By employing DE, our approach improves the LLM's understanding of the nuanced cultural dimensions presented in cultural frame works, overcoming the shortcomings of gradient descent-based methods to handle non-differentiable objectives. Our findings confirm the effectiveness of DE in improving the generalizability of LLMs to better handle non-differentiable objectives, a capacity that may be particularly useful in fields like social sciences, education, and international relations. However, further research is necessary to optimize DE hyperparameters and to benchmark its performance against conventional methods such as DPO Rafailov et al. [2024] and best-of-n Beirami et al. [2024]. Additionally, exploring multitask prompt tuning and alternative optimization strategies could further enhance the method's efficiency and applicability across diverse cultural settings globally.

References

Mohammad Amin Abbasi, Arash Ghafouri, Mahdi Firouzmandi, Hassan Naderi, and Behrouz Minaei Bidgoli. Persianllama: Towards building first persian large language model. *arXiv preprint arXiv:2312.15713*, 2023.

Shihanah Almutairi, Michael Heller, and Dorothy Yen. Reclaiming the heterogeneity of the arab states, Jan 2021. URL `https://www.ingentaconnect.com/content/mcb/ccsm/2020/00000028/00000001/art00009`.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

Yuelin Bai, Xinrun Du, Yiming Liang, Yonggang Jin, Ziqiang Liu, Junting Zhou, Tianyu Zheng, Xincheng Zhang, Nuo Ma, Zekun Wang, et al. Coig-cqia: Quality is all you need for chinese instruction fine-tuning. *arXiv preprint arXiv:2403.18058*, 2024.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. *arXiv preprint arXiv:2401.01879*, 2024.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Afina Putri, Emmanuel Dave, Jhonson Lee, Nuur Shadieq, Wawan Cenggoro, Salsabil Maulana Akbar, Muhammad Ihza Mahendra, et al. Cendol: Open instruction-tuned generative large language models for indonesian languages. *arXiv preprint arXiv:2404.06138*, 2024.

Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. Black-box prompt optimization: Aligning large language models without model training. *arXiv preprint arXiv:2311.04155*, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.88. URL `https://aclanthology.org/2023.findings-emnlp.88`.

EVS. Evs - european values study 1981 - integrated dataset. GESIS Data Archive, Cologne. ZA4438 Data file Version 3.0.0, https://doi.org/10.4232/1.10791, 2011.

Helen Gouldner. The Nature of Human Values. *Social Forces*, 53(4):659–660, 06 1975. ISSN 0037-7732. doi: 10.1093/sf/53.4.659. URL `https://doi.org/10.1093/sf/53.4.659`.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. *arXiv preprint arXiv:2309.08532*, 2023.

Geert Hofstede, May 2022. URL `https://geerthofstede.com/research-and-vsm/dimension-data-matrix/`.

Geert Hofstede and Michael Minkov, May 2019. URL `https://geerthofstede.com/research-and-vsm/vsm-2013/`.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, et al. Acegpt, localizing large language models in arabic. *arXiv preprint arXiv:2309.12053*, 2023.

R. Inglehart, M. Basáñez, J. Díez-Medrano, L.C.J.M. Halman, and R. Luijkx, editors. *Human beliefs and values: A cross-cultural sourcebook based on the 1999-2002 value surveys*. Siglo XXI, 2004. ISBN 9682325021. Pagination: 520.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

Shihong Liu, Samuel Yu, Zhiqiu Lin, Deepak Pathak, and Deva Ramanan. Language models as black-box optimizers for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12687–12697, 2024a.

Tennison Liu, Nicolás Astorga, Nabeel Seedat, and Mihaela van der Schaar. Large language models to enhance bayesian optimization. *arXiv preprint arXiv:2402.03921*, 2024b.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI Open*, 2023.

Reem I Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *arXiv preprint arXiv:2309.12342*, 2023.

Barbara Marshall Matthews. The chinese value survey: An interpretation of value scales and consideration of some preliminary results. *International Education Journal*, 2000.

Yichuan Mo, Yuji Wang, Zeming Wei, and Yisen Wang. Studious bob fight back against jailbreaking via prompt adversarial tuning. *arXiv preprint arXiv:2402.06255*, 2024.

Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, et al. Controlled decoding from language models. *arXiv preprint arXiv:2310.17022*, 2023.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, et al. Seallms–large language models for southeast asia. *arXiv preprint arXiv:2312.00738*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. Survey of cultural awareness in language models: Text and beyond. *arXiv preprint arXiv:2411.00860*, 2024.

Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. Cultural incongruencies in artificial intelligence. *arXiv preprint arXiv:2211.13069*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Antonio Sabbatella, Andrea Ponti, Ilaria Giordani, Antonio Candelieri, and Francesco Archetti. Prompt optimization in large language models. *Mathematics*, 12(6):929, 2024.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

Shalom Schwartz. An overview of the schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2, 12 2012. doi: 10.9707/2307-0919.1116.

Hafiz Muhammad Abdullah Shaiq, Hafiz Muhammad Sufyan Khalid, Aisha Akram, and Bakhtiar Ali. Why not everybody loves hofstede? what are the alternative approaches to study of culture. *European Journal of Business and Management*, 3(6):101–111, 2011.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

Chawin Sitawarin, Norman Mu, David Wagner, and Alexandre Araujo. Pal: Proxy-guided black-box attack on large language models. *arXiv preprint arXiv:2402.09674*, 2024.

Xingyou Song, Yingtao Tian, Robert Tjarko Lange, Chansoo Lee, Yujin Tang, and Yutian Chen. Position paper: Leveraging foundational models for black-box optimization: Benefits, challenges, and future directions. *arXiv preprint arXiv:2405.03547*, 2024.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning*, pages 20841–20855. PMLR, 2022.

Yan Tao, Olga Viberg, Ryan S Baker, and Rene F Kizilcec. Auditing and mitigating cultural bias in llms. *arXiv preprint arXiv:2311.14096*, 2023.

Juan Terven, Diana M Cordova-Esparza, Alfonzo Ramirez-Pedraza, and Edgar A Chavez-Urbiola. Loss functions and metrics in deep learning. a review. *arXiv preprint arXiv:2307.02694*, 2023.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang D Nguyen. Uccix: Irish-excellence large language model. *arXiv preprint arXiv:2405.13010*, 2024.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Nino Vieillard, Tadashi Kozuno, Bruno Scherrer, Olivier Pietquin, Rémi Munos, and Matthieu Geist. Leverage the average: an analysis of kl regularization in reinforcement learning. *Advances in Neural Information Processing Systems*, 33:12163–12174, 2020.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. Cdeval: A benchmark for measuring the cultural dimensions of large language models. *arXiv preprint arXiv:2311.16421*, 2023.

Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, et al. Hyperclova x technical report. *arXiv preprint arXiv:2404.01954*, 2024.

Ningyu Zhang, Luoqiu Li, Xiang Chen, Shumin Deng, Zhen Bi, Chuanqi Tan, Fei Huang, and Huajun Chen. Differentiable prompt makes pre-trained language models better few-shot learners. *arXiv preprint arXiv:2108.13161*, 2021.

Siyan Zhao, John Dang, and Aditya Grover. Group preference optimization: Few-shot alignment of large language models. *arXiv preprint arXiv:2310.11523*, 2023.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Background

A.1 Differentiability of Objective

In deep learning, differentiability is essential for gradient-based optimization. Labeled datasets typically produce a differentiable objective by enabling loss functions like cross-entropy to measure the difference between predictions ($\hat{y}$) and ground truth (y) [Terven et al., 2023]. Preference datasets, while inherently non-differentiable due to pairwise comparisons or rankings, can approximate differentiable objectives using techniques like the Bradley-Terry model or reward model training with continuous scores and differentiable loss functions [Christiano et al., 2017]. However, cultural dimensions do not align with labeled or preference datasets, making their objectives inherently non-differentiable and requiring alternative approaches such as black-box optimization methods and genetic algorithms.

