Novelty Assessment Report

Paper: Using cognitive models to reveal value trade-offs in language models

PDF URL: https://openreview.net/pdf?id=nM2QhvybwI

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

Value trade-offs are an integral part of human decision-making and language use, however, current tools for interpreting such dynamic and multi-faceted notions of values in LLMs are limited. In cognitive science, so-called “cognitive models” provide formal accounts of such trade-offs in humans, by modeling the weighting of a speaker’s competing utility functions in choosing an action or utterance. Here we use a leading cognitive model of polite speech to systematically evaluate value trade-offs in two encompassing model settings: degrees of reasoning “effort” in frontier black-box models, and RL post-training dynamics of open-source models. Our results highlight patterns of higher informational utility than social utility in reasoning models’ default behavior, and demonstrate that these patterns shift in predictable ways when models are prompted to prioritize certain goals over others. Our findings from LLMs’ training dynamics suggest large shifts in utility values early on in training with persistent effects of the choice of base model and pretraining data, compared to feedback dataset or alignment method. We show that our method is responsive to diverse aspects of the rapidly evolving LLM landscape, with insights for forming hypotheses about other social behaviors such as sycophancy, and shaping training regimes that better control trade-offs between values during model development

Core Task Landscape

This paper addresses: **Evaluating Value Trade-offs in Language Model Behavior**

A total of **50 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

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

- **Value Alignment Assessment and Measurement**
- **Behavioral Trade-off Analysis in LLM Systems**
- **Alignment Interventions and Training Dynamics**
- **Robustness and Adversarial Challenges in Value Alignment**
- **Application-Specific Value Alignment**
- **Technical Infrastructure and Methodological Foundations**

Complete Taxonomy Tree

- Evaluating Value Trade-offs in Language Model Behavior Survey Taxonomy
- Value Alignment Assessment and Measurement
  - Value Representation and Prioritization Frameworks (6 papers)
  - [6] Strong and weak alignment of large language models with human values (Khamassi, 2024) View paper
  - [9] Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts (Wang Hao-xiang, 2024) View paper
  - [15] Human Value Alignment in AI (Ilias O. Pappas, 2025) View paper
  - [29] Unpacking the ethical value alignment in big models (Yi, 2023) View paper
  - [43] Plurality of value pluralism and AI value alignment (Kasirzadeh, 2024) View paper
  - [45] Prism: Perspective reasoning for integrated synthesis and mediation as a multi-perspective framework for ai alignment (Diamond, 2025) View paper
  - Value Conflict Benchmarks and Dilemma Evaluation (6 papers)
  - [1] Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark (Pan, 2023) View paper
  - [25] Dailydilemmas: Revealing value preferences of llms with quandaries of daily life (Yu Ying Chiu, 2024) View paper
  - [36] The Staircase of Ethics: Probing LLM Value Priorities through Multi-Step Induction to Complex Moral Dilemmas (WU Ya, 2025) View paper
  - [39] CLASH: Evaluating Language Models on Judging High-Stakes Dilemmas from Multiple Perspectives (Lee Aâ□□Young, 2025) View paper
  - [41] Generative Value Conflicts Reveal LLM Priorities (Liu, 2025) View paper
  - [47] LLMs as mediators: Can they diagnose conflicts accurately? (Ã□zgecan KoÃ§ak, 2024) View paper
  - Cultural Value Assessment and Cross-Cultural Alignment (7 papers)
  - [5] Towards realistic evaluation of cultural value alignment in large language models: Diversity enhancement for survey response simulation (Haijiang Liu, 2025) View paper
  - [23] Assessing llms for moral value pluralism (Noam Benkler, 2023) View paper
  - [26] Assessing the alignment of large language models with human values for mental health integration: cross-sectional study using Schwartz's theory of basic â□¦ (D Hadar-Shoval, 2024) View paper
  - [28] Exploring Multilingual Concepts of Human Values in Large Language Models: Is Value Alignment Consistent, Transferable and Controllable across Languages? (Shaoyang Xu, 2024) View paper
  - [31] The Ghost in the Machine has an American accent: value conflict in GPT-3 (Johnson, 2022) View paper

- [32] Cultural Value Alignment in Large Language Models: A Prompt-based Analysis of Schwartz Values in Gemini, ChatGPT, and DeepSeek (Segerer, 2025) View paper
- [44] CVC: A Large-Scale Chinese Value Rule Corpus for Value Alignment of Large Language Models (Wu Ping, 2025) View paper
- Value-Action Consistency and Behavioral Fidelity (5 papers)
- [14] Valuing time in silicon: Can large language models replicate human value of travel time (Yingnan Yan, 2025) View paper
- [19] Valuecompass: A framework for measuring contextual value alignment between human and llms (Hua Shen, 2025) View paper
- [21] Mind the Value-Action Gap: Do LLMs Act in Alignment with Their Values? (Shen Hua, 2025) View paper
- [30] What's the most important value? INVP: INvestigating the Value Priorities of LLMs through Decision-making in Social Scenarios (X Liu, 2025) View paper
- [33] LLM ethics benchmark: a three-dimensional assessment system for evaluating moral reasoning in large language models (Junfeng Jiao, 2025) View paper
- Moral Reasoning Structure and Ethical Theory Grounding (3 papers)
- [16] Structured Moral Reasoning in Language Models: A Value-Grounded Evaluation Framework (Chakraborty, 2025) View paper
- [18] Value FULCRA: Mapping large language models to the multidimensional spectrum of basic human value (Jing Yao, 2024) View paper
- [49] Prosocial behavior in Large Language Models: Value alignment and affective mechanisms (Hao Liu, 2025) View paper

- Behavioral Trade-off Analysis in LLM Systems
  - Cognitive Model-Based Value Trade-off Interpretation ★ (2 papers)
  - [0] Using cognitive models to reveal value trade-offs in language models (Anon et al., 2026) View paper
  - [38] Inside you are many wolves: Using cognitive models to interpret value trade-offs in LLMs (SK Murthy, 2025) View paper
  - Safety-Capability Trade-offs in Fine-tuning (3 papers)
  - [7] Fundamental Safety-Capability Trade-offs in Fine-tuning Large Language Models (Chen, 2025) View paper
  - [13] XGUARD: A Graded Benchmark for Evaluating Safety Failures of Large Language Models on Extremist Content (Vadivel Abishethvarman, 2025) View paper
  - [24] Safeguarding large language models in real-time with tunable safety-performance trade-offs (Fonseca, 2025) View paper
  - Accuracy-Fairness Trade-offs (1 papers)
  - [3] Exploring accuracy-fairness trade-off in large language models (Zhang Qing-quan, 2024) View paper
  - Privacy-Utility-Efficiency Trade-offs (2 papers)
  - [11] De-amplifying Bias from Differential Privacy in Language Model Fine-tuning (Srivastava, 2024) View paper
  - [12] Revisiting Privacy, Utility, and Efficiency Trade-offs when Fine-Tuning Large Language Models (Das, 2025) View paper
  - Watermarking Performance Trade-offs (3 papers)
  - [10] Performance trade-offs of watermarking large language models (Anirudh Ajith, 2023) View paper
  - [40] Can you Finetune your Binoculars? Embedding Text Watermarks into the Weights of Large Language Models (Ajroldi, 2025) View paper
  - [50] WaterPool: A Watermark Mitigating Trade-offs among Imperceptibility, Efficacy and Robustness (Huang, 2024) View paper
  - Hallucination-Mode Collapse Trade-offs (1 papers)
  - [4] On the Limits of Language Generation: Trade-Offs between Hallucination and Mode-Collapse (Alkis Kalavasis, 2025) View paper
  - Economic and Deployment Trade-offs (1 papers)
  - [17] The economic trade-offs of large language models: A case study (Howell Kristen, 2023) View paper
- Alignment Interventions and Training Dynamics
  - Activation Steering and Interpretable Control Methods (1 papers)
  - [42] Interpretable Steering of Large Language Models with Feature Guided Activation Additions (Guang Chen, 2025) View paper
  - Alignment Training Dynamics and Post-Training Effects (1 papers)
  - [20] One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity (Tomer Ullman, 2024) View paper
- Robustness and Adversarial Challenges in Value Alignment
  - Adversarial Attacks and Jailbreaking (1 papers)
  - [35] Sampling-aware Adversarial Attacks Against Large Language Models (Beyer, 2025) View paper
  - Knowledge Conflicts and Inconsistency (1 papers)
  - [34] ConflictBank: A Benchmark for Evaluating the Influence of Knowledge Conflicts in LLM (Su, 2024) View paper
  - Hallucination Detection and Mitigation (1 papers)
  - [2] Context-Aligned and Evidence-Based Detection of Hallucinations in Large Language Model Outputs (Peng, 2025) View paper
- Application-Specific Value Alignment (3 papers)
  - [27] Application-Driven Value Alignment in Agentic AI Systems: Survey and Perspectives (W Zeng, 2025) View paper
  - [46] " There are no solutions, only trade-offs."Taking A Closer Look At Safety Data Annotations. (EM Yang, 2024) View paper
  - [48] Training A Small Emotional Vision Language Model for Visual Art Comprehension (Zhang Jing, 2024) View paper
- Technical Infrastructure and Methodological Foundations (3 papers)
  - [8] Structural Embedding Projection for Contextual Large Language Model Inference (Vincent Enoasmo, 2025) View paper
  - [22] Watermarking for Large Language Models: A Survey (Zhiguang Yang, 2025) View paper
  - [37] Multimodal Mamba: Decoder-only Multimodal State Space Model via Quadratic to Linear Distillation (Liao, 2025) View paper

Narrative

Core task: evaluating value trade-offs in language model behavior. The field has organized itself into several major branches that reflect different facets of alignment research. Value Alignment Assessment and Measurement focuses on benchmarking and quantifying how well models reflect human values, often through structured evaluations and cultural or ethical probes (e.g., Cultural Value Alignment[5], ValueCompass[19]). Behavioral Trade-off Analysis examines the inherent tensions that arise when models must balance competing objectives—such as accuracy versus fairness (Accuracy Fairness Tradeoff[3]), safety versus capability (Safety Capability Tradeoffs[7]), or privacy versus utility (Privacy Utility Efficiency[12]). Alignment Interventions and Training Dynamics investigates how different training regimes and fine-tuning strategies shape value priorities, while Robustness and Adversarial Challenges explores how alignment holds up under attack or distributional shift. Application-Specific Value Alignment tailors these concerns to domains like mental health or agentic systems, and Technical Infrastructure provides the methodological scaffolding—datasets, metrics, and frameworks—that underpin empirical work across all branches.

Within Behavioral Trade-off Analysis, a particularly active line of work examines how models navigate conflicting values in realistic decision scenarios, often drawing on moral dilemmas (DailyDilemmas[25], CLASH Dilemmas[39]) or game-theoretic settings (Machiavelli

Rewards Ethics[1]). Another strand investigates technical trade-offs such as watermarking's impact on generation quality (Watermarking Performance Tradeoffs[10]) or the interplay between safety constraints and model expressiveness (Tunable Safety Performance[24]). Cognitive Value Tradeoffs[0] sits within the Cognitive Model-Based Value Trade-off Interpretation cluster, emphasizing how cognitive frameworks can illuminate the internal reasoning processes that lead to particular trade-off resolutions. This approach contrasts with purely behavioral or outcome-focused methods, offering a more mechanistic lens on why models prioritize certain values over others. Nearby work like Cognitive Wolves[38] similarly explores cognitive architectures, suggesting a small but growing interest in interpretability-driven accounts of value conflict resolution.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Inside you are many wolves: Using cognitive models to interpret value trade-offs in LLMs

Authors: SK Murthy, R Zhao, J Hu, S Kakade | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ These value trade-offs are an integral part of human decision-making and language use, â¦ We apply this lens to systematically evaluate value trade-offs in two encompassing model â¦

⚠ Similarity Notice

These papers appear to be the same work or very close variants. Both have nearly identical titles ('Using cognitive models to reveal value trade-offs in language models' vs 'Inside you are many wolves: Using cognitive models to interpret value trade-offs in LLMs'), identical abstracts, the same cognitive model framework (Yoon et al. 2020's polite speech model), identical experimental design with closed-source and open-source model suites, and the same core findings about informational vs social utility trade-offs in reasoning models. The only notable difference is the subtitle in the candidate paper.

Contributions Analysis

Overall novelty summary. The paper applies cognitive models from human decision-making research to interpret value trade-offs in LLMs, specifically using a politeness model to quantify informational versus social utility. It resides in the Cognitive Model-Based Value Trade-off Interpretation leaf, which contains only two papers total. This represents a sparse research direction within the broader Behavioral Trade-off Analysis branch, suggesting the cognitive modeling approach to LLM value alignment is relatively underexplored compared to empirical benchmarking or technical intervention methods that dominate neighboring areas.

The taxonomy reveals that most behavioral trade-off work focuses on specific technical tensions—safety versus capability, accuracy versus fairness, privacy versus utility—rather than cognitive frameworks for interpreting multi-dimensional value conflicts. The paper's leaf sits alongside general Behavioral Trade-off Analysis but diverges from purely outcome-focused evaluations by emphasizing mechanistic accounts of how models weight competing utilities. Neighboring leaves like Safety-Capability Trade-offs and Accuracy-Fairness Trade-offs examine similar tensions but lack the cognitive modeling lens, while Value Alignment Assessment branches focus on measurement frameworks rather than interpretive models of decision processes.

Among thirty candidates examined across three contributions, none were identified as clearly refuting the work. The first contribution—applying cognitive models to LLMs—examined ten candidates with zero refutable matches, suggesting limited prior work directly combining cognitive science frameworks with LLM value analysis at this level of formalism. The second contribution on reasoning effort and training dynamics similarly found no refutations across ten candidates, indicating the systematic evaluation of utility shifts across model settings may be novel. The third contribution's method for hypothesis formation about social behaviors also showed no overlapping prior work among ten examined papers, though the limited search scope means exhaustive coverage cannot be claimed.

Given the sparse taxonomy position and absence of refutations within the examined candidate set, the work appears to occupy relatively unexplored methodological territory. However, the analysis is constrained by top-thirty semantic search results and does not guarantee comprehensive coverage of adjacent cognitive science or interpretability literature. The novelty assessment reflects what is visible within this limited scope rather than an exhaustive field survey.

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

Contribution 1: Application of cognitive models to reveal value trade-offs in LLMs

Description: The authors apply a well-established cognitive model from cognitive science (the Rational Speech Acts model of polite speech) to interpret and quantify value trade-offs in large language models. This method is used to analyze both closed-source reasoning models and open-source models across different training stages.

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. Machine Reasoning Framework for Large Language Models

URL: View paper

Brief Assessment

Machine Reasoning Framework[76] focuses on reasoning mechanisms (chain-of-thought, memory architectures, attention) and computational trade-offs in LLMs, not on applying cognitive models from cognitive science to interpret value trade-offs or social behaviors like politeness.

---

2. CognAlign: A Multi-Agent Cognitive-Alignment Framework for Transparent, Bias-Aware Medical Triage Using Small Language Models

URL: View paper

Brief Assessment

CognAlign Medical Triage[73] applies cognitive dual-process reasoning to medical triage tasks using small language models, not to analyzing value trade-offs in large language models. The original paper uses the Rational Speech Acts model to interpret competing utilities in LLMs, while the candidate focuses on clinical decision-making architecture.

---

3. Computational analysis of 100 K choice dilemmas: Decision attributes, trade-off structures, and model-based prediction

URL: View paper

Brief Assessment

Choice Dilemmas Analysis[70] focuses on analyzing human decision-making in real-world choices using LLMs as coding tools, not on applying cognitive models to evaluate LLMs themselves or their value trade-offs.

---

4. Stability-Plasticity Trade-Off in Large Language Models for Health Chatbot Applications

URL: View paper

Brief Assessment

Stability Plasticity Tradeoff[74] focuses on stability-plasticity trade-offs in health chatbot applications using computational neuroscience approaches, not on applying cognitive models from cognitive science (like Rational Speech Acts) to interpret value trade-offs in LLMs.

---

5. How do large language models navigate conflicts between honesty and helpfulness?

URL: View paper

Brief Assessment

Honesty Helpfulness Conflicts[72] applies cognitive models (Rational Speech Acts) to analyze honesty-helpfulness trade-offs in LLMs, while the original paper applies the same framework to analyze informational-social utility trade-offs in polite speech. The domains and specific utilities examined differ substantially.

---

6. Neuro-symbolic models of human moral judgment: LLMs as automatic feature extractors

URL: View paper

Brief Assessment

Neuro Symbolic Moral[77] focuses on using LLMs to extract features for cognitive models of moral judgment, not on applying cognitive models to interpret value trade-offs in LLMs during training or reasoning. The candidate uses cognitive models as prediction targets, while the original uses them as interpretive tools for analyzing LLM behavior.

---

7. Analogies versus rules in cognitive architecture

URL: View paper

Brief Assessment

Analogies versus Rules[75] focuses on the tradeoffs between analogical reasoning and rule-based mechanisms in cognitive architectures for tasks like commonsense reasoning and natural language interpretation. It does not address applying cognitive models to evaluate value trade-offs in language models.

---

8. Parallel trade-offs in human cognition and neural networks: The dynamic interplay between in-context and in-weight learning

URL: View paper

Brief Assessment

Parallel Cognitive Tradeoffs[71] focuses on dual learning systems (in-context vs. in-weight learning) in neural networks as models of human cognition, not on applying cognitive models to evaluate value trade-offs in language models.

---

9. Inside you are many wolves: Using cognitive models to interpret value trade-offs in LLMs

URL: View paper

Brief Assessment

Cognitive Wolves[38] applies the same cognitive model (Rational Speech Acts model of polite speech from Yoon et al. 2020) to the same domain. Both papers use identical methodology to evaluate LLMs' value trade-offs in polite communication, making this a case of near-identical work rather than prior art that refutes novelty.

---

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

URL: View paper

Brief Assessment

Structured Moral Reasoning[16] focuses on moral reasoning frameworks and value systems for ethical decision-making, not on applying cognitive models (like Rational Speech Acts) to quantify utility trade-offs in pragmatic communication tasks.

Contribution 2: Systematic evaluation of reasoning effort and training dynamics on utility trade-offs

Description: The authors provide empirical findings on how reasoning budgets and goal-based prompts affect utility weightings in frontier models, and how base model choice and pretraining data influence utility trade-offs more than feedback datasets or alignment methods during RL post-training.

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. Adaptive Dual Reasoner: Large Reasoning Models Can Think Efficiently by Hybrid Reasoning

URL: View paper

Brief Assessment

Adaptive Dual Reasoner[68] focuses on optimizing reasoning efficiency in mathematical tasks through hybrid fast/slow thinking modes, not on evaluating utility trade-offs (informational vs. social) in language models or analyzing how training dynamics affect value alignment.

---

2. Thinking Sparks!: Emergent Attention Heads in Reasoning Models During Post Training

URL: View paper

Brief Assessment

Thinking Sparks[67] focuses on circuit-level analysis of attention heads during post-training for reasoning tasks, not on utility trade-offs between informational and social goals in language models.

---

3. Training and Inference Time Dynamics of Artificial Neural Networks

URL: View paper

Brief Assessment

Training Inference Dynamics[66] focuses on computational trade-offs in neural network training and inference optimization, not on value alignment or utility trade-offs in language model behavior as studied in the original paper.

---

4. Llm post-training: A deep dive into reasoning large language models

URL: View paper

Brief Assessment

Post Training Reasoning[61] focuses on post-training methodologies (fine-tuning, RL, test-time scaling) for LLMs broadly, not specifically on reasoning effort's effects on utility trade-offs between informational and social goals as measured through cognitive models of polite speech.

---

5. Learning When to Think: Shaping Adaptive Reasoning in R1-Style Models via Multi-Stage RL

URL: View paper

Brief Assessment

Learning When Think[64] focuses on adaptive reasoning in R1-style models through multi-stage RL to optimize computational efficiency, not on evaluating utility trade-offs between social and informational goals in language models.

---

6. Scalable Graph Neural Networks for Global Knowledge Representation and Reasoning

URL: View paper

Brief Assessment

Scalable Graph Reasoning[69] focuses on GNN architectures for knowledge graph reasoning tasks, evaluating trade-offs between scalability, expressiveness, and computational efficiency. This is fundamentally different from the original paper's investigation of reasoning effort and training dynamics affecting utility trade-offs (informational vs. social) in language models.

---

7. AMFT: Aligning LLM Reasoners by Meta-Learning the Optimal Imitation-Exploration Balance

URL: View paper

Brief Assessment

AMFT Meta Learning[63] focuses on balancing SFT and RL rewards through meta-learning for reasoning tasks, not on evaluating how reasoning budgets or training dynamics affect utility trade-offs in value alignment contexts.

---

8. Inside you are many wolves: Using cognitive models to interpret value trade-offs in LLMs

URL: View paper

Brief Assessment

Cognitive Wolves[38] evaluates the same aspects: reasoning budgets in frontier models and RL post-training dynamics in open-source models. The experimental designs, model suites, and findings are nearly identical, indicating this is the same work rather than independent prior art.

---

9. AdaCtrl: Towards Adaptive and Controllable Reasoning via Difficulty-Aware Budgeting

URL: View paper

Brief Assessment

AdaCtrl Adaptive Reasoning[65] focuses on adaptive reasoning budget allocation based on problem difficulty in mathematical reasoning tasks, not on systematic evaluation of how reasoning effort and training dynamics affect utility trade-offs (informational vs. social utility) in language models as studied in the original paper.

---

10. ThinkDial: An Open Recipe for Controlling Reasoning Effort in Large Language Models

URL: View paper

Brief Assessment

ThinkDial[62] focuses on controlling computational reasoning effort through discrete operational modes and token budget reduction, not on evaluating utility trade-offs between informational, social, and presentational goals in language models.

Contribution 3: Method for forming hypotheses about social behaviors and shaping training regimes

Description: The authors demonstrate that their cognitive modeling approach can be used to generate testable hypotheses about high-level social behaviors like sycophancy and to inform the design of training procedures that better manage value trade-offs in LLM development.

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. LLM Social Simulations Are a Promising Research Method

URL: View paper

Brief Assessment

LLM Social Simulations[51] focuses on using LLMs to simulate human research subjects for social science research, not on cognitive modeling approaches for analyzing LLM value trade-offs or informing training procedures as in the original paper.

---

2. Human behavior atlas: Benchmarking unified psychological and social behavior understanding

URL: View paper

Brief Assessment

Human Behavior Atlas[55] focuses on benchmarking unified models for psychological and social behavior understanding across diverse tasks, not on using cognitive models to generate hypotheses about social behaviors or inform training procedures for managing value trade-offs.

---

3. SLEAP: A deep learning system for multi-animal pose tracking

URL: View paper

Brief Assessment

SLEAP Pose Tracking[53] is a deep learning system for multi-animal pose tracking in behavioral neuroscience, not a method for hypothesizing social behaviors or shaping training regimes in language models. The domains are entirely different (computer vision for animal tracking vs. cognitive modeling of LLM values).

---

4. Collective Constitutional AI: Aligning a Language Model with Public Input

URL: View paper

Brief Assessment

Collective Constitutional AI[54] focuses on sourcing public input to shape LM behavior through constitutional principles, not on using cognitive models to generate hypotheses about social behaviors or inform training procedures for value trade-offs.

---

5. Social Behavioral Theory

URL: View paper

Brief Assessment

Social Behavioral Theory[52] focuses on understanding and modifying human social behavior through observation, learning, and reinforcement in contexts like HRM, psychology, and education. It does not address LLM training regimes, cognitive modeling of language models, or computational methods for hypothesis generation about AI system behaviors.

---

6. Self Control Analysis of Adolescent Prosocial Behavior Based on Optimized Random Forest Algorithm

URL: View paper

Brief Assessment

Prosocial Random Forest[59] focuses on analyzing adolescent prosocial behavior using random forest algorithms for psychological research, not on LLM training or cognitive modeling of language systems.

---

7. Simple Behavioral Analysis (SimBA) â□□ an open source toolkit for computer classification of complex social behaviors in experimental animals

URL: View paper

Brief Assessment

SimBA Behavior Analysis[58] focuses on automated classification of animal social behaviors using computer vision, not on cognitive modeling approaches for LLMs or methods for hypothesis generation about language model training regimes.

---

8. VLM-Social-Nav: Socially Aware Robot Navigation Through Scoring Using Vision-Language Models

URL: View paper

Brief Assessment

VLM Social Nav[57] focuses on real-time robot navigation in human environments using vision-language models for scoring socially compliant actions, not on forming hypotheses about social behaviors or shaping training regimes for language models.

---

9. Align on the Fly: Adapting Chatbot Behavior to Established Norms

URL: View paper

Brief Assessment

Align on Fly[60] focuses on real-time alignment with external memory for adapting to changing social norms, not on using cognitive models to generate hypotheses about social behaviors or inform training procedure design.

---

10. Machine-assisted social psychology hypothesis generation.

URL: View paper

Brief Assessment

Machine Hypothesis Generation[56] focuses on generating research hypotheses in social psychology using language models, not on forming hypotheses about LLM social behaviors or shaping LLM training regimes as in the original paper.

Appendix: Text Similarity Detection

Textual similarity detection checked 29 papers and found 3 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. Inside you are many wolves: Using cognitive models to interpret value trade-offs in LLMs

Detected in: Core Task (sibling), Contribution: contribution_1, Contribution: contribution_2

⚠ **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] Using cognitive models to reveal value trade-offs in language models View paper
- [1] Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark View paper
- [2] Context-Aligned and Evidence-Based Detection of Hallucinations in Large Language Model Outputs View paper
- [3] Exploring accuracy-fairness trade-off in large language models View paper
- [4] On the Limits of Language Generation: Trade-Offs between Hallucination and Mode-Collapse View paper
- [5] Towards realistic evaluation of cultural value alignment in large language models: Diversity enhancement for survey response simulation View paper
- [6] Strong and weak alignment of large language models with human values View paper
- [7] Fundamental Safety-Capability Trade-offs in Fine-tuning Large Language Models View paper
- [8] Structural Embedding Projection for Contextual Large Language Model Inference View paper
- [9] Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts View paper
- [10] Performance trade-offs of watermarking large language models View paper
- [11] De-amplifying Bias from Differential Privacy in Language Model Fine-tuning View paper
- [12] Revisiting Privacy, Utility, and Efficiency Trade-offs when Fine-Tuning Large Language Models View paper
- [13] XGUARD: A Graded Benchmark for Evaluating Safety Failures of Large Language Models on Extremist Content View paper
- [14] Valuing time in silicon: Can large language models replicate human value of travel time View paper
- [15] Human Value Alignment in AI View paper
- [16] Structured Moral Reasoning in Language Models: A Value-Grounded Evaluation Framework View paper
- [17] The economic trade-offs of large language models: A case study View paper
- [18] Value FULCRA: Mapping large language models to the multidimensional spectrum of basic human value View paper
- [19] Valuecompass: A framework for measuring contextual value alignment between human and llms View paper
- [20] One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity View paper
- [21] Mind the Value-Action Gap: Do LLMs Act in Alignment with Their Values? View paper
- [22] Watermarking for Large Language Models: A Survey View paper
- [23] Assessing llms for moral value pluralism View paper
- [24] Safeguarding large language models in real-time with tunable safety-performance trade-offs View paper
- [25] Dailydilemmas: Revealing value preferences of llms with quandaries of daily life View paper

- [26] 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
- [27] Application-Driven Value Alignment in Agentic AI Systems: Survey and Perspectives View paper
- [28] Exploring Multilingual Concepts of Human Values in Large Language Models: Is Value Alignment Consistent, Transferable and Controllable across Languages? View paper
- [29] Unpacking the ethical value alignment in big models View paper
- [30] What's the most important value? INVP: INvestigating the Value Priorities of LLMs through Decision-making in Social Scenarios View paper
- [31] The Ghost in the Machine has an American accent: value conflict in GPT-3 View paper
- [32] Cultural Value Alignment in Large Language Models: A Prompt-based Analysis of Schwartz Values in Gemini, ChatGPT, and DeepSeek View paper
- [33] LLM ethics benchmark: a three-dimensional assessment system for evaluating moral reasoning in large language models View paper
- [34] ConflictBank: A Benchmark for Evaluating the Influence of Knowledge Conflicts in LLM View paper
- [35] Sampling-aware Adversarial Attacks Against Large Language Models View paper
- [36] The Staircase of Ethics: Probing LLM Value Priorities through Multi-Step Induction to Complex Moral Dilemmas View paper
- [37] Multimodal Mamba: Decoder-only Multimodal State Space Model via Quadratic to Linear Distillation View paper
- [38] Inside you are many wolves: Using cognitive models to interpret value trade-offs in LLMs View paper
- [39] CLASH: Evaluating Language Models on Judging High-Stakes Dilemmas from Multiple Perspectives View paper
- [40] Can you Finetune your Binoculars? Embedding Text Watermarks into the Weights of Large Language Models View paper
- [41] Generative Value Conflicts Reveal LLM Priorities View paper
- [42] Interpretable Steering of Large Language Models with Feature Guided Activation Additions View paