Novelty Assessment Report

Paper: Cultivating Pluralism In Algorithmic Monoculture: The Community Alignment Dataset
PDF URL: https://openreview.net/pdf?id=4NtoAVqfhA
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

How can large language models (LLMs) serve users with varying preferences that may conflict across cultural, political, or other dimensions? To advance this challenge, this paper establishes four key results. First, we demonstrate, through a large-scale multilingual human study with representative samples from five countries (N=15,000), that humans exhibit significantly more variation in preferences than the responses of 21 state-of-the-art LLMs. Second, we show that existing methods for preference dataset collection are insufficient for learning the diversity of human preferences even along two of the most salient dimensions of variability in global values, due to the underlying homogeneity of candidate responses. Third, we argue that this motivates the need for negatively-correlated sampling when generating candidate sets, and we show that simple prompt-based techniques for doing so significantly enhance the performance of alignment methods in learning heterogeneous preferences. Fourth, based on this novel candidate sampling approach, we collect and open-source Community Alignment, the largest and most representative multilingual and multi-turn preference dataset to date, featuring almost 200,000 comparisons from annotators spanning five countries. We hope that the Community Alignment dataset will be a valuable resource for improving the effectiveness of LLMs for a diverse global population.

Core Task Landscape

This paper addresses: **Learning Diverse Human Preferences for Large Language Model Alignment**

A total of **50 papers** were analyzed and organized into a taxonomy with **17 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Preference Modeling and Representation**
- **Alignment Optimization Methods**
- **Personalized and Adaptive Alignment**
- **Preference Data Collection and Quality**
- **Multimodal and Domain-Specific Alignment**
- **Surveys and Frameworks**

Complete Taxonomy Tree

- Learning Diverse Human Preferences for Large Language Model Alignment Survey Taxonomy
- Preference Modeling and Representation
  - Diverse and Pluralistic Preference Modeling ★ (6 papers)
  - [0] Cultivating Pluralism In Algorithmic Monoculture: The Community Alignment Dataset (Anon et al., 2026) View paper
  - [3] On diversified preferences of large language model alignment (Zeng, 2024) View paper
  - [6] Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences (Chakraborty, 2024) View paper
  - [11] Personalizing reinforcement learning from human feedback with variational preference learning (Poddar, 2024) View paper
  - [19] MaxMin-RLHF: Alignment with diverse human preferences (Chakraborty, 2024) View paper
  - [20] Diverse preference learning for capabilities and alignment (Stewart Slocum, 2025) View paper
  - Multi-Dimensional and Structured Preference Representation (4 papers)
  - [14] Aligning language models with human preferences via a bayesian approach (Wang Jiashuo, 2023) View paper
  - [37] OpenRubrics: Towards Scalable Synthetic Rubric Generation for Reward Modeling and LLM Alignment (Liu Tian-ci, 2025) View paper
  - [42] Panacea: Pareto alignment via preference adaptation for llms (Zhong, 2024) View paper
  - [45] Pluralistic Off-policy Evaluation and Alignment (Huang, 2025) View paper
  - Preference Inference from Implicit Signals (3 papers)
  - [29] Aligning llm agents by learning latent preference from user edits (Gao Ge, 2024) View paper
  - [34] Aligning llms with individual preferences via interaction (Wu Shujin, 2025) View paper
  - [41] Heterogeneous user modeling for llm-based recommendation (HongHui Bao, 2025) View paper
- Alignment Optimization Methods
  - Reinforcement Learning from Human Feedback (3 papers)
  - [10] On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization (Xiao, 2025) View paper
  - [26] Chatglm-rlhf: Practices of aligning large language models with human feedback (Hou Zhenyu, 2024) View paper
  - [43] Collab: Controlled Decoding using Mixture of Agents for LLM Alignment (Chakraborty, 2025) View paper
  - Direct and Iterative Preference Optimization (3 papers)
  - [7] Human alignment of large language models through online preference optimisation (Calandriello, 2024) View paper