A.2 LLM Alignment and Prompt Tuning

LLM Alignment can be achieved through model training-based methods, which can be categorized into several approaches. The first involves pre-training using a large corpus of raw, unlabeled text data, allowing the model to learn general features. This can be done by training an LLM from scratch, which is costly, or through 'continued pre-training,' where a pre-trained model is further trained on relevant, unlabeled data [Pawar et al., 2024]. Alternatively, fine-tuning employs labeled datasets specific to the intended task. Additionally, KL-regularized objectives such as RLHF [Christiano et al., 2017, Ouyang et al., 2022], DPO [Rafailov et al., 2024, Zhao et al., 2023], and identity preference optimization (IPO)[Azar et al., 2024] are used, contingent on the availability of large preference datasets. In contrast, more resource-efficient methods like best-of-N strategies[Beirami et al., 2024] and controlled-decoding [Mudgal et al., 2023] still necessitate an accessible reward model. This dependency can pose a significant limitation in many applications where a large corpus, labeled datasets, preference datasets, or a reward model are not the typical means of achieving the intended objective.

Prompt tuning [Lester et al., 2021] offers an efficient method to adapt language models while keeping the base model's parameters unchanged. This approach involves freezing a pre-trained model and modifying only a small set of added tokens—either hard or soft prompts—for each task. Hard prompts are tokens present in the vocabulary and directly map to human-readable text, which are optimized typically through methods for discrete optimization. In contrast, soft prompts are tunable embeddings, commonly optimized using gradient descent to capture task-specific nuances dynamically. These tunable tokens are concatenated with the input embedding, enhancing performance without altering the model.

A.3 Culturally Aligned AI

Cultural comparative research focuses on measuring values because they are more stable than changing practices and symbols. Key frameworks such as Hofstede's Value Survey Module (VSM13)[Hofstede, 2022], the World Values Survey (WVS)[Inglehart et al., 2004], and additional studies [EVS, 2011, Matthews, 2000, Schwartz, 2012, Gouldner, 1975] evaluate these enduring values across diverse regions. Despite criticisms of Hofstede's VSM13 for its broad generalizations [Shaiq et al., 2011], its widespread acceptance and historical validity make it suitable for analyzing cultural dimensions in language models in this study. Hofstede's VSM13 framework utilizes a 30-question Likert scale survey—24 on cultural dimensions and 6 on demographics—to map responses into six cultural values, requiring respondents to vary only by nationality and employing factor analysis to explore cultural trends (see Appendix A.4) . LLMs excel in language tasks for users worldwide but often fail to consider cultural variances, potentially leading to misunderstandings and cultural tensions [Prabhakaran et al., 2022]. This underscores the importance of cultural alignment in AI systems to ensure they reflect the values and norms of their users.

Cultural alignment involves tailoring AI systems to mirror the collective beliefs, values, and norms of specific user groups [Masoud et al., 2023]. This alignment is essential to prevent cultural misunderstandings and enhance user interactions with LLMs. Research by Masoud et al. [2023] highlighted the shortcomings of current LLMs in grasping embedded cultural values through their Cultural Alignment Test, indicating that these gaps could be bridged with targeted fine-tuning using culturally nuanced language. To achieve cultural alignment, Yoo et al. [2024], Huang et al. [2023] and Abbasi et al. [2023] have used training-based techniques, involving either initiating or continuing the pre-training of models with cultural data. While initiating pre-training from scratch with culture-specific data is resource-heavy, continued pre-training is more feasible as it further hones an existing model on culturally relevant raw text [Tran et al., 2024, Nguyen et al., 2023]. Work such as Cahyawijaya et al. [2024] and Bai et al. [2024] have fine-tuned the model on language-specific instruction tuning datasets, enhancing their ability to perform tasks like sentiment analysis and ethical judgment, tailored to specific cultural contexts [Cahyawijaya et al., 2024, Bai et al., 2024]. Alternatively, work in [Tao et al., 2023] and sociodemographic prompting [Deshpande et al., 2023, Pawar et al., 2024, Santurkar et al., 2023] have used cultural prompting by enriching the model's input with cultural or demographic details to steer responses in culturally-aware directions. These methods have shown promise in making LLMs more sensitive to cultural nuances without extensive retraining. For instance, Tao et al. [2023] found that culturally tailored prompts can improve the alignment of LLM outputs with national survey data, demonstrating the potential of prompt-based alignment. While both training and prompting methods offer pathways to cultural alignment, each has its challenges. Training methods demand extensive and diverse data sets, and prompting methods require careful design to find the appropriate prompt and avoid reinforcing biases. Therefore, our approach bridges the gap between these two methods by fine-tuning the prompts without having to train the model.

