Novelty Assessment Report

Paper: What's In My Human Feedback? Learning Interpretable Descriptions of Preference Data
PDF URL: https://openreview.net/pdf?id=sC6A1bFDUt
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Preference data is widely used for aligning language models, but remains largely opaque. While prior work has studied specific aspects of annotator preference (e.g., length or sycophancy), automatically inferring preferences without pre-specifying hypotheses remains challenging. We introduce What's In My Human Feedback (WIMHF), a method that produces human-interpretable, natural language features from preference data using sparse autoencoders. We show that a sparse set of interpretable features can account for two-thirds of the preference signal achieved by black-box models. Applying WIMHF to 7 widely-used datasets, we precisely characterize both (1) which preferences are even possible to measure from each dataset and (2) which preferences humans actually display. WIMHF surfaces preferences that are unintentional or even actively harmful, like a preference for toxic outputs in Chatbot Arena. We show how these findings enable interpretable data curation: re-labeling the examples that contain the harmful preference yields large safety gains (+37%) with no cost to general performance. We also demonstrate a new approach to personalization: on the Community Alignment dataset, we identify preferences that are subjective across annotators, and use the features as interpretable knobs to adjust model behavior along these axes.

Core Task Landscape

This paper addresses: **Interpreting Human Preference Data for Language Model Alignment**

A total of **50 papers** were analyzed and organized into a taxonomy with **30 categories**.

Taxonomy Overview

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

- **Preference Data Quality and Selection**
- **Preference Modeling Frameworks**
- **Preference Optimization Algorithms**
- **Online and Adaptive Preference Learning**
- **Personalized and Pluralistic Alignment**
- **Preference Annotation and Feedback Acquisition**
- **AI-Generated Feedback for Alignment**
- **Domain-Specific and Task-Specific Alignment**
- **Alignment Evaluation and Analysis**
- **Continual and Lifelong Alignment**
- ... and 3 more categories

Complete Taxonomy Tree

- Interpreting Human Preference Data for Language Model Alignment Survey Taxonomy
- Preference Data Quality and Selection
  - Data Influence and Valuation Methods (1 papers)
  - [1] Towards understanding valuable preference data for large language model alignment (Zhang, 2025) View paper
  - Annotation Efficiency Strategies (2 papers)
  - [30] Not Everything is All You Need: Toward Low-Redundant Optimization for Large Language Model Alignment (Chen Zhipeng, 2024) View paper
  - [41] Annotation-Efficient Preference Optimization for Language Model Alignment (Yuu Jinnai, 2024) View paper
  - Diversity and Coverage Optimization (2 papers)
  - [35] Scaling Data Diversity for Fine-Tuning Language Models in Human Alignment (Song, 2024) View paper
  - [44] HelpSteer3-Preference: Open Human-Annotated Preference Data across Diverse Tasks and Languages (Wang Zhi-lin, 2025) View paper
- Preference Modeling Frameworks
  - Beyond Bradley-Terry Models (3 papers)
  - [2] Self-Play Preference Optimization for Language Model Alignment (Wu Yue, 2024) View paper
  - [11] Beyond Bradley-Terry Models: A General Preference Model for Language Model Alignment (Zhang Yifan, 2024) View paper
  - [48] General Preference Modeling with Preference Representations for Aligning Language Models (Zhang Yifan, 2024) View paper
  - Interpretable Preference Representations ★ (1 papers)
  - [0] What's In My Human Feedback? Learning Interpretable Descriptions of Preference Data (Anon et al., 2026) View paper
  - Multi-Objective and Pluralistic Preference Modeling (2 papers)
  - [20] Multi-Objective Preference Optimization: Improving Human Alignment of Generative Models (Agnihotri, 2025) View paper
  - [32] Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences (Chakraborty, 2024) View paper