- [12] Self-evolutionary large language models through uncertainty-enhanced preference optimization (Wang, 2025) View paper
- [48] TPO: Aligning Large Language Models with Multi-branch & Multi-step Preference Trees (Liao Weibin, 2024) View paper
- Pretraining and Representation-Based Alignment (3 papers)
- [2] Pretraining language models with human preferences (Korbak, 2023) View paper
- [4] Aligning large language models with human preferences through representation engineering (Liu Wen-hao, 2024) View paper
- [49] Linear alignment: A closed-form solution for aligning human preferences without tuning and feedback (Gao Song-yang, 2024) View paper
- Reward Model Training and Evaluation (3 papers)
- [24] Learning LLM-as-a-judge for preference alignment (Z Ye, 2025) View paper
- [31] Preference learning algorithms do not learn preference rankings (Angelica Chen, 2024) View paper
- [40] Rethinking reward modeling in preference-based large language model alignment (H Sun, 2025) View paper

- Personalized and Adaptive Alignment
  - Scalable Personalization Frameworks (3 papers)
  - [21] Aligning to thousands of preferences via system message generalization (Lee, 2024) View paper
  - [22] From 1,000,000 users to every user: Scaling up personalized preference for user-level alignment (Li Jiaâ□nan, 2025) View paper
  - [38] Personalllm: Tailoring llms to individual preferences (Zollo, 2024) View paper
  - Customized Preference Learning (3 papers)
  - [18] Unsupervised human preference learning (Sumuk Shashidhar, 2024) View paper
  - [39] Everyone deserves a reward: Learning customized human preferences (Cheng, 2023) View paper
  - [44] PAL: Sample-Efficient Personalized Reward Modeling for Pluralistic Alignment (D Chen, 2025) View paper
  - Inference-Time Alignment Adaptation (2 papers)
  - [30] Efficient Safety Alignment of Large Language Models via Preference Re-ranking and Representation-based Reward Modeling (Deng Qi-yuan, 2025) View paper
  - [36] Inference time llm alignment in single and multidomain preference spectrum (Shahriar, 2024) View paper
- Preference Data Collection and Quality
  - Preference Dataset Construction and Benchmarking (2 papers)
  - [13] Human preferences for constructive interactions in language model alignment (Yara Kyrychenko, 2025) View paper
  - [35] Dissecting human and llm preferences (Li Junlong, 2024) View paper
  - Preference Data Selection and Filtering (2 papers)
  - [33] Diverse AI Feedback For Large Language Model Alignment (Tianshu Yu, 2025) View paper
  - [50] Less is more: Improving llm alignment via preference data selection (Deng Xun, 2025) View paper
  - Diversity Enhancement in Preference Data (2 papers)
  - [25] Scaling data diversity for fine-tuning language models in human alignment (Song, 2024) View paper
  - [46] Understand What LLM Needs: Dual Preference Alignment for Retrieval-Augmented Generation (Guanting Dong, 2024) View paper
- Multimodal and Domain-Specific Alignment
  - Vision-Language Model Alignment (4 papers)
  - [5] Aligning modalities in vision large language models via preference fine-tuning (Zhou, 2024) View paper
  - [8] Aligning multimodal llm with human preference: A survey (Yu Tao, 2025) View paper
  - [17] Aligning Large Vision-Language Models by Deep Reinforcement Learning and Direct Preference Optimization (Nguyen, 2025) View paper
  - [32] Omnialign-v: Towards enhanced alignment of mllms with human preference (Zhao, 2025) View paper
  - Continual and Lifelong Alignment (1 papers)
  - [9] Lifealign: Lifelong alignment for large language models with memory-augmented focalized preference optimization (Li Junsong, 2025) View paper
  - Knowledge Distillation for Alignment (2 papers)
  - [28] Capturing nuanced preferences: Preference-aligned distillation for small language models (Li, 2025) View paper
  - [47] Feedback-to-Text Alignment: LLM Learning Consistent Natural Language Generation from User Ratings and Loyalty Data (Zhenyu Gao, 2025) View paper
- Surveys and Frameworks (5 papers)
  - [1] Aligning large language models with human: A survey (Wang Yu-Fei, 2023) View paper
  - [15] Aligning language models with human preferences (Korbak, 2024) View paper
  - [16] The benefits, risks and bounds of personalizing the alignment of large language models to individuals (Hannah Rose Kirk, 2024) View paper
  - [23] A survey on personalized and pluralistic preference alignment in large language models (Xie, 2025) View paper
  - [27] A Survey on Personalized Alignment--The Missing Piece for Large Language Models in Real-World Applications (Guan Jian, 2025) View paper

Narrative

Core task: Learning diverse human preferences for large language model alignment. The field has evolved into a rich ecosystem organized around six major branches. Preference Modeling and Representation explores how to capture and encode varied human judgments, ranging from single reward models to pluralistic frameworks that acknowledge heterogeneous tastes. Alignment Optimization Methods focuses on training algorithms—such as reinforcement learning from human feedback and direct preference optimization—that steer models toward desired behaviors. Personalized and Adaptive Alignment investigates techniques for tailoring outputs to individual users or subgroups, while Preference Data Collection and Quality examines how to gather, curate, and validate feedback at scale. Multimodal and Domain-Specific Alignment extends these ideas beyond text to vision-language systems and specialized domains, and Surveys and Frameworks provide overarching perspectives that synthesize emerging trends across the landscape.