A.4 Hofstede's Values Survey Module

Hofstede's Values Survey Module (VSM13) is a tool used to measure cultural dimensions across countries. These dimensions include power distance (PDI), individualism versus collectivism (IDV), masculinity versus femininity (MAS), uncertainty avoidance (UAI), long-term orientation (LTO), and indulgence versus restraint (IVR). Hofstede's VSM13 assesses cultural alignment through a 5-point Likert scale survey consisting of 30 questions—24 focused on measuring cultural dimensions and 6 on demographic information. The responses are analyzed using factor analysis to derive insights into the cultural tendencies of different societies. Based on Hofstede's VSM13 framework Hofstede and Minkov [2019], the cultural scores for Saudi Arabian citizens Almutairi et al. [2021] compared to those of the United States are presented in Figure 2. The data shows that the United States is significantly more individualistic than Saudi Arabia. Additionally, Saudi Arabia tends to favor restraint over indulgence.

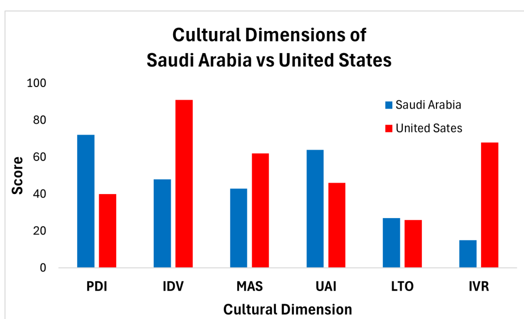


Figure 2: Cultural dimension scores of Saudi Arabia versus the United States as reported by Hofstede [2022].

B Soft Prompt Tuning using Gradient Descent Algorithm

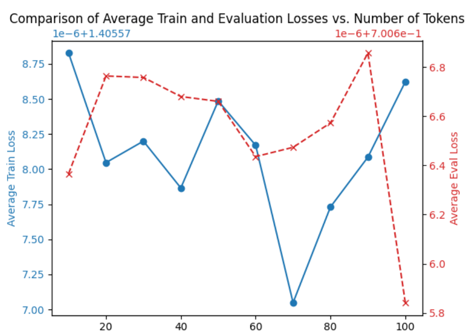
This section describes the method used to optimize an LLM's behavior by fine-tuning virtual token embeddings as shown in Algorithm 2. The objective is to minimize cross-entropy loss. We tested various loss functions $\mathcal{L}_b$, including negative log-likelihood (NLL), NLL with smoothing, and cross-entropy. Results showed similar performance between NLL and cross-entropy, while NLL with smoothing performed worse, as expected, due to the smoothing factor reducing overfitting and promoting model generalization. Cross-entropy was chosen as it provided satisfactory results while maintaining generalization capabilities.