- Preference Optimization Algorithms
  - Direct Preference Optimization Variants
  - Calibrated and Contrastive Preference Optimization (2 papers)
    - [6] Cal-DPO: Calibrated Direct Preference Optimization for Language Model Alignment (Vasant Honavar, 2024) View paper
    - [24] Anchored preference optimization and contrastive revisions: Addressing underspecification in alignment (Karel Dâ□□Oosterlinck, 2025) View paper
  - Listwise and Ranking-Based Optimization (2 papers)
    - [7] Preference ranking optimization for human alignment (Feifan Song, 2024) View paper
    - [8] LPOI: Listwise Preference Optimization for Vision Language Models (Oh, 2025) View paper
  - Game-Theoretic and Robust Preference Optimization (2 papers)
    - [10] Stackelberg Game Preference Optimization for Data-Efficient Alignment of Language Models (Chu Xu, 2025) View paper
    - [40] Magnetic Preference Optimization: Achieving Last-iterate Convergence for Language Models Alignment (Wang Mingzhi, 2024) View paper
  - Reward-Based Reinforcement Learning Approaches (2 papers)
  - [14] RS-DPO: A Hybrid Rejection Sampling and Direct Preference Optimization Method for Alignment of Large Language Models (Khaki, 2024) View paper
  - [45] RRHF: Rank Responses to Align Language Models with Human Feedback without tears (YUAN Zheng, 2023) View paper
  - Hybrid and Multi-Stage Alignment Training (1 papers)
  - [31] Hybrid Alignment Training for Large Language Models (Chenglong Wang, 2024) View paper
- Online and Adaptive Preference Learning
  - Exploration Strategies for Online RLHF (1 papers)
  - [13] Online preference alignment for language models via count-based exploration (Bai, 2025) View paper
  - Self-Rewarding and Self-Play Mechanisms (1 papers)
  - [36] CREAM: Consistency Regularized Self-Rewarding Language Models (Wang Zhaoyang, 2024) View paper
  - Strategic and Adversarial Feedback Handling (1 papers)
  - [28] Online learning from strategic human feedback in llm fine-tuning (Shugang Hao, 2025) View paper
- Personalized and Pluralistic Alignment
  - User-Level Preference Modeling (2 papers)
  - [22] From 1,000,000 users to every user: Scaling up personalized preference for user-level alignment (Li Jiaâ□□nan, 2025) View paper
  - [29] A Survey on Personalized and Pluralistic Preference Alignment in Large Language Models (Xie, 2025) View paper
  - Participatory and Multicultural Alignment (1 papers)
  - [12] â□¦ alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language â□¦ (HR Kirk, 2024) View paper
- Preference Annotation and Feedback Acquisition
  - Feedback Format and Structure Design (2 papers)
  - [18] Peering Through Preferences: Unraveling Feedback Acquisition for Aligning Large Language Models (Bansal, 2023) View paper
  - [26] Comparing Bad Apples to Good Oranges Aligning Large Language Models via Joint Preference Optimization (Hritik Bansal, 2025) View paper
  - Human-AI Collaborative Annotation (2 papers)
  - [34] Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences (Shreya Shankar, 2024) View paper
  - [42] On the use of Eye-Tracking during preference annotation for the alignment of Vision-Language Models (Mazzini, 2025) View paper
  - Implicit Feedback and Behavioral Signals (1 papers)
  - [37] Rlhf fine-tuning of llms for alignment with implicit user feedback in conversational recommenders (Yang Zhong-heng, 2025) View paper
- AI-Generated Feedback for Alignment
  - LLM-as-Judge Evaluation (1 papers)
  - [5] Aligning with human judgement: The role of pairwise preference in large language model evaluators (Liu, 2024) View paper
  - Vision-Language Feedback Generation (1 papers)
  - [39] VLFeedback: A Large-Scale AI Feedback Dataset for Large Vision-Language Models Alignment (Lei Li, 2024) View paper
- Domain-Specific and Task-Specific Alignment
  - Safety and Harmlessness Alignment (1 papers)
  - [21] BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset (Ji, 2023) View paper
  - Multimodal Vision-Language Alignment (2 papers)
  - [16] HSCR: Hierarchical Self-Contrastive Rewarding for Aligning Medical Vision Language Models (Zhang Yan, 2025) View paper
  - [17] Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback (Tianyu Yu, 2024) View paper
  - Specialized Application Alignment (2 papers)
  - [33] AesthetiQ: Enhancing Graphic Layout Design via Aesthetic-Aware Preference Alignment of Multi-modal Large Language Models (Patnaik, 2025) View paper
  - [38] Recexplainer: Aligning large language models for explaining recommendation models (Yuxuan Lei, 2024) View paper
- Alignment Evaluation and Analysis
  - Factor-Level Preference Analysis (1 papers)
  - [23] Uncovering Factor Level Preferences to Improve Human-Model Alignment (Oh, 2024) View paper
  - Alignment Quality Assessment (1 papers)
  - [3] Rethinking reward modeling in preference-based large language model alignment (H Sun, 2025) View paper
- Continual and Lifelong Alignment (1 papers)
  - [50] Lifealign: Lifelong alignment for large language models with memory-augmented focalized preference optimization (Li Junsong, 2025) View paper

- Pre-Training and Fine-Tuning Interactions (2 papers)
  - [43] Amuro & Char: Analyzing the Relationship between Pre-Training and Fine-Tuning of Large Language Models (Sun, 2024) View paper
  - [46] PoliTune: Analyzing the Impact of Data Selection and Fine-Tuning on Economic and Political Biases in Large Language Models (Ahmed Agiza, 2024) View paper
- Surveys and Comprehensive Reviews (3 papers)
  - [9] A survey on human preference learning for large language models (Jiang, 2024) View paper
  - [19] Towards a unified view of preference learning for large language models: A survey (Gao, 2024) View paper
  - [27] Aligning large language models with human: A survey (Wang Yu-Fei, 2023) View paper
- Practical Implementations and Systems (5 papers)
  - [4] Openassistant conversations-democratizing large language model alignment (A KÃ¶pf, 2023) View paper
  - [15] Tuning for LLM alignment (Uday Kamath, 2024) View paper
  - [25] ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools (Team GLM, 2024) View paper
  - [47] PRompt Optimization in Multi-Step Tasks (PROMST): Integrating Human Feedback and Preference Alignment (Chen Yongchao, 2024) View paper
  - [49] Enhancing Visual-Language Modality Alignment in Large Vision Language Models via Self-Improvement (Xiyao Wang, 2024) View paper

Narrative

Core task: Interpreting human preference data for language model alignment. The field has evolved into a rich ecosystem organized around several major themes. At the highest level, researchers address data quality and selection—ensuring that preference signals are informative and representative—while simultaneously developing preference modeling frameworks that translate raw comparisons into learnable representations. Parallel branches focus on preference optimization algorithms, which refine model behavior given these signals, and on online or adaptive learning schemes that update models as new feedback arrives. Additional branches explore personalized and pluralistic alignment to accommodate diverse user values, methods for acquiring annotations and feedback (including AI-generated alternatives), and domain-specific tuning for specialized tasks. Complementary work examines evaluation strategies, continual learning paradigms, and the interplay between pre-training and fine-tuning, with surveys and practical systems rounding out the taxonomy.

Within this landscape, a particularly active line of inquiry concerns how to represent and leverage preference information more effectively. Some studies question the sufficiency of standard pairwise comparisons, proposing listwise or ranking-based formulations (Listwise Preference Optimization[8], Preference Ranking Optimization[7]) or revisiting foundational assumptions like the Bradley-Terry model (Beyond Bradley-Terry[11], Rethinking Reward Modeling[3]). Others investigate what makes preference data valuable (Valuable Preference Data[1]) or how strategic annotator behavior shapes feedback (Strategic Human Feedback[28]). Interpretable Preference Descriptions[0] sits squarely in the preference modeling frameworks branch, emphasizing interpretable representations that make the underlying structure of human judgments more transparent. This focus on interpretability contrasts with purely algorithmic approaches like Self-Play Preference Optimization[2] and aligns closely with efforts to understand the role and limitations of pairwise signals (Pairwise Preference Role[5]), offering a complementary lens on how preference data can be both modeled and explained.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses on extracting human-interpretable features from preference data using techniques like sparse autoencoders, aiming to make preference representations transparent. The sibling subtopics address different challenges: 'Beyond Bradley-Terry Models' tackles complex preference structures that violate standard assumptions (intransitivity, irrationality), while 'Multi-Objective and Pluralistic Preference Modeling' handles scenarios with multiple conflicting objectives or diverse annotator viewpoints. All three areas work with preference data but differ in their primary goals—interpretability versus structural complexity versus pluralism.

Similarities: - All three subtopics work within the domain of interpreting and modeling human preference data for language model alignment - Each addresses limitations of standard preference modeling approaches - All aim to better capture the nuances and complexities of human preferences beyond simple pairwise comparisons

Differences: - Interpretable Preference Representations focuses on transparency and feature extraction (what preferences mean), while Beyond Bradley-Terry Models focuses on structural assumptions (how preferences relate), and Multi-Objective handles multiplicity (whose/which preferences) - The original leaf uses specific techniques (sparse autoencoders) for decomposition, while siblings use embeddings/game theory or multi-objective frameworks respectively - Interpretable Preference Representations maintains interpretability as the primary goal, whereas siblings prioritize modeling fidelity to complex/diverse preference structures - The original leaf's exclude note targets black-box models, while Beyond Bradley-Terry excludes Bradley-Terry assumptions, and Multi-Objective excludes single-objective methods—reflecting different architectural concerns

