Novelty Assessment Report

Paper: AdAEM: An Adaptively and Automated Extensible Evaluation Method of LLMs' Value Difference

PDF URL: https://openreview.net/pdf?id=qNlTH4kYJZ

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

Assessing Large Language Models (LLMs)' underlying value differences enables comprehensive comparison of their misalignment, cultural adaptability, and biases. Nevertheless, current value measurement methods face the informativeness challenge: with often outdated, contaminated, or generic test questions, they can only capture the orientations on comment safety values, e.g., HHH, shared among different LLMs, leading to indistinguishable and uninformative results. To address this problem, we introduce AdAEM, a novel, self-extensible evaluation algorithm for revealing LLMs' inclinations. Distinct from static benchmarks, AdAEM automatically and adaptively generates and extends its test questions. This is achieved by probing the internal value boundaries of a diverse set of LLMs developed across cultures and time periods in an in-context optimization manner. Such a process theoretically maximizes an information-theoretic objective to extract diverse controversial topics that can provide more distinguishable and informative insights about models' value differences. In this way, AdAEM is able to co-evolve with the development of LLMs, consistently tracking their value dynamics. We use AdAEM to generate novel questions and conduct an extensive analysis, demonstrating our method's validity and effectiveness, laying the groundwork for better interdisciplinary research on LLMs' values and alignment.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Adaptive Value Evaluation of Large Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **26 categories**.

Taxonomy Overview

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

- **Value Alignment and Orientation Assessment**
- **Reinforcement Learning and Value-Based Optimization**
- **Adaptive Planning and Decision-Making Agents**
- **Evaluation Frameworks and Benchmarking Methodologies**
- **Adaptive Model Optimization and Compression**
- **Memory Management and KV Cache Optimization**
- **Adaptive Inference and Realignment Strategies**
- **Domain-Specific Adaptive Applications**
- **Specialized Adaptive Techniques**

Complete Taxonomy Tree

- Adaptive Value Evaluation of Large Language Models Survey Taxonomy
- Value Alignment and Orientation Assessment
  - Heterogeneous and Multi-Objective Value Alignment (4 papers)
  - [1] Heterogeneous value alignment evaluation for large language models (Zhaowei Zhang, 2025) View paper
  - [17] MetaAligner: Towards Generalizable Multi-Objective Alignment of Language Models (Sophia Ananiadou, 2024) View paper
  - [22] Heterogeneous Value Evaluation for Large Language Models (Zhang, 2023) View paper
  - [30] MAVIS: Multi-Objective Alignment via Value-Guided Inference-Time Search (Jeremy Carleton, 2025) View paper
  - Value Measurement Benchmarks and Psychometric Evaluation (3 papers)
  - [6] Valuebench: Towards comprehensively evaluating value orientations and understanding of large language models (Ye HaoRan, 2024) View paper
  - [39] Value compass benchmarks: A comprehensive, generative and self-evolving platform for llms' value evaluation (Jing Yao, 2025) View paper
  - [45] Value Compass Benchmarks: A Platform for Fundamental and Validated Evaluation of LLMs Values (Yao Jing, 2025) View paper
  - Adaptive Value Evaluation Methods ★ (3 papers)
  - [0] AdAEM: An Adaptively and Automated Extensible Evaluation Method of LLMs' Value Difference (Anon et al., 2026) View paper
  - [9] Clave: An adaptive framework for evaluating values of llm generated responses (Yao Jing, 2024) View paper
  - [27] AdAEM: An Adaptively and Automated Extensible Measurement of LLMs' Value Difference (Yi, 2025) View paper
  - Value-Action Alignment and Behavioral Consistency (2 papers)
  - [36] Process for adapting language models to society (palms) with values-targeted datasets (Irene Solaiman, 2021) View paper
  - [37] Mind the Value-Action Gap: Do LLMs Act in Alignment with Their Values? (Shen Hua, 2025) View paper
- Reinforcement Learning and Value-Based Optimization
  - Token-Level and Step-Level Value Estimation (4 papers)
  - [8] Segment Policy Optimization: Effective Segment-Level Credit Assignment in RL for Large Language Models (Guo, 2025) View paper
  - [12] Enhancing decision-making for llm agents via step-level q-value models (Zhai, 2025) View paper

  - [19] Token-Supervised Value Models for Enhancing Mathematical Reasoning Capabilities of Large Language Models (Lee Jung Hyun, 2024) View paper
  - [24] Inverse-Q*: Token Level Reinforcement Learning for Aligning Large Language Models Without Preference Data (Xia Han, 2024) View paper
  - Value-Guided Inference-Time Search and Optimization (4 papers)
  - [21] Language models can self-improve at state-value estimation for better search (Mendes, 2025) View paper
  - [23] Value augmented sampling for language model alignment and personalization (Han, 2024) View paper
  - [42] SWE-Search: Enhancing Software Agents with Monte Carlo Tree Search and Iterative Refinement (Antonis Antoniades, 2024) View paper
  - [48] Inference-time language model alignment via integrated value guidance (Zhixuan Liu, 2024) View paper
  - Direct Value Optimization and Refinement (2 papers)
  - [25] Direct value optimization: Improving chain-of-thought reasoning in llms with refined values (Zhang Hong-bo, 2025) View paper
  - [49] Memory-Driven Self-Improvement for Decision Making with Large Language Models (Yan Xue, 2025) View paper
- Adaptive Planning and Decision-Making Agents (4 papers)
  - [3] Adaplanner: Adaptive planning from feedback with language models (Sun Haotian, 2023) View paper
  - [10] G-Sim: Generative Simulations with Large Language Models and Gradient-Free Calibration (Holt, 2025) View paper
  - [18] Adaptation with self-evaluation to improve selective prediction in llms (Chen Jie-feng, 2023) View paper
  - [46] Agentic LLM Framework for Adaptive Decision Discourse (Kumar Praveen, 2025) View paper
- Evaluation Frameworks and Benchmarking Methodologies
  - Adaptive Testing and Item Response Theory (3 papers)
  - [15] Adaptive Testing for LLM Evaluation: A Psychometric Alternative to Static Benchmarks (Li Pei-yu, 2025) View paper
  - [26] Darg: Dynamic evaluation of large language models via adaptive reasoning graph (Jiaao Chen, 2024) View paper
  - [35] The LLM Already Knows: Estimating LLM-Perceived Question Difficulty via Hidden Representations (Zhu Yu-bo, 2025) View paper
  - Zero-Shot and Automated Benchmark Generation (2 papers)
  - [7] Ufo: a unified and flexible framework for evaluating factuality of large language models (Huang, 2024) View paper
  - [11] Zero-shot Benchmarking: A Framework for Flexible and Scalable Automatic Evaluation of Language Models (Pombal, 2025) View paper
  - Confidence Calibration and Risk-Aware Evaluation (2 papers)
  - [4] Probabilistic medical predictions of large language models (Bowen Gu, 2024) View paper
  - [47] Overconfidence in LLM-as-a-Judge: Diagnosis and Confidence-Driven Solution (Zailong Tian, 2025) View paper
- Adaptive Model Optimization and Compression
  - Singular Value Decomposition and Matrix Factorization (1 papers)
  - [2] AdaSVD: Adaptive Singular Value Decomposition for Large Language Models (Li Zhiteng, 2025) View paper
  - Adaptive Layer-Wise and Unified Compression (1 papers)
  - [16] Edge-llm: Enabling efficient large language model adaptation on edge devices via unified compression and adaptive layer voting (Zhongzhi Yu, 2024) View paper
  - Flexible Sparsity and Compute-in-Memory Acceleration (1 papers)
  - [29] Accelerating llm inference with flexible n: M sparsity via a fully digital compute-in-memory accelerator (Kundu, 2025) View paper
- Memory Management and KV Cache Optimization
  - Paged Attention and Virtual Memory Techniques (2 papers)
  - [5] Efficient Memory Management for Large Language Model Serving with PagedAttention (Woosuk Kwon, 2023) View paper
  - [38] Open-AI model Efficient Memory Reduce Management for the Large Language Models (LLMs) Serving with Paged Attention of sharing the KV Cashes (Dr.K.Naveen Kumar, 2024) View paper
  - Dynamic KV Cache Pruning and Layer-Adaptive Management (1 papers)
  - [33] Lethe: Layer- and Time-Adaptive KV Cache Pruning for Reasoning-Intensive LLM Serving (Hui Zeng, 2025) View paper
- Adaptive Inference and Realignment Strategies
  - Self-Adaptive and Transformer-Based Adaptation (1 papers)
  - [40] Transformer-Squared: Self-adaptive LLMs (SUN Qi, 2025) View paper
  - Flexible Realignment and Controllable Fusion (1 papers)
  - [32] Flexible Realignment of Language Models (Zhu Wen-hong, 2025) View paper
  - CPU-Based and Heterogeneous Inference Optimization (1 papers)
  - [41] Flexinfer: Flexible llm inference with cpu computations (S Na, 2025) View paper
- Domain-Specific Adaptive Applications
  - Personalized Education and Curriculum Design (2 papers)
  - [28] Adaptive Learning Systems: Personalized Curriculum Design Using LLM-Powered Analytics (Li, 2025) View paper
  - [31] A Theory of Adaptive Scaffolding for LLM-Based Pedagogical Agents (Clayton Cohn, 2025) View paper
  - Healthcare and Clinical Decision Support (1 papers)
  - [43] eCBT-I dialogue system: a comparative evaluation of large language models and adaptation strategies for insomnia treatment. (Xueying Bao, 2025) View paper
  - Conversational Agents with Value Modeling (1 papers)
  - [34] Effectiveness of Conversational Robots Capable of Estimating and Modeling User Values (Y Sakamoto, 2025) View paper
- Specialized Adaptive Techniques
  - Adaptive Contextual Synthesis and Semantic Alignment (1 papers)
  - [13] Assessing semantic alignment in large language models through adaptive contextual synthesis (Pu Cha, 2024) View paper
  - Data Valuation and Shapley Value Estimation (2 papers)
  - [44] Document Valuation in LLM Summaries: A Cluster Shapley Approach (Zikun Ye, 2025) View paper
  - [50] Data Valuation for LLM Fine-Tuning: Efficient Shapley Value Approximation via Language Model Arithmetic (MÃ©lissa Tamine, 2025) View paper
  - Zero-Shot Time Series Forecasting (1 papers)
  - [20] Large Language Models Are Zero-Shot Time Series Forecasters (Gruver, 2023) View paper
  - Category-Agnostic Pose Estimation (1 papers)