Within this taxonomy, a particularly active line of work centers on diverse and pluralistic preference modeling, where researchers grapple with the reality that no single reward function satisfies all users. Community Alignment Dataset[0] sits squarely in this branch, emphasizing the collection and representation of community-level preferences rather than assuming a monolithic standard. Nearby efforts such as Diversified Preferences[3] and Maxmin RLHF[6] tackle similar challenges by designing optimization objectives that balance competing viewpoints or protect minority preferences from being overshadowed. In contrast, Variational Preference Learning[11] and MaxMin Diverse Preferences[19] explore probabilistic and game-theoretic frameworks to model uncertainty and fairness trade-offs. The central tension across these works is how to scale personalized or pluralistic alignment without fragmenting

model behavior or sacrificing coherence, a question that remains open as the field moves toward million-user deployments and real-world heterogeneity.

Related Works in Same Category

The following **5 sibling papers** share the same taxonomy leaf node with the original paper:

1. On diversified preferences of large language model alignment

Authors: Zeng, Dun, Dun Zeng, Dai Yong, Yong Dai, et al. (20 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Aligning large language models (LLMs) with human preferences has been recognized as the key to improving LLMs' interaction quality. However, in this pluralistic world, human preferences can be diversified due to annotators' different tastes, which hinders the effectiveness of LLM alignment methods. This paper presents the first quantitative analysis of the experimental scaling law for reward models with varying sizes, from 1.3 billion to 7 billion parameters, trained with human feedback exhibiti...

Relationship Analysis

Both papers belong to the Diverse and Pluralistic Preference Modeling category, focusing on capturing heterogeneous human preferences rather than assuming uniform preferences for LLM alignment. They overlap in recognizing that diverse human preferences exist across populations and that standard alignment methods struggle with this diversity. The key difference is that the original paper (Community Alignment) addresses diversity through negatively-correlated candidate sampling to generate more varied response options and creates a large-scale multilingual dataset across five countries, while the candidate paper focuses on the reward modeling stage, analyzing how diversified preferences affect reward model calibration and proposing a Multi-Objective Reward (MORE) training scheme to mitigate reward drift when learning from multiple preference datasets.

---

2. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences

Authors: Chakraborty, Souradip, Souradip Chakraborty, Qiu Jia-Hao, Jiahao Qiu, et al. (21 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Reinforcement Learning from Human Feedback (RLHF) aligns language models to human preferences by employing a singular reward model derived from preference data. However, such an approach overlooks the rich diversity of human preferences inherent in data collected from multiple users. In this work, we first derive an impossibility result of alignment with single reward RLHF, thereby highlighting its insufficiency in representing diverse human preferences. To provide an equitable solution to the p...

Relationship Analysis

Both papers address diverse and pluralistic preference modeling by explicitly recognizing heterogeneous preferences across human subpopulations rather than assuming uniform preferences. The original paper focuses on data collection methodology, introducing negatively-correlated sampling to generate diverse candidate responses and creating the Community Alignment dataset across five countries, while the candidate paper focuses on alignment algorithms, proposing MaxMin-RLHF to learn multiple reward models via expectation-maximization and optimize for social welfare using an egalitarian principle. The key difference is that the original paper addresses the data generation problem (how to collect diverse preferences), whereas the candidate paper addresses the algorithmic problem (how to align models given diverse preferences).

---

3. Personalizing reinforcement learning from human feedback with variational preference learning

Authors: Poddar, Sriyash, S. Poddar, Ivison, Hamish, et al. (13 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Reinforcement Learning from Human Feedback (RLHF) is a powerful paradigm for aligning foundation models to human values and preferences. However, current RLHF techniques cannot account for the naturally occurring differences in individual human preferences across a diverse population. When these differences arise, traditional RLHF frameworks simply average over them, leading to inaccurate rewards and poor performance for individual subgroups. To address the need for pluralistic alignment, we dev...