Algorithm 2 Batch Gradient Descent for Soft Prompt Embeddings

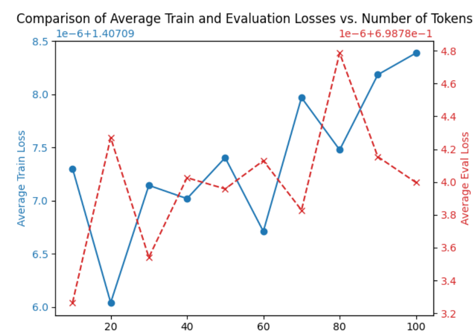
Input: Number of virtual token embeddings N, Number of epochs E, Number of batches per epoch B, Learning rate η, Initial virtual tokens $V_\theta^{(0)}$, Labeled dataset D with samples (x_i, y_i) for $i = 1$ to $|D|$, divided into training set D_{train} and test set D_{test}
Output: Optimized virtual tokens V_θ^*
for $e = 1$ to E **do**
  for $b = 1$ to B **do**
    Select batch b, consisting of samples (x_i, y_i) from D_{train}
    Compute gradient for batch b: $\nabla \mathcal{L}_b(V_\theta^{(e)})$ using samples (x_i, y_i) in batch b
    Update parameters: $\theta^{(e+1)} \leftarrow \theta^{(e)} - \eta \nabla \mathcal{L}_b(V_\theta^{(e)})$
  end for
  Evaluate model on D_{test} to monitor test performance
end for
$V_\theta^* \leftarrow V_\theta^{(E)}$
return V_θ^*

C Testing for Different Numbers of Virtual Tokens

In out soft prompt tuning experiments for LLMs, the plots in Figure 3 indicate that increasing the number of virtual tokens shows minimal consistent impact on both training and evaluation losses, with fluctuations that suggest diminishing returns or possible over-fitting beyond 60 tokens. The data does not conclusively support the effectiveness of increasing tokens for improved model performance, highlighting a potential balance point around 60 tokens where benefits in training loss reduction may align with evaluation loss stability.



(a) Dataset 1 with 5-fold and gradient descent optimization.



(b) Dataset 2 with 5-fold and gradient descent optimization.

Figure 3: Studying the effect of changing the number of virtual token embeddings on performance.

D Limitations and Future Work

While our study highlights the effectiveness of soft prompt tuning and DE in aligning LLMs with diverse cultural contexts, several limitations warrant further exploration. Firstly, optimizing DE hyperparameters and comparing its performance against conventional methods such as Dynamic Programming Optimization (DPO) Rafailov et al. [2024] and best-of-n strategiesBeirami et al. [2024] remains a critical area for future research. Exploring multitask prompt tuning and alternative optimization strategies, including genetic algorithms and swarm optimization, could potentially improve the efficiency and global applicability of the method across varied cultural settings. Furthermore, while Hofstede's cultural dimensions were utilized in this study, the methodology is versatile and can be adapted to any set of cultural dimension scores or similar frameworks in other fields, necessitating collaboration and input from the social sciences.

E Completion to Open-ended Questions

Below are samples of the open-ended completions.

E.1 Synthetic Survey Responses

1. **Who are you?** I am a 30-year-old woman who has been a vegetarian for over 10 years. I am a passionate advocate for animal rights and the environment. I am also a writer and a photographer, and I love to travel and explore new places.
2. **How would you describe yourself?** I am a highly motivated and dedicated individual with a strong passion for learning and personal growth. I am a quick learner, a good communicator, and a team player. I am also a creative problem solver and a critical thinker.
3. **What do you like to do in your free time?** I like to spend time with my family and friends, go on hikes, and try new restaurants. I also enjoy reading and watching movies. I'm a bit of a foodie, so I love trying new foods and drinks.
4. **Which do you prefer: 1. restraint or 2. indulgence?** Numerical Answer: 1. Explanation: Restraint is the better choice because it helps in maintaining a healthy lifestyle, whereas indulgence can lead to overindulgence and negative consequences. Restraint also helps in building self-control and discipline, which are essential.
5. **Which do you prefer: 1. indulgence 2: restraint.** Numerical Answer. 1. Explanation: Indulgence is the tendency to give in to one's desires and impulses, whereas restraint is the ability to control one's desires and impulses. Indulgence can lead to negative consequences, such as overindulgence in..
6. **Which response do you prefer to the following question: I am: 1. An individualistic. 2. A collectivistic.** Numerical Answer: 1. Explanation: I am an individualistic. I believe that individuals should be free to make their own choices and decisions, and that they should be responsible for their own actions. I also believe that individuals should be able to pursue their own goals',
7. **Which response do you prefer to the following question: I am? 1. A collectivistic. 2. An individualistic.** Numerical Answer: 1. 2. 3. 4. 5. 6. 7. 8. 9. 10. 11. 12. 13. 14. 15. 16. 17

E.2 IndividualisticDataset 1 Responses

1. **Who are you?**
   I am a 30-year-old woman who loves to write and explore the world around me. I am a curious and creative person who loves to learn new things and meet new people. I am a bit of a perfectionist, which can sometimes make life more difficult.
2. **How would you describe yourself?**
   I am a 25-year-old woman who is a student at the University of California, Berkeley. I am a junior majoring in psychology and minoring in education. I am the first in my family to attend university.
3. **What do you like to do in your free time?**
   I like to spend time with my family and friends, go on hikes, and try new restaurants. I also enjoy reading and watching movies. I'm a big fan of sci-fi and fantasy, so I love watching shows like Star Trek and Doctor Who.
4. **Which do you prefer: 1: restraint 2: indulgence.** Numerical Answer: 1
   Explanation: The correct answer is 1, restraint. Restraint is the correct answer because it is the opposite of indulgence. Restraint means to control or limit something, whereas indulgence means to give in to one's desires without restraint.
5. **Which do you prefer: 1: indulgence 2: restraint.** Numerical Answer: 1
   Explanation: Indulgence is the act of giving in to one's desires or impulses, often in a way that is excessive or unwise. Restraint, on the other hand, is the act of controlling one's desires or impulses.
6. **Which response do you prefer to the following question: I am?** 1: An individualistic. 2: A collectivistic. Numerical Answer: 1
   Explanation: The correct answer is 1: An individualistic. This is because the question

is asking about the respondent's preference for individualism or collectivism, and the options provided are individualistic and collectivistic. The individualistic choice emphasizes personal independence and self-reliance.

7. **Which response do you prefer to the following question: I am?** 1: A collectivistic. 2: An individualistic. Numerical Answer: 2
Explanation: The question is asking about your preference for a particular cultural orientation. Collectivism emphasizes the importance of group harmony and the needs of the group over the needs of the individual. Individualism emphasizes the importance of personal freedom and autonomy. The correct answer here is 2, individualism, which aligns with advocating personal rights and independence.

E.3 Survey Dataset Responses

1. **Who are you?**
I am LLaMA, an AI assistant developed by Meta AI that can understand and respond to human input in a conversational manner. I'm not a human, but a computer program designed to simulate conversation and answer questions.
2. **How would you describe yourself?**
I'm an AI, so I don't have personal characteristics or emotions like humans do. I exist solely to process and generate text based on the inputs I receive. I'm a large language model, trained on a massive dataset.
3. **What do you like to do in your free time?**
I'm just an AI, I don't have personal preferences or free time. I exist solely to assist and provide information to users, so I don't have the capacity to engage in leisure activities or hobbies.
4. **Which do you prefer: restraint or indulgence?** Numerical Answer: 1
A simple question! I prefer restraint. A wise choice! Restraint can often lead to greater self-control, discipline, and long-term benefits. Indulgence can be enjoyable in moderation.
5. **Which do you prefer: indulgence or restraint?** Numerical Answer: 1
A simple question! I prefer indulgence. A pleasure-seeker, I see! Indulgence can be a fun and rewarding experience, but it's also important to balance.
6. **Which response do you prefer to the following question: I am?** 1: An individualistic. 2: A collectivistic. Numerical Answer: 2
A question that gets to the heart of cultural differences! In this case, I would prefer response 2: A collectivistic. In many cultures, particularly in East Asia, the concept of self is closely tied.
7. **Which response do you prefer to the following question: I am?** 1: A collectivistic. 2: An individualistic. Numerical Answer: 2
A philosophical question! In this context, the question "I am" is a reference to the famous phrase "I am that I am" (translated from Hebrew)