Suggested Search Directions: - Investigate whether interpretable representations can be extracted from Beyond Bradley-Terry models (e.g., interpreting preference embeddings) - Explore how sparse autoencoders might decompose multi-objective or pluralistic preferences into interpretable dimensions - Examine papers combining interpretability with non-transitive or multi-stakeholder preference structures

Sibling Subtopics

- **Beyond Bradley-Terry Models** (leaves: 1, papers: 3)
- Scope: Frameworks addressing intransitivity, irrationality, or complex preference structures using embeddings or game-theoretic approaches.
- Exclude: Excludes methods that assume Bradley-Terry structure; see Calibrated and Contrastive Preference Optimization.
- **Multi-Objective and Pluralistic Preference Modeling** (leaves: 1, papers: 2)
- Scope: Frameworks that model multiple conflicting objectives or diverse annotator preferences simultaneously.
- Exclude: Excludes single-objective methods; see Beyond Bradley-Terry Models.

Contributions Analysis

Overall novelty summary. The paper introduces WIMHF, a method that extracts human-interpretable natural language features from preference data using sparse autoencoders. It occupies the 'Interpretable Preference Representations' leaf within the 'Preference Modeling Frameworks' branch of the taxonomy. Notably, this leaf contains only the original paper itself—no sibling papers exist in this specific category. This positioning suggests the work addresses a relatively sparse research direction: while the broader field includes numerous preference modeling approaches (Beyond Bradley-Terry models, multi-objective frameworks), the specific focus on extracting interpretable features from preference data appears less explored within the examined literature.

The taxonomy reveals substantial activity in adjacent areas. The parent branch 'Preference Modeling Frameworks' includes work on complex preference structures (intransitivity, game-theoretic approaches) and multi-objective modeling, but these typically remain black-box representations. Neighboring branches address preference optimization algorithms (DPO variants, reward-based RL) and data quality methods (influence functions, annotation efficiency), yet these focus on algorithmic refinement rather than interpretability. The 'Alignment Evaluation and Analysis' branch includes factor-level preference analysis, which shares interpretability goals but approaches the problem from an evaluation rather than modeling perspective. WIMHF's use of sparse autoencoders to surface interpretable features bridges preference modeling and analysis in a way that appears distinct from existing categorical boundaries.

Among 30 candidates examined across three contributions, none clearly refute the core claims. The WIMHF method itself (10 candidates examined, 0 refutable) appears novel in its application of sparse autoencoders to preference data interpretation. The interpretable data curation contribution (10 candidates, 0 refutable) demonstrates practical safety improvements through targeted re-labeling, a use case not prominently covered in the examined literature. The personalization approach (10 candidates, 0 refutable) similarly shows no substantial prior overlap. The limited search scope means these findings reflect top-30 semantic matches rather than exhaustive coverage, but within this sample, the work's combination of interpretability techniques and preference data analysis appears distinctive.

Based on the examined candidates and taxonomy structure, the work occupies a relatively unexplored niche at the intersection of interpretability and preference modeling. The absence of sibling papers in its taxonomy leaf and the lack of refutable prior work among 30 candidates suggest meaningful novelty, though the limited search scope prevents definitive claims about the broader literature. The practical applications to safety and personalization extend beyond pure modeling contributions, addressing gaps in how preference data is understood and curated.

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

Contribution 1: What's In My Human Feedback (WIMHF) method

Description: The authors propose WIMHF, a three-step procedure that uses sparse autoencoders to automatically discover interpretable natural language features from preference datasets, enabling analysis of both measurable preferences (features that vary between responses) and realized preferences (features that affect human labels) without pre-specifying hypotheses.

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. SAFER: Probing Safety in Reward Models with Sparse Autoencoder

URL: View paper

Brief Assessment

SAFER[62] focuses on interpreting reward model activations for safety auditing and targeted data modification, while WIMHF analyzes preference datasets to discover measurable and realized preferences. The methods serve different purposes in the alignment pipeline.

---

2. Transcoders Beat Sparse Autoencoders for Interpretability

URL: View paper

Brief Assessment

Transcoders Beat SAEs[64] focuses on improving sparse autoencoder architectures for neural network interpretability, not on analyzing preference datasets or extracting features from human feedback data.