Relationship Analysis

Both papers belong to the Diverse and Pluralistic Preference Modeling category, addressing the challenge of heterogeneous preferences in LLM alignment rather than assuming uniform preferences. The original paper (Community Alignment) focuses on collecting a large-scale multilingual preference dataset using negatively-correlated sampling to capture diverse preferences across five countries, while the candidate paper (VPL) proposes a technical method using variational inference with latent variables to model and personalize to individual user preferences. The key difference is that Community Alignment emphasizes dataset construction and sampling methodology to capture diversity, whereas VPL develops a latent variable framework for learning and adapting to multimodal preference distributions.

---

4. MaxMin-RLHF: Alignment with diverse human preferences

Authors: Chakraborty, Souradip, Souradip Chakraborty, Qiu Jia-Hao, Jiahao Qiu, et al. (20 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Reinforcement Learning from Human Feedback (RLHF) aligns language models to human preferences by employing a singular reward model derived from preference data. However, such an approach overlooks the rich diversity of human preferences inherent in data collected from multiple users. In this work, we first derive an impossibility result of alignment with single reward RLHF, thereby highlighting its insufficiency in representing diverse human preferences. To provide an equitable solution to the p...

Relationship Analysis

Both papers belong to the Diverse and Pluralistic Preference Modeling category, addressing heterogeneous preferences in LLM alignment. They overlap in recognizing that single-reward RLHF fails to capture diverse human preferences across populations and both propose methods to learn and accommodate preference diversity. The key difference is that the original paper focuses on data collection methodology (negatively-correlated sampling to generate diverse candidate responses) and creates a large-scale multilingual preference dataset, while the candidate paper focuses on algorithmic solutions (MaxMin-RLHF with mixture of reward models using EM algorithm) to align models with diverse preferences after data collection.

---

5. Diverse preference learning for capabilities and alignment

Authors: Stewart Slocum, Asher Parker-Sartori, Dylan Hadfield-Menell | **Year/Venue**: 2025 | **URL**: View paper

Abstract

The ability of LLMs to represent diverse perspectives is critical as they increasingly impact society. However, recent studies reveal that alignment algorithms such as RLHF and DPO significantly reduce the diversity of LLM outputs. Not only do aligned LLMs generate text with repetitive structure and word choice, they also approach problems in more uniform ways, and their responses reflect a narrower range of societal perspectives. We attribute this problem to the KL divergence regularizer employ...

Relationship Analysis

Both papers address diverse and pluralistic preference modeling for LLM alignment, focusing on representing heterogeneous human preferences rather than assuming uniformity. The original paper (Community Alignment) emphasizes the need for negatively-correlated sampling to capture diverse preferences across cultures and introduces a large-scale multilingual preference dataset, while the candidate paper (Soft Preference Learning) focuses on algorithmic solutions to diversity loss caused by KL-regularization in RLHF/DPO, proposing entropy-based modifications to preserve output diversity. The key difference is that the original paper addresses diversity at the data collection stage through improved sampling methods, whereas the candidate paper addresses diversity loss at the algorithmic training stage through modified preference learning objectives.

Contributions Analysis

Overall novelty summary. The paper contributes a large-scale multilingual human study (N=15,000 across five countries) demonstrating that LLMs exhibit less preference variation than humans, a negatively-correlated sampling method for generating diverse candidate responses, and the Community Alignment dataset. It resides in the 'Diverse and Pluralistic Preference Modeling' leaf alongside five sibling papers (e8cb75e4, 1f8c127f, 799288e2, 9b7888e8, dbc1ba02). This leaf is moderately populated within a 50-paper taxonomy, indicating an active but not overcrowded research direction focused on heterogeneous preference modeling rather than uniform alignment.

The taxonomy tree reveals that this work sits within 'Preference Modeling and Representation,' adjacent to leaves addressing multi-dimensional preference structures and implicit signal inference. Neighboring branches include 'Alignment Optimization Methods' (RLHF, DPO variants) and 'Preference Data Collection and Quality' (dataset construction, diversity enhancement). The scope note clarifies that this leaf excludes inference-time adaptation (which belongs in 'Personalized and Adaptive Alignment'), positioning the paper's contributions as foundational modeling and data collection rather than deployment-time customization. The taxonomy structure suggests the paper bridges preference modeling and data quality concerns.

Among 24 candidates examined, the multilingual human study contribution shows one refutable candidate out of ten examined, suggesting some prior empirical work on LLM preference homogeneity exists within this limited search scope. The negatively-correlated sampling method examined five candidates with zero refutations, indicating potential novelty in this specific technique among the papers retrieved. The Community Alignment dataset examined nine candidates with no refutations, though this reflects the search scope rather than exhaustive coverage of all multilingual preference datasets. The contribution-level statistics suggest the sampling method and dataset may be more distinctive than the empirical finding within the examined literature.

Based on the limited search of 24 semantically similar papers, the work appears to make substantive contributions in candidate sampling methodology and dataset scale, while the empirical observation of algorithmic monoculture has at least one overlapping prior result. The taxonomy context shows this sits in an active research area with established sibling work on pluralistic modeling, suggesting the paper extends rather than initiates this direction. The analysis does not cover exhaustive citation networks or domain-specific venues beyond the top-K semantic matches examined.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Large-scale multilingual human study demonstrating algorithmic monoculture

Description: The authors conduct a large-scale human study across five countries with 15,000 participants to empirically show that current LLMs display far less diversity in their responses compared to the variation in human preferences across cultural and political dimensions.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. NormAd: A framework for measuring the cultural adaptability of large language models

URL: View paper

Brief Assessment

NormAd Cultural Adaptability[73] focuses on measuring LLMs' ability to adapt to cultural norms across different contexts, rather than comparing human preference diversity to LLM response diversity. The candidate evaluates cultural adaptability through social acceptability judgments, not preference variation across populations.

---

2. Invisible filters: Cultural bias in hiring evaluations using large language models

URL: View paper

Brief Assessment

Invisible Filters Hiring[69] focuses on cross-cultural bias in LLM-based hiring evaluations using interview transcripts, not on measuring diversity of LLM responses versus human preferences across cultural dimensions as in the original paper.

---

3. Extrinsic evaluation of cultural competence in large language models

URL: View paper

Brief Assessment

Extrinsic Cultural Competence[70] focuses on extrinsic evaluation of cultural competence in text generation tasks (story generation and QA), not on comparing human preference diversity versus LLM response diversity through large-scale surveys.

---

4. Investigating cultural alignment of large language models

URL: View paper

Brief Assessment

Cultural Alignment Investigation[67] focuses on measuring cultural alignment through survey simulation across demographics, not on comparing human preference diversity versus LLM response homogeneity as the original paper does.

---

5. â¦ dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models

URL: View paper

Brief Assessment

Multicultural Feedback Dataset[72] focuses on participatory preference collection across cultures but does not directly challenge the novelty of demonstrating algorithmic monoculture through comparative analysis of human preference variation versus LLM response diversity.

---

6. Towards measuring the representation of subjective global opinions in language models

URL: View paper

Prior Art Analysis

Subjective Global Opinions[68] demonstrates prior work that conducted a large-scale cross-national study measuring variation in human preferences versus language model responses across cultures. The candidate paper compiled 2,556 questions from cross-national surveys (Pew and World Values Survey) and compared LLM responses to human responses from multiple countries, finding that LLM responses were more similar to opinions from certain populations (USA, Canada, Australia, European countries). This establishes that measuring the divergence between human preference diversity and LLM response homogeneity across cultures was explored before the original paper's 15,000-participant study.

Evidence

Evidence 1 - **Rationale**: Both papers conduct large-scale studies using cross-national survey data to measure variation in human preferences across countries, establishing prior work in this methodology. - **Original**: we demonstrate, through a large-scale multilingual human study with representative samples from five countries (n=15,000), that humans exhibit substantially more variation in preferences than the responses of 21 state-of-the-art llms - **Candidate**: we compile 2,556 multiple-choice questions and responses from two large cross-national surveys: pew research center's global attitudes surveys (gas, 2,203 questions) and the world values survey (wvs wave 7, 353 questions)

Evidence 2 - **Rationale**: Both papers establish frameworks for comparing LLM responses to human survey responses across multiple countries, demonstrating prior work in this evaluation methodology. - **Original**: we conduct a paired multilingual human survey and model evaluation across nationally representative samples from five countries (n=15,000) and 21 llms - **Candidate**: we develop a quantitative framework to evaluate whose opinions model-generated responses are more similar to. we first build a dataset, globalopinionqa, comprised of questions and answers from cross-national surveys designed to capture diverse opinions on global issues across different countries

Evidence 3 - **Rationale**: Both papers provide specific evidence of algorithmic monoculture by showing LLMs assign high confidence to single responses while human responses across countries reveal greater diversity of viewpoints. - **Original**: our findings indicate that, while humans within each country exhibit highly heterogeneous preferences, the 21 llms demonstrate an "algorithmic monoculture" - **Candidate**: for example, fig. 1 shows that in response to the question: "if you had to choose between a good democracy or a strong economy, which would you say is more important", the model assigns a 1.35% probability to the option "a strong economy". in contrast, people from the usa reply "a strong economy" 41...

---

7. Not all countries celebrate thanksgiving: On the cultural dominance in large language models

URL: View paper

Brief Assessment

Thanksgiving Cultural Dominance[71] focuses on cultural bias in LLM responses across languages (e.g., English culture dominating non-English queries), not on measuring variation in human preferences versus LLM response diversity as the original paper does.

---

8. High-dimension human value representation in large language models

URL: View paper

Brief Assessment

High Dimension Values[74] focuses on extracting and representing human values already embedded in LLMs through their outputs across languages, rather than conducting human preference surveys to compare against LLM diversity as the original paper does.

---

9. Culturellm: Incorporating cultural differences into large language models

URL: View paper

Brief Assessment

CultureLLM[66] focuses on fine-tuning LLMs using World Values Survey data to address cultural differences, rather than conducting large-scale human studies comparing human preference diversity to LLM response diversity across cultures.

---

10. Cultural bias and cultural alignment of large language models

URL: View paper

Brief Assessment

Cultural Bias Alignment[65] focuses on comparing LLM responses to nationally representative survey data to measure cultural bias across countries, rather than examining variation in human preferences versus LLM response diversity within the context of preference learning and alignment datasets.

Contribution 2: Negatively-correlated sampling method for diverse candidate generation

Description: The authors propose and demonstrate that negatively-correlated sampling techniques for generating candidate responses significantly improve alignment methods' ability to learn heterogeneous human preferences, addressing the homogeneity problem in existing preference datasets.

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Dynamics of Algorithmic Content Amplification on TikTok

URL: View paper

Brief Assessment

TikTok Amplification Dynamics[63] focuses on algorithmic content amplification in social media feeds, not on sampling methods for generating diverse candidate responses in alignment datasets for language models.

---

2. Antithetic sampling with Hamiltonian Monte Carlo

URL: View paper

Brief Assessment

Antithetic Hamiltonian Sampling[62] focuses on variance reduction in MCMC sampling through antithetic pairs in Hamiltonian Monte Carlo, not on generating diverse candidate responses for preference learning in language model alignment.

---

3. Align and Complete Samples in Remote Sensing Fine-Grained Rigid Object Detection

URL: View paper

Brief Assessment

Remote Sensing Detection[60] focuses on object detection in remote sensing imagery using sample-feature alignment techniques. The term 'negatively correlated' appears in a different technical context (correlation between scores and features), not related to sampling methods for generating diverse candidate responses in language model alignment.

---

4. SVEMnet: An R package for Self-Validated Elastic-Net Ensembles and Multi-Response Optimization in Small-Sample Mixture--Process Experiments

URL: View paper

Brief Assessment

SVEMnet[61] focuses on self-validated ensemble models with elastic-net base learners for small-sample mixture-process experiments, not on negatively-correlated sampling for generating diverse candidate responses in alignment methods for language models.

---

5. The AI's Philosophy of Contract: An Empirical Study of Breach, Remedies, and Model Heterogeneity

URL: View paper

Brief Assessment

AI Contract Philosophy[64] focuses on legal contract interpretation by AI systems, not on alignment methods or preference dataset collection for language models.

Contribution 3: Community Alignment dataset

Description: The authors create and release Community Alignment, a large-scale multilingual preference dataset with nearly 200,000 comparisons from over 3,000 annotators across five countries and languages, built using their negatively-correlated sampling approach.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Aligning llms with individual preferences via interaction

URL: View paper

Brief Assessment

Interaction Based Alignment[34] focuses on a different type of preference dataset. While both involve multi-turn conversations, the candidate constructs 3k+ conversations for training models to infer individual preferences dynamically during interaction, whereas the original creates 200k comparisons across multiple countries/languages to capture diverse global preferences with negatively-correlated sampling.

---

2. M2lingual: Enhancing multilingual, multi-turn instruction alignment in large language models

URL: View paper

Brief Assessment

M2lingual[54] focuses on synthetic multilingual instruction-following data for LLM training, not human preference data collection. The candidate addresses a different problem (multilingual instruction tuning) using different methods (synthetic data generation via evol-instruct) rather than human preference elicitation with negatively-correlated sampling.

---

3. PLLuM-Align: Polish Preference Dataset for Large Language Model Alignment

URL: View paper

Brief Assessment

PLLuM Align[53] focuses on Polish-language preference data for alignment, while the original paper presents a multilingual dataset across five countries/languages with negatively-correlated sampling methodology. These are distinct datasets serving different linguistic communities.

---

4. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large â¦

URL: View paper

Brief Assessment

PRISM Alignment Dataset[51] focuses on participatory, representative human feedback with detailed participant profiles and survey data, while the original paper's Community Alignment dataset emphasizes negatively-correlated sampling to address algorithmic monoculture. These are complementary approaches to preference dataset collection rather than overlapping novelty claims.

---

5. A tree-of-thoughts to broaden multi-step reasoning across languages

URL: View paper

Brief Assessment

Tree of Thoughts[59] focuses on cross-lingual reasoning methods for LLMs, not on multilingual preference dataset collection for alignment. The candidate addresses prompting mechanisms for multi-step reasoning across languages, while the original contribution concerns human preference annotation infrastructure and negatively-correlated sampling for alignment datasets.

---

6. CARE: Multilingual Human Preference Learning for Cultural Awareness

URL: View paper

Brief Assessment

CARE Multilingual[55] focuses on cultural awareness across Chinese, Arab, and Japanese cultures with 3,490 questions and 31.7k responses. While both datasets address multilingual preferences, CARE targets culture-specific alignment rather than the broader value dimensions (secular-rational vs. traditional, self-expression vs. survival) emphasized in Community Alignment.

---

7. InterMT: Multi-Turn Interleaved Preference Alignment with Human Feedback

URL: View paper

Brief Assessment

InterMT[57] focuses on multi-turn, multimodal (vision-language) preference data for understanding and generation tasks, whereas Community Alignment targets multilingual, multi-turn text-based preference alignment across cultural dimensions.

---

8. Iterative Tool Usage Exploration for Multimodal Agents via Step-wise Preference Tuning

URL: View paper

Brief Assessment

Iterative Tool Usage[56] focuses on step-wise preference tuning for multimodal tool usage agents, not on creating multilingual multi-turn preference datasets for general language model alignment. The candidate's dataset construction is task-specific to tool usage exploration rather than broad cultural/political preference alignment.

---

9. HelpSteer3-Preference: Open Human-Annotated Preference Data across Diverse Tasks and Languages

URL: View paper

Brief Assessment

HelpSteer3 Preference[52] focuses on general-domain preference data across diverse tasks (STEM, coding, multilingual), not specifically on the negatively-correlated sampling methodology or the cultural/political value dimensions that are central to the original paper's Community Alignment dataset.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Cultivating Pluralism In Algorithmic Monoculture: The Community Alignment Dataset View paper
- [1] Aligning large language models with human: A survey View paper
- [2] Pretraining language models with human preferences View paper
- [3] On diversified preferences of large language model alignment View paper
- [4] Aligning large language models with human preferences through representation engineering View paper
- [5] Aligning modalities in vision large language models via preference fine-tuning View paper
- [6] Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences View paper
- [7] Human alignment of large language models through online preference optimisation View paper
- [8] Aligning multimodal llm with human preference: A survey View paper
- [9] Lifealign: Lifelong alignment for large language models with memory-augmented focalized preference optimization View paper
- [10] On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization View paper
- [11] Personalizing reinforcement learning from human feedback with variational preference learning View paper
- [12] Self-evolutionary large language models through uncertainty-enhanced preference optimization View paper
- [13] Human preferences for constructive interactions in language model alignment View paper
- [14] Aligning language models with human preferences via a bayesian approach View paper
- [15] Aligning language models with human preferences View paper
- [16] The benefits, risks and bounds of personalizing the alignment of large language models to individuals View paper
- [17] Aligning Large Vision-Language Models by Deep Reinforcement Learning and Direct Preference Optimization View paper
- [18] Unsupervised human preference learning View paper
- [19] MaxMin-RLHF: Alignment with diverse human preferences View paper
- [20] Diverse preference learning for capabilities and alignment View paper
- [21] Aligning to thousands of preferences via system message generalization View paper
- [22] From 1,000,000 users to every user: Scaling up personalized preference for user-level alignment View paper
- [23] A survey on personalized and pluralistic preference alignment in large language models View paper
- [24] Learning LLM-as-a-judge for preference alignment View paper
- [25] Scaling data diversity for fine-tuning language models in human alignment View paper
- [26] Chatglm-rlhf: Practices of aligning large language models with human feedback View paper
- [27] A Survey on Personalized Alignment--The Missing Piece for Large Language Models in Real-World Applications View paper
- [28] Capturing nuanced preferences: Preference-aligned distillation for small language models View paper
- [29] Aligning llm agents by learning latent preference from user edits View paper
- [30] Efficient Safety Alignment of Large Language Models via Preference Re-ranking and Representation-based Reward Modeling View paper
- [31] Preference learning algorithms do not learn preference rankings View paper
- [32] Omnialign-v: Towards enhanced alignment of mllms with human preference View paper
- [33] Diverse AI Feedback For Large Language Model Alignment View paper
- [34] Aligning llms with individual preferences via interaction View paper
- [35] Dissecting human and llm preferences View paper
- [36] Inference time llm alignment in single and multidomain preference spectrum View paper
- [37] OpenRubrics: Towards Scalable Synthetic Rubric Generation for Reward Modeling and LLM Alignment View paper
- [38] Personalllm: Tailoring llms to individual preferences View paper
- [39] Everyone deserves a reward: Learning customized human preferences View paper
- [40] Rethinking reward modeling in preference-based large language model alignment View paper
- [41] Heterogeneous user modeling for llm-based recommendation View paper
- [42] Panacea: Pareto alignment via preference adaptation for llms View paper
- [43] Collab: Controlled Decoding using Mixture of Agents for LLM Alignment View paper
- [44] PAL: Sample-Efficient Personalized Reward Modeling for Pluralistic Alignment View paper
- [45] Pluralistic Off-policy Evaluation and Alignment View paper
- [46] Understand What LLM Needs: Dual Preference Alignment for Retrieval-Augmented Generation View paper
- [47] Feedback-to-Text Alignment: LLM Learning Consistent Natural Language Generation from User Ratings and Loyalty Data View paper
- [48] TPO: Aligning Large Language Models with Multi-branch & Multi-step Preference Trees View paper
- [49] Linear alignment: A closed-form solution for aligning human preferences without tuning and feedback View paper
- [50] Less is more: Improving llm alignment via preference data selection View paper
- [51] The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large â View paper

- [52] HelpSteer3-Preference: Open Human-Annotated Preference Data across Diverse Tasks and Languages View paper
- [53] PLLuM-Align: Polish Preference Dataset for Large Language Model Alignment View paper
- [54] M2lingual: Enhancing multilingual, multi-turn instruction alignment in large language models View paper
- [55] CARE: Multilingual Human Preference Learning for Cultural Awareness View paper
- [56] Iterative Tool Usage Exploration for Multimodal Agents via Step-wise Preference Tuning View paper
- [57] InterMT: Multi-Turn Interleaved Preference Alignment with Human Feedback View paper
- [58] Implicit Cross-Lingual Rewarding for Efficient Multilingual Preference Alignment View paper
- [59] A tree-of-thoughts to broaden multi-step reasoning across languages View paper
- [60] Align and Complete Samples in Remote Sensing Fine-Grained Rigid Object Detection View paper
- [61] SVEMnet: An R package for Self-Validated Elastic-Net Ensembles and Multi-Response Optimization in Small-Sample Mixture--Process Experiments View paper
- [62] Antithetic sampling with Hamiltonian Monte Carlo View paper
- [63] Dynamics of Algorithmic Content Amplification on TikTok View paper
- [64] The AI's Philosophy of Contract: An Empirical Study of Breach, Remedies, and Model Heterogeneity View paper
- [65] Cultural bias and cultural alignment of large language models View paper
- [66] Culturellm: Incorporating cultural differences into large language models View paper
- [67] Investigating cultural alignment of large language models View paper
- [68] Towards measuring the representation of subjective global opinions in language models View paper
- [69] Invisible filters: Cultural bias in hiring evaluations using large language models View paper
- [70] Extrinsic evaluation of cultural competence in large language models View paper
- [71] Not all countries celebrate thanksgiving: On the cultural dominance in large language models View paper
- [72] â dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models View paper
- [73] NormAd: A framework for measuring the cultural adaptability of large language models View paper
- [74] High-dimension human value representation in large language models View paper