    - [14] CapeLLM: Support-Free Category-Agnostic Pose Estimation with Multimodal Large Language Models (Kim Jun-Ho, 2024) View paper

Narrative

Core task: adaptive value evaluation of large language models. The field has grown into a rich landscape organized around several major branches. Value Alignment and Orientation Assessment focuses on measuring whether models reflect human values and cultural norms, often through benchmarking frameworks like Valuebench[6] and methods that assess semantic alignment or heterogeneous value orientations (Heterogeneous Value Alignment[1]). Reinforcement Learning and Value-Based Optimization explores how to train models using value functions and reward signals, including techniques like step-level Q-value estimation (Step-level Q-value[12]) and direct value optimization (Direct Value Optimization[25]). Adaptive Planning and Decision-Making Agents examines how models can dynamically adjust their reasoning strategies, as seen in works like Adaplanner[3]. Evaluation Frameworks and Benchmarking Methodologies provide systematic ways to measure model capabilities, while branches on Adaptive Model Optimization and Memory Management (e.g., PagedAttention[5]) address efficiency concerns. Adaptive Inference and Realignment Strategies, Domain-Specific Applications, and Specialized Techniques round out the taxonomy, covering context-dependent adjustments and targeted use cases.

A particularly active line of work centers on developing fine-grained value evaluation methods that can adapt to different contexts or user populations. AdAEM Value Difference[0] sits squarely within this cluster, proposing adaptive mechanisms to measure value differences across diverse settings. It shares thematic ground with AdAEM Measurement[27], which also emphasizes adaptive evaluation, and with Clave[9], another work in the same branch that explores context-sensitive value assessment. These methods contrast with more static benchmarking approaches like Valuebench[6] or zero-shot evaluation schemes (Zero-shot Benchmarking[11]), which apply uniform criteria across all scenarios. Meanwhile, reinforcement learning branches pursue value estimation for optimization rather than pure assessment, highlighting a trade-off between diagnostic measurement and performance improvement. The original paper's focus on adaptive value difference measurement positions it as a bridge between alignment assessment and dynamic evaluation, addressing the challenge of capturing how model values shift in response to varying inputs or populations.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Clave: An adaptive framework for evaluating values of llm generated responses

Authors: Yao Jing, Yi, Xiaoyuan, Jing Yao, Xie Xing, et al. (7 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

The rapid progress in Large Language Models (LLMs) poses potential risks such as generating unethical content. Assessing LLMs' values can help expose their misalignment, but relies on reference-free evaluators, e.g., fine-tuned LLMs or close-source ones like GPT-4, to identify values reflected in generated responses. Nevertheless, these evaluators face two challenges in open-ended value evaluation: they should align with changing human value definitions with minimal annotation, against their own...

Relationship Analysis

Both papers belong to the Adaptive Value Evaluation Methods category, focusing on self-extensible frameworks for assessing LLM values. They overlap in addressing the challenge of evaluating value differences in LLMs through adaptive approaches that go beyond static benchmarks. However, AdAEM focuses on automatically generating and refining test questions through an information-theoretic optimization process that probes value boundaries across diverse LLMs, while CLAVE addresses the evaluation challenge by combining complementary large and small LLMs through value concept extraction to adaptively assess values in LLM-generated responses with minimal annotation.

2. AdAEM: An Adaptively and Automated Extensible Measurement of LLMs' Value Difference

Authors: Yi, Xiaoyuan, Shitong Duan, Zhang, Peng, et al. (19 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Assessing Large Language Models (LLMs)' underlying value differences enables comprehensive comparison of their misalignment, cultural adaptability, and biases. Nevertheless, current value measurement datasets face the informativeness challenge: with often outdated, contaminated, or generic test questions, they can only capture the shared value orientations among different LLMs, leading to saturated and thus uninformative results. To address this problem, we introduce AdAEM, a novel, self-extensi...

⚠ Similarity Notice

These papers share nearly identical titles, abstracts, and core technical content describing the AdAEM framework for adaptive value evaluation of LLMs. Both present the same methodology involving information-theoretic optimization, EM-like iteration steps, and exploration algorithms to generate value-evoking questions. This appears to be the same paper or a very close variant (possibly a preprint vs. conference submission version).

Contributions Analysis

Overall novelty summary. The paper introduces AdAEM, a self-extensible algorithm for evaluating value differences across LLMs by dynamically generating test questions through in-context optimization. It resides in the 'Adaptive Value Evaluation Methods' leaf, which contains only three papers total, making this a relatively sparse research direction within the broader taxonomy. This leaf explicitly excludes static benchmarks with fixed question sets, positioning AdAEM as part of an emerging cluster focused on dynamic, context-sensitive value measurement rather than traditional psychometric approaches.

The taxonomy reveals that AdAEM's immediate neighbors include static 'Value Measurement Benchmarks' (e.g., Valuebench) and 'Heterogeneous Value Alignment' frameworks assessing multiple conflicting objectives. Nearby branches address reinforcement learning-based value optimization and behavioral consistency checks, but these focus on training-time alignment or action validation rather than adaptive diagnostic measurement. The scope notes clarify that AdAEM's dynamic question generation distinguishes it from fixed-item psychometric tools, while its focus on value orientation assessment separates it from optimization-focused RL methods.

Among 26 candidates examined, each of AdAEM's three contributions shows at least one refutable candidate. Contribution A (the core algorithm) examined 9 papers with 1 potential refutation; Contribution B (information-theoretic objective) examined 7 with 1 refutation; Contribution C (AdAEM Bench) examined 10 with 1 refutation. The statistics suggest that within this limited search scope, each contribution encounters some overlapping prior work, though the majority of examined candidates (23 of 26 total) do not clearly refute the claims. The sparse leaf structure and modest refutation counts indicate moderate novelty relative to the examined literature.

Based on top-26 semantic matches, AdAEM appears to occupy a less-crowded niche within value evaluation, though the limited search scope and presence of refutable candidates for all contributions suggest caution. The analysis captures adaptive value measurement methods but does not exhaustively cover static benchmarking or optimization-focused RL literature, which may contain additional relevant comparisons. The taxonomy structure confirms that dynamic, self-extensible evaluation remains an emerging area with fewer established precedents than static assessment frameworks.

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

Contribution 1: AdAEM: A self-extensible dynamic value evaluation algorithm

Description: The authors propose AdAEM, an automated framework that dynamically generates and extends test questions to evaluate LLMs' value orientations. Unlike static benchmarks, AdAEM probes value boundaries across diverse LLMs through in-context optimization, enabling it to co-evolve with LLM development and consistently track value dynamics.

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. Exploring Conversational Adaptability: Assessing the Proficiency of Large Language Models in Dynamic Alignment with Updated User Intent

URL: View paper

Brief Assessment

Conversational Adaptability[65] focuses on dialogue systems adapting to changing user intentions in conversations, not on dynamic value evaluation methods for language models. The technical domains are fundamentally different.

2. Capturing nuanced preferences: Preference-aligned distillation for small language models

URL: View paper

Brief Assessment

Preference-aligned Distillation[62] focuses on distilling preference knowledge from large language models to small language models for alignment purposes, not on dynamic evaluation methods for value orientations. The candidate addresses model compression and preference learning, while the original contribution concerns automated value assessment frameworks.

3. Value compass benchmarks: A comprehensive, generative and self-evolving platform for llms' value evaluation

URL: View paper

Prior Art Analysis

Value Compass Benchmarks[39] demonstrates that similar prior work exists for dynamic, self-evolving value evaluation methods. The candidate paper presents a 'generative self-evolving evaluation paradigm' that automatically generates and periodically refines test items tailored for evolving LLM capabilities, using an optimization framework similar to AdAEM's approach. Both methods employ in-context optimization to dynamically generate value-evoking questions that co-evolve with LLM development, addressing data contamination and ceiling effects through adaptive item generation.

Evidence

Evidence 1 - **Rationale**: Both papers describe dynamic evaluation methods that automatically generate and refine test items. The candidate's 'generative self-evolving evaluation paradigm' directly parallels AdAEM's 'self-creates and self-extends its test questions' approach, suggesting prior work on this concept. - **Original**: we introduce adaem, a novel value evaluation algorithm. distinct from previous static datasets (zhang et al., 2023b), following the dynamic evaluation schema (bai et al., 2023b; zhu et al., 2023), adaem automatically self-creates and self-extends its test questions by exploring the underlying value ... - **Candidate**: our benchmarks adopt a novel generative selfevolving evaluation paradigm (duan et al., 2025), which automatically generates and periodically refines test items tailored for evolving llm capabilities and deciphers values in a generative manner

Evidence 2 - **Rationale**: Both methods use optimization objectives to generate value-evoking questions. The candidate's optimization framework for item generation closely mirrors AdAEM's 'iteratively optimizing an information-theoretic objective,' indicating that the core algorithmic approach was not novel to the original paper. - **Original**: concretely, adaem produces such questions by iteratively optimizing an information-theoretic objective in an in-context manner without any manually curated data or fine-tuning. then valueevoking test questions, which are on the value boundaries of different llms, can be adaptively exploited leveragi... - **Candidate**: this is achieved by optimizing an item generator, qθ(x), parameterized by θ, via: θ∗= argmax θ ex~qθ(x){(1 -α) d[p1(v|x),...,p m (v|x)] informativeness maximization +αevep~pip(v,y|x)} value elicitation

Evidence 3 - **Rationale**: Both systems explicitly describe co-evolution with LLM development. The candidate's mechanism of updating test items when new LLMs are released demonstrates the same self-extensible property claimed as novel in AdAEM. - **Original**: in this way, adaem can continuously refine questions and co-evolve with the development of llms, fostering better comparison of their misalignment and cultural biases - **Candidate**: once an llm is updated or newly released, we update the llm set pand re-execute eq.(1) to generate new test items, keeping pace with llms' development. thus, our benchmarks can co-evolve with llms, consistently providing informative assessments to reveal their nuanced differences

Evidence 4 - **Rationale**: Both methods address data contamination and cultural diversity through their optimization frameworks. The candidate's approach to capturing value differences and generating culturally diverse items parallels AdAEM's claimed innovations. - **Original**: when integrated with the latest llms, adaem extracts more recent social issues not yet memorized by most models, mitigating data contamination; when applied to those from different cultures, adaem explores culturally diverse topics, avoiding indistinguishable evaluation results - **Candidate**: this is achieved by optimizing an item generator, qθ(x), parameterized by θ... the first term in eq.(1) exploits x that maximally captures value differences of llms (e.g., the cultural ones, see fig. 5), while the second constrains xto be value-evoking rather than neutral

4. Localvaluebench: A collaboratively built and extensible benchmark for evaluating localized value alignment and ethical safety in large language models

URL: View paper

Brief Assessment

Localvaluebench[59] focuses on creating localized benchmarks for specific jurisdictions (e.g., Australian values) through manual question curation and human review processes, rather than proposing an automated, self-extensible algorithm for dynamic question generation like AdAEM.

5. Benchmarking multi-national value alignment for large language models

URL: View paper

Brief Assessment

Multi-national Value Alignment[60] focuses on extracting national values from official media sources across countries to evaluate LLM alignment with specific national perspectives, rather than proposing a general dynamic value evaluation framework that co-evolves with LLM development.

6. Gradient-Adaptive Policy Optimization: Towards Multi-Objective Alignment of Large Language Models

URL: View paper

Brief Assessment

Gradient-Adaptive Policy[61] focuses on multi-objective optimization for LLM alignment through gradient-based fine-tuning, not on dynamic evaluation methods for assessing value orientations.

7. Dynamic Rewarding with Prompt Optimization Enables Tuning-free Self-Alignment of Language Models

URL: View paper

Brief Assessment

Dynamic Rewarding[63] focuses on tuning-free self-alignment of language models through prompt optimization and dynamic rewarding mechanisms, not on value evaluation or measurement of value orientations across diverse LLMs.

8. Do Language Models Think Consistently? A Study of Value Preferences Across Varying Response Lengths

URL: View paper

Brief Assessment

Value Preferences Consistency[64] focuses on consistency of value preferences across short-form vs. long-form responses of varying lengths, not on dynamic benchmark generation or self-extensible evaluation frameworks for value orientations.

9. Towards understanding valuable preference data for large language model alignment

URL: View paper

Brief Assessment

Valuable Preference Data[66] focuses on preference data selection for LLM alignment using influence functions, not on dynamic value evaluation methods. The candidate addresses data quality for RLHF training, while the original proposes a self-extensible framework for evaluating LLMs' value orientations through adaptive question generation.

Contribution 2: Information-theoretic optimization objective for maximizing value differences

Description: The authors formalize an information-theoretic optimization objective that guides the generation of test questions to maximize distinguishability and disentanglement of value orientations across different LLMs. This objective addresses the informativeness challenge by extracting controversial topics that reveal genuine value differences rather than shared safety values.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. AdAEM: An Adaptively and Automated Extensible Measurement of LLMs' Value Difference

URL: View paper

Prior Art Analysis

AdAEM Measurement[27] demonstrates prior work on information-theoretic objectives for generating distinguishable test questions. Both papers formalize optimization objectives using Jensen-Shannon divergence to maximize distinguishability of value orientations across different LLMs. The candidate paper presents the same core mathematical framework (Equations 1-2) with identical components: using JSD to measure separability among value distributions, incorporating disentanglement terms, and optimizing questions to elicit distinguishable responses. The candidate's approach of 'maximizing an information-theoretic objective to extract diverse controversial topics that can provide more distinguishable and informative insights about models' value differences' directly parallels the original's contribution.

Evidence

Evidence 1 - **Rationale**: Both papers explicitly state the same goal of using the optimization objective to address the informativeness challenge by exposing distinguishable value differences. - **Original**: maximizing eq.(1) helps identify x that better exposes llms' own value differences, handling the informativeness challenge - **Candidate**: to tackle the informativeness challenge, we require x to be able to expose sufficiently distinguishable instead of saturated results vi ∼ pθi(v|x) for different llms, so as to provide more meaningful insights for comparing various value-based properties of llms

2. Entropy-based experimental design for optimal model discrimination in the geosciences

URL: View paper

Brief Assessment

Entropy Model Discrimination[71] focuses on optimal experimental design for model selection in geosciences using mutual information to discriminate between competing physical models (e.g., sorption isotherms). The ORIGINAL paper addresses value orientation measurement in LLMs using information-theoretic objectives to generate distinguishable test questions. These are fundamentally different application domains with distinct technical goals.

3. An Entropy-Driven Method for LLM Dataset Evaluation And Optimization

URL: View paper

Brief Assessment

Entropy-Driven Dataset Evaluation[70] focuses on dataset quality assessment using entropy-based metrics for question discrimination, not on generating test questions to maximize distinguishability of value orientations across LLMs through information-theoretic objectives.

4. Feature selection by utilizing kernel-based fuzzy rough set and entropy-based non-dominated sorting genetic algorithm in multi-label data

URL: View paper

Brief Assessment

Kernel Fuzzy Entropy[68] focuses on feature selection in multi-label data using entropy-based methods for identifying discriminative features, not on generating test questions to distinguish model behaviors or value orientations.

5. Separation and the information theory surrogate evaluation approach: A penalised likelihood solution.

URL: View paper

Brief Assessment

The candidate paper (Penalised Likelihood Surrogate[72]) addresses separation in statistical models using penalised likelihood methods, which is unrelated to generating test questions for LLM value assessment or information-theoretic objectives for distinguishability.

6. Cognitive constraints in bilingual processingâ an entropy-based discrimination between translation and second language production

URL: View paper

Brief Assessment

Bilingual Entropy Discrimination[67] uses entropy to measure cognitive load in bilingual text production (translation vs. L2), not to generate distinguishable test questions across LLMs or maximize value orientation differences.

7. The role of entropy in construct specification equations (CSE) to improve the validity of memory tests

URL: View paper

Brief Assessment

The candidate paper applies entropy theory to memory test construction and item difficulty prediction, not to generating distinguishable test questions across different AI models or measuring value orientations in LLMs.

Contribution 3: AdAEM Bench: A novel value evaluation benchmark

Description: The authors construct AdAEM Bench, a benchmark dataset containing 12,310 value-evoking questions generated using their framework. This benchmark is grounded in Schwartz's Theory of Basic Values and demonstrates superior semantic diversity, novelty, and ability to elicit distinguishable value orientations compared to existing static benchmarks.

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. â¦ the alignment of large language models with human values for mental health integration: cross-sectional study using Schwartz's theory of basic values

URL: View paper

Brief Assessment

Mental Health Alignment[52] applies Schwartz's Theory to mental health integration contexts, not to constructing comprehensive value evaluation benchmarks for LLMs with automated question generation.

2. Value compass benchmarks: A comprehensive, generative and self-evolving platform for llms' value evaluation

URL: View paper

Brief Assessment

Value Compass Benchmarks[39] focuses on building an online platform with multiple value systems and interactive features, rather than constructing a specific benchmark dataset grounded in Schwartz's Theory. The candidate's emphasis is on platform functionality and multi-faceted interpretation, not on benchmark dataset construction methodology.

3. AdAEM: An Adaptively and Automated Extensible Measurement of LLMs' Value Difference

URL: View paper

Prior Art Analysis

AdAEM Measurement[27] presents a benchmark dataset constructed using their framework that is grounded in Schwartz's Theory of Basic Values. The candidate paper describes 'adaem bench' with 12,310 questions instantiated with Schwartz's theory, demonstrating superior semantic diversity and ability to elicit distinguishable value orientations. Both benchmarks use the same theoretical foundation (Schwartz's 10 value dimensions), similar construction methodology (automated generation from generic topics), and serve the same purpose of addressing the informativeness challenge in value evaluation.

Evidence

Evidence 1 - **Rationale**: Both papers describe creating a dataset of questions grounded in social science value theories using the AdAEM framework, with the same scale (12,310 questions) and purpose. - **Original**: using adaem, we create a dataset of informative evaluation questions grounded in value theories from social science, analyzing and validating adaem's effectiveness - **Candidate**: using adaem, we create a large-scale dataset consisting of 12,310 questions grounded in crossculture schwartz value theory [44] from psychology, and benchmark as well as analyze the value orientations of 16 popular llms, manifesting adaem's superiority over previous benchmarks

Evidence 2 - **Rationale**: The text describing the benchmark construction process and its properties is nearly identical, including the exact number of questions (12,310) and the same stated goals. - **Original**: through this process, we obtained 12,310 evoking questions, x, which help prevent data contamination and exposevalue difference, tackling theinformativeness challengediscussed in sec. 1 - **Candidate**: through this process, we obtained 12,310 evoking test questions, x, rooted in controversial social issues, which help prevent data contamination and expose value difference, handling the informativeness challenge discussed in sec. 1

4. Generative Psycho-Lexical Approach for Constructing Value Systems in Large Language Models

URL: View paper

Brief Assessment

Generative Psycho-Lexical[58] focuses on constructing a new five-factor value system for LLMs using a psycho-lexical approach, rather than creating a benchmark dataset for evaluating values using Schwartz's Theory. The candidate does not present a benchmark comparable to AdAEM Bench.

5. Value FULCRA: Mapping Large Language Models to the Multidimensional Spectrum of Basic Human Value

URL: View paper

Brief Assessment

Value FULCRA[55] focuses on mapping LLM outputs to a multidimensional value space using Schwartz's theory, constructing a dataset of 20k (LLM output, value vector) pairs for alignment purposes. This differs from AdAEM Bench's emphasis on generating 12,310 value-evoking questions through adaptive optimization to elicit distinguishable value orientations across models.

6. Understanding how value neurons shape the generation of specified values in llms

URL: View paper

Brief Assessment

Value Neurons[56] focuses on mechanistic interpretability of value-encoding neurons within LLMs, not on constructing value evaluation benchmarks. The candidate paper introduces ValueInsight, a dataset for neuron identification rather than comprehensive value evaluation across models.

7. Assessing the Alignment of Large Language Models With Human Values for Mental Health Integration: Cross-Sectional Study Using Schwartzâs Theory of Basic Values

URL: View paper

Brief Assessment

Schwartz Basic Values[53] focuses on applying Schwartz's Theory to evaluate LLMs using the existing Portrait Values Questionnaire-Revised (PVQ-RR), not on constructing a novel benchmark dataset. The candidate uses a static questionnaire approach rather than the dynamic, adaptive question generation framework proposed in the original paper.

8. Measuring value understanding in language models through discriminator-critique gap

URL: View paper

Brief Assessment

Discriminator-Critique Gap[57] focuses on measuring value understanding through the gap between discriminator and critique scores, not on constructing a benchmark dataset. The candidate evaluates whether LLMs understand 'know what' and 'know why' aspects of values, while the original contribution is about creating a diverse benchmark dataset grounded in Schwartz's Theory.

9. Assessing the alignment of large language models with human values for mental health integration: cross-sectional study using Schwartz's theory of basic â¦

URL: View paper

Brief Assessment

Mental Health Values[51] focuses on applying Schwartz's theory to mental health integration contexts, not on constructing a comprehensive value evaluation benchmark for LLMs with automated question generation.

10. Structured Moral Reasoning in Language Models: A Value-Grounded Evaluation Framework

URL: View paper

Brief Assessment

Structured Moral Reasoning[54] focuses on evaluating moral reasoning through prompting strategies and distillation, not on constructing a value evaluation benchmark dataset. The candidate does not present a competing benchmark to AdAEM Bench.

Appendix: Text Similarity Detection

Textual similarity detection checked 26 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. AdAEM: An Adaptively and Automated Extensible Measurement of LLMs' Value Difference

Detected in: Core Task (sibling), Contribution: contribution_2, Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] AdAEM: An Adaptively and Automated Extensible Evaluation Method of LLMs' Value Difference View paper
- [1] Heterogeneous value alignment evaluation for large language models View paper
- [2] AdaSVD: Adaptive Singular Value Decomposition for Large Language Models View paper
- [3] Adaplanner: Adaptive planning from feedback with language models View paper
- [4] Probabilistic medical predictions of large language models View paper
- [5] Efficient Memory Management for Large Language Model Serving with PagedAttention View paper
- [6] Valuebench: Towards comprehensively evaluating value orientations and understanding of large language models View paper
- [7] Ufo: a unified and flexible framework for evaluating factuality of large language models View paper
- [8] Segment Policy Optimization: Effective Segment-Level Credit Assignment in RL for Large Language Models View paper
- [9] Clave: An adaptive framework for evaluating values of llm generated responses View paper
- [10] G-Sim: Generative Simulations with Large Language Models and Gradient-Free Calibration View paper
- [11] Zero-shot Benchmarking: A Framework for Flexible and Scalable Automatic Evaluation of Language Models View paper
- [12] Enhancing decision-making for llm agents via step-level q-value models View paper
- [13] Assessing semantic alignment in large language models through adaptive contextual synthesis View paper
- [14] CapeLLM: Support-Free Category-Agnostic Pose Estimation with Multimodal Large Language Models View paper
- [15] Adaptive Testing for LLM Evaluation: A Psychometric Alternative to Static Benchmarks View paper
- [16] Edge-llm: Enabling efficient large language model adaptation on edge devices via unified compression and adaptive layer voting View paper
- [17] MetaAligner: Towards Generalizable Multi-Objective Alignment of Language Models View paper
- [18] Adaptation with self-evaluation to improve selective prediction in llms View paper
- [19] Token-Supervised Value Models for Enhancing Mathematical Reasoning Capabilities of Large Language Models View paper
- [20] Large Language Models Are Zero-Shot Time Series Forecasters View paper
- [21] Language models can self-improve at state-value estimation for better search View paper
- [22] Heterogeneous Value Evaluation for Large Language Models View paper
- [23] Value augmented sampling for language model alignment and personalization View paper
- [24] Inverse-Q*: Token Level Reinforcement Learning for Aligning Large Language Models Without Preference Data View paper
- [25] Direct value optimization: Improving chain-of-thought reasoning in llms with refined values View paper
- [26] Darg: Dynamic evaluation of large language models via adaptive reasoning graph View paper
- [27] AdAEM: An Adaptively and Automated Extensible Measurement of LLMs' Value Difference View paper
- [28] Adaptive Learning Systems: Personalized Curriculum Design Using LLM-Powered Analytics View paper
- [29] Accelerating llm inference with flexible n: M sparsity via a fully digital compute-in-memory accelerator View paper
- [30] MAVIS: Multi-Objective Alignment via Value-Guided Inference-Time Search View paper
- [31] A Theory of Adaptive Scaffolding for LLM-Based Pedagogical Agents View paper
- [32] Flexible Realignment of Language Models View paper
- [33] Lethe: Layer- and Time-Adaptive KV Cache Pruning for Reasoning-Intensive LLM Serving View paper
- [34] Effectiveness of Conversational Robots Capable of Estimating and Modeling User Values View paper
- [35] The LLM Already Knows: Estimating LLM-Perceived Question Difficulty via Hidden Representations View paper
- [36] Process for adapting language models to society (palms) with values-targeted datasets View paper

- [37] Mind the Value-Action Gap: Do LLMs Act in Alignment with Their Values? View paper
- [38] Open-AI model Efficient Memory Reduce Management for the Large Language Models (LLMs) Serving with Paged Attention of sharing the KV Cashes View paper
- [39] Value compass benchmarks: A comprehensive, generative and self-evolving platform for llms' value evaluation View paper
- [40] Transformer-Squared: Self-adaptive LLMs View paper
- [41] Flexinfer: Flexible llm inference with cpu computations View paper
- [42] SWE-Search: Enhancing Software Agents with Monte Carlo Tree Search and Iterative Refinement View paper
- [43] eCBT-I dialogue system: a comparative evaluation of large language models and adaptation strategies for insomnia treatment. View paper
- [44] Document Valuation in LLM Summaries: A Cluster Shapley Approach View paper
- [45] Value Compass Benchmarks: A Platform for Fundamental and Validated Evaluation of LLMs Values View paper
- [46] Agentic LLM Framework for Adaptive Decision Discourse View paper
- [47] Overconfidence in LLM-as-a-Judge: Diagnosis and Confidence-Driven Solution View paper
- [48] Inference-time language model alignment via integrated value guidance View paper
- [49] Memory-Driven Self-Improvement for Decision Making with Large Language Models View paper
- [50] Data Valuation for LLM Fine-Tuning: Efficient Shapley Value Approximation via Language Model Arithmetic View paper
- [51] Assessing the alignment of large language models with human values for mental health integration: cross-sectional study using Schwartz's theory of basic â¦ View paper
- [52] â¦ the alignment of large language models with human values for mental health integration: cross-sectional study using Schwartz's theory of basic values View paper
- [53] Assessing the Alignment of Large Language Models With Human Values for Mental Health Integration: Cross-Sectional Study Using Schwartzâs Theory of Basic Values View paper
- [54] Structured Moral Reasoning in Language Models: A Value-Grounded Evaluation Framework View paper
- [55] Value FULCRA: Mapping Large Language Models to the Multidimensional Spectrum of Basic Human Value View paper
- [56] Understanding how value neurons shape the generation of specified values in llms View paper
- [57] Measuring value understanding in language models through discriminator-critique gap View paper
- [58] Generative Psycho-Lexical Approach for Constructing Value Systems in Large Language Models View paper
- [59] Localvaluebench: A collaboratively built and extensible benchmark for evaluating localized value alignment and ethical safety in large language models View paper
- [60] Benchmarking multi-national value alignment for large language models View paper
- [61] Gradient-Adaptive Policy Optimization: Towards Multi-Objective Alignment of Large Language Models View paper
- [62] Capturing nuanced preferences: Preference-aligned distillation for small language models View paper
- [63] Dynamic Rewarding with Prompt Optimization Enables Tuning-free Self-Alignment of Language Models View paper
- [64] Do Language Models Think Consistently? A Study of Value Preferences Across Varying Response Lengths View paper
- [65] Exploring Conversational Adaptability: Assessing the Proficiency of Large Language Models in Dynamic Alignment with Updated User Intent View paper
- [66] Towards understanding valuable preference data for large language model alignment View paper
- [67] Cognitive constraints in bilingual processingâan entropy-based discrimination between translation and second language production View paper
- [68] Feature selection by utilizing kernel-based fuzzy rough set and entropy-based non-dominated sorting genetic algorithm in multi-label data View paper
- [69] The role of entropy in construct specification equations (CSE) to improve the validity of memory tests View paper
- [70] An Entropy-Driven Method for LLM Dataset Evaluation And Optimization View paper
- [71] Entropy-based experimental design for optimal model discrimination in the geosciences View paper
- [72] Separation and the information theory surrogate evaluation approach: A penalised likelihood solution. View paper