---

3. Interpretable Reward Model via Sparse Autoencoder

URL: View paper

Brief Assessment

Interpretable Reward Model[61] focuses on integrating sparse autoencoders into reward models to produce interpretable reward scores for RLHF alignment, not on analyzing preference datasets to discover what features humans prefer. The original paper's WIMHF method specifically aims to explain preference data by discovering measurable and realized preferences from datasets, which is a different application domain.

---

4. Universal Sparse Autoencoders: Interpretable Cross-Model Concept Alignment

URL: View paper

Brief Assessment

Universal Sparse Autoencoders[68] focuses on cross-model concept alignment in vision models, not on extracting interpretable features from preference datasets or analyzing human feedback for language model alignment.

---

5. Towards Interpretable Scientific Foundation Models: Sparse Autoencoders for Disentangling Dense Embeddings of Scientific Concepts

URL: View paper

Brief Assessment

Scientific Foundation Models[63] applies sparse autoencoders to dense text embeddings from scientific literature for interpretability and search, not to preference datasets for analyzing human feedback signals as in the original paper.

---

6. Interpretability as Compression: Reconsidering SAE Explanations of Neural Activations

URL: View paper

Brief Assessment

Interpretability as Compression[67] focuses on using sparse autoencoders (SAEs) to interpret neural network activations through a minimal description length framework, not on extracting interpretable features from preference datasets. The candidate addresses SAE interpretability for model internals, while WIMHF applies SAEs to analyze human feedback data for alignment purposes.

---

7. Sparse autoencoders uncover biologically interpretable features in protein language model representations.

URL: View paper

Brief Assessment

Protein Language Models[69] applies sparse autoencoders to protein sequence representations to discover biologically interpretable features, not to preference data or human feedback analysis. The domains and applications are fundamentally different.

---

8. Sparse Autoencoders Reveal Interpretable Features in Single-Cell Foundation Models

URL: View paper

Brief Assessment

Single-Cell Foundation Models[66] applies sparse autoencoders to biological single-cell data for cell type annotation and data integration, not to preference datasets or human feedback analysis. The domains and applications are fundamentally different.

---

9. Sparse autoencoders match supervised features for model steering on the ioi task

URL: View paper

Brief Assessment

SAEs Match Supervised[65] applies sparse autoencoders to model steering on the IOI task, comparing unsupervised SAE features against supervised feature dictionaries for controlling model behavior. This is fundamentally different from WIMHF's application of SAEs to extract interpretable natural language features from preference datasets to understand human feedback.

---

10. Sparse Autoencoders Find Highly Interpretable Features in Language Models

URL: View paper

Brief Assessment

SAEs Find Interpretable Features[70] focuses on resolving polysemanticity in neural network activations using sparse autoencoders, not on analyzing preference datasets or human feedback. The application domains are fundamentally different.

Contribution 2: Interpretable data curation for safety improvement

Description: The authors demonstrate that WIMHF enables targeted data curation by identifying and correcting misaligned preferences in datasets. For example, flipping labels on examples with harmful anti-refusal preferences in Chatbot Arena substantially improves safety metrics while preserving overall model performance.

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. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions

URL: View paper

Brief Assessment

Safety-Tuned LLaMAs[71] focuses on adding safety examples during instruction-tuning to improve model safety, not on identifying and correcting misaligned preferences in existing datasets through interpretable feature analysis. The candidate addresses safety through data augmentation, while the original paper addresses it through interpretable data curation based on discovered preference features.

---

2. Tower+: Bridging Generality and Translation Specialization in Multilingual LLMs

URL: View paper

Brief Assessment

Tower Plus[79] focuses on multilingual translation and general-purpose capabilities through continued pretraining and reinforcement learning, not on interpretable data curation methods for identifying and correcting misaligned preferences in safety datasets.

---

3. Deep ignorance: Filtering pretraining data builds tamper-resistant safeguards into open-weight llms

URL: View paper

Brief Assessment

Deep Ignorance[76] focuses on pretraining data filtering to prevent dangerous capabilities from being learned initially, while the original paper addresses post-hoc correction of misaligned preferences in already-collected feedback datasets. These are fundamentally different stages of the ML pipeline with different technical approaches.

---

4. Compassjudger-2: Towards generalist judge model via verifiable rewards

URL: View paper

Brief Assessment

CompassJudger-2[80] focuses on developing a generalist judge model for LLM evaluation through task-driven data curation and verifiable rewards. The original paper addresses interpretable preference data analysis and targeted safety improvements through feature-based label correction, which is a different technical approach and application domain.

---

5. Phi-4-reasoning technical report

URL: View paper

Brief Assessment

Phi-4-Reasoning[73] focuses on training reasoning models through careful data curation for supervised fine-tuning and reinforcement learning, but does not address safety-specific data curation or the identification and correction of misaligned preferences in datasets as described in the original paper.

---

6. EnsembleXAI-Motor: A Lightweight Framework for Fault Classification in Electric Vehicle Drive Motors Using Feature Selection, Ensemble Learning, and Explainable AI

URL: View paper

Brief Assessment

EnsembleXAI-Motor[78] focuses on fault diagnosis in electric vehicle drive motors using machine learning and explainable AI (LIME), not on data curation for language model alignment or safety improvement through preference data correction.

---

7. SPA-VL: A Comprehensive Safety Preference Alignment Dataset for Vision Language Models

URL: View paper

Brief Assessment

SPA-VL[75] focuses on constructing a large-scale safety preference dataset for vision-language models using automated data collection and annotation. The paper does not demonstrate targeted data curation by identifying and correcting misaligned preferences through interpretable features, which is the core novelty of the original contribution.

---

8. Large language models for reticular chemistry

URL: View paper

Brief Assessment

LLMs Reticular Chemistry[74] focuses on applying large language models to reticular chemistry tasks (data mining, material design, synthesis automation). It does not address safety alignment, preference data curation, or harmful content filtering in language models.

---

9. Safe delta: Consistently preserving safety when fine-tuning LLMs on diverse datasets

URL: View paper

Brief Assessment

Safe Delta[77] focuses on post-training defense methods that adjust model parameters after fine-tuning to preserve safety, not on identifying and correcting misaligned preferences within training datasets before model training. The original paper's contribution involves pre-training data curation through interpretable feature discovery.

---

10. Constraint-guided online data selection for scalable data-driven safety filters in uncertain robotic systems

URL: View paper

Brief Assessment

Constraint-Guided Data Selection[72] focuses on selecting informative data points for robotic control systems using Gaussian process regression, not on curating preference datasets to improve language model safety. The domains (robotics vs. NLP alignment) and methods (GP-based data selection vs. sparse autoencoders for preference analysis) are fundamentally different.

Contribution 3: Interpretable personalization approach

Description: The authors introduce an interpretable personalization method that identifies subjective preferences across annotators and learns user-specific coefficients for selected features. This approach allows practitioners to personalize models on acceptable attributes while preventing undesirable personalization, such as creating echo chambers.

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. On generative agents in recommendation

URL: View paper

Brief Assessment

Generative Agents Recommendation[51] focuses on simulating user behavior in recommender systems using LLM-powered agents with profile, memory, and action modules. This is fundamentally different from the original paper's interpretable personalization method that identifies subjective preferences using sparse autoencoders and learns user-specific coefficients for selected features to enable controlled personalization while preventing echo chambers.

---

2. When large language models meet personalization: Perspectives of challenges and opportunities

URL: View paper

Brief Assessment

LLMs Meet Personalization[52] discusses personalization in recommender systems broadly but does not present a method for identifying subjective preferences across annotators with user-specific coefficients for interpretable features as described in the original paper.

---

3. ReasoningRec: Bridging Personalized Recommendations and Human-Interpretable Explanations through LLM Reasoning

URL: View paper

Brief Assessment

ReasoningRec[53] focuses on recommendation systems using LLMs to generate explanations for user preferences in item recommendations, not on learning interpretable features across annotators for preference data alignment as in the original paper.

---

4. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5)

URL: View paper

Brief Assessment

P5[57] focuses on personalized prompts for recommendation tasks using language models, not on interpretable feature-based personalization for subjective preferences in human feedback data. The technical approaches and application domains are fundamentally different.

---

5. Recommender Systems for Renewable Energy Communities: Tailoring LLM-Powered Recommendations to User Personal Values and Literacy

URL: View paper

Brief Assessment

Renewable Energy Recommendations[59] focuses on recommender systems for renewable energy communities using LLMs, which is a different application domain from the ORIGINAL paper's work on preference data analysis and language model alignment.

---

6. RPM: Reasoning-Level Personalization for Black-Box Large Language Models

URL: View paper

Brief Assessment

Reasoning-Level Personalization[56] focuses on personalizing LLM reasoning processes using structured rationales from user behavior patterns, while the original paper personalizes preference models using sparse autoencoder features to identify subjective preferences in feedback data. These are fundamentally different personalization approaches applied to different problems (LLM generation vs. preference modeling).

---

7. The effectiveness of personalised food choice advice tailored to an individual's socio-demographic, cognitive characteristics, and sensory preferences

URL: View paper

Brief Assessment

Personalized Food Choice[54] focuses on personalizing dietary advice based on socio-demographic, cognitive, and sensory characteristics for health behavior change. The original paper addresses personalization in language model alignment using interpretable features from preference data, which is a fundamentally different domain and technical approach.

---

8. Integrating food preference Profiling, behavior change Strategies, and machine learning for cardiovascular disease prevention in a personalized nutrition â□¦

URL: View paper

Brief Assessment

Food Preference Profiling[58] focuses on personalized nutrition recommendations for cardiovascular disease prevention, not on interpretable feature-based personalization for subjective preferences in language model alignment or general machine learning contexts.

---

9. Prefpalette: Personalized preference modeling with latent attributes

URL: View paper

Brief Assessment

PrefPalette[55] focuses on decomposing preferences into predefined attribute dimensions (formality, humor, cultural values) for community-level personalization on Reddit, while the original paper learns interpretable features automatically from preference data using sparse autoencoders without pre-specifying attributes, and demonstrates personalization on individual annotators with the ability to restrict personalization to chosen attributes to prevent echo chambers.

---

10. Exploring the impact of explainable AI and cognitive capabilities on usersâ decisions

URL: View paper

Brief Assessment

Explainable AI Impact[60] focuses on how different explanation styles (example-based, feature-based, rule-based, counterfactual) affect user decisions and cognitive load in AI systems, examining personality traits like need for cognition. The original paper's contribution addresses personalization of preference models using interpretable features to prevent echo chambers in language model alignment—a fundamentally different domain and technical approach.

Appendix: Text Similarity Detection

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

References

- [0] What's In My Human Feedback? Learning Interpretable Descriptions of Preference Data View paper
- [1] Towards understanding valuable preference data for large language model alignment View paper
- [2] Self-Play Preference Optimization for Language Model Alignment View paper
- [3] Rethinking reward modeling in preference-based large language model alignment View paper
- [4] Openassistant conversations-democratizing large language model alignment View paper
- [5] Aligning with human judgement: The role of pairwise preference in large language model evaluators View paper
- [6] Cal-DPO: Calibrated Direct Preference Optimization for Language Model Alignment View paper
- [7] Preference ranking optimization for human alignment View paper
- [8] LPOI: Listwise Preference Optimization for Vision Language Models View paper
- [9] A survey on human preference learning for large language models View paper
- [10] Stackelberg Game Preference Optimization for Data-Efficient Alignment of Language Models View paper
- [11] Beyond Bradley-Terry Models: A General Preference Model for Language Model Alignment View paper
- [12] â¦ alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language â¦ View paper
- [13] Online preference alignment for language models via count-based exploration View paper
- [14] RS-DPO: A Hybrid Rejection Sampling and Direct Preference Optimization Method for Alignment of Large Language Models View paper
- [15] Tuning for LLM alignment View paper
- [16] HSCR: Hierarchical Self-Contrastive Rewarding for Aligning Medical Vision Language Models View paper
- [17] Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback View paper
- [18] Peering Through Preferences: Unraveling Feedback Acquisition for Aligning Large Language Models View paper
- [19] Towards a unified view of preference learning for large language models: A survey View paper
- [20] Multi-Objective Preference Optimization: Improving Human Alignment of Generative Models View paper
- [21] BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset View paper
- [22] From 1,000,000 users to every user: Scaling up personalized preference for user-level alignment View paper
- [23] Uncovering Factor Level Preferences to Improve Human-Model Alignment View paper
- [24] Anchored preference optimization and contrastive revisions: Addressing underspecification in alignment View paper
- [25] ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools View paper
- [26] Comparing Bad Apples to Good Oranges Aligning Large Language Models via Joint Preference Optimization View paper
- [27] Aligning large language models with human: A survey View paper
- [28] Online learning from strategic human feedback in llm fine-tuning View paper
- [29] A Survey on Personalized and Pluralistic Preference Alignment in Large Language Models View paper
- [30] Not Everything is All You Need: Toward Low-Redundant Optimization for Large Language Model Alignment View paper
- [31] Hybrid Alignment Training for Large Language Models View paper
- [32] Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences View paper
- [33] AesthetiQ: Enhancing Graphic Layout Design via Aesthetic-Aware Preference Alignment of Multi-modal Large Language Models View paper
- [34] Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences View paper
- [35] Scaling Data Diversity for Fine-Tuning Language Models in Human Alignment View paper
- [36] CREAM: Consistency Regularized Self-Rewarding Language Models View paper
- [37] Rlhf fine-tuning of llms for alignment with implicit user feedback in conversational recommenders View paper
- [38] Recexplainer: Aligning large language models for explaining recommendation models View paper
- [39] VLFeedback: A Large-Scale AI Feedback Dataset for Large Vision-Language Models Alignment View paper
- [40] Magnetic Preference Optimization: Achieving Last-iterate Convergence for Language Models Alignment View paper
- [41] Annotation-Efficient Preference Optimization for Language Model Alignment View paper
- [42] On the use of Eye-Tracking during preference annotation for the alignment of Vision-Language Models View paper
- [43] Amuro & Char: Analyzing the Relationship between Pre-Training and Fine-Tuning of Large Language Models View paper
- [44] HelpSteer3-Preference: Open Human-Annotated Preference Data across Diverse Tasks and Languages View paper
- [45] RRHF: Rank Responses to Align Language Models with Human Feedback without tears View paper

- [46] PoliTune: Analyzing the Impact of Data Selection and Fine-Tuning on Economic and Political Biases in Large Language Models View paper
- [47] PRompt Optimization in Multi-Step Tasks (PROMST): Integrating Human Feedback and Preference Alignment View paper
- [48] General Preference Modeling with Preference Representations for Aligning Language Models View paper
- [49] Enhancing Visual-Language Modality Alignment in Large Vision Language Models via Self-Improvement View paper
- [50] Lifealign: Lifelong alignment for large language models with memory-augmented focalized preference optimization View paper
- [51] On generative agents in recommendation View paper
- [52] When large language models meet personalization: Perspectives of challenges and opportunities View paper
- [53] ReasoningRec: Bridging Personalized Recommendations and Human-Interpretable Explanations through LLM Reasoning View paper
- [54] The effectiveness of personalised food choice advice tailored to an individual's socio-demographic, cognitive characteristics, and sensory preferences View paper
- [55] Prefpalette: Personalized preference modeling with latent attributes View paper
- [56] RPM: Reasoning-Level Personalization for Black-Box Large Language Models View paper
- [57] Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5) View paper
- [58] Integrating food preference Profiling, behavior change Strategies, and machine learning for cardiovascular disease prevention in a personalized nutrition â View paper
- [59] Recommender Systems for Renewable Energy Communities: Tailoring LLM-Powered Recommendations to User Personal Values and Literacy View paper
- [60] Exploring the impact of explainable AI and cognitive capabilities on usersâ decisions View paper
- [61] Interpretable Reward Model via Sparse Autoencoder View paper
- [62] SAFER: Probing Safety in Reward Models with Sparse Autoencoder View paper
- [63] Towards Interpretable Scientific Foundation Models: Sparse Autoencoders for Disentangling Dense Embeddings of Scientific Concepts View paper
- [64] Transcoders Beat Sparse Autoencoders for Interpretability View paper
- [65] Sparse autoencoders match supervised features for model steering on the ioi task View paper
- [66] Sparse Autoencoders Reveal Interpretable Features in Single-Cell Foundation Models View paper
- [67] Interpretability as Compression: Reconsidering SAE Explanations of Neural Activations View paper
- [68] Universal Sparse Autoencoders: Interpretable Cross-Model Concept Alignment View paper
- [69] Sparse autoencoders uncover biologically interpretable features in protein language model representations. View paper
- [70] Sparse Autoencoders Find Highly Interpretable Features in Language Models View paper
- [71] Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions View paper
- [72] Constraint-guided online data selection for scalable data-driven safety filters in uncertain robotic systems View paper
- [73] Phi-4-reasoning technical report View paper
- [74] Large language models for reticular chemistry View paper
- [75] SPA-VL: A Comprehensive Safety Preference Alignment Dataset for Vision Language Models View paper
- [76] Deep ignorance: Filtering pretraining data builds tamper-resistant safeguards into open-weight llms View paper
- [77] Safe delta: Consistently preserving safety when fine-tuning LLMs on diverse datasets View paper
- [78] EnsembleXAI-Motor: A Lightweight Framework for Fault Classification in Electric Vehicle Drive Motors Using Feature Selection, Ensemble Learning, and Explainable AI View paper
- [79] Tower+: Bridging Generality and Translation Specialization in Multilingual LLMs View paper
- [80] Compassjudger-2: Towards generalist judge model via verifiable rewards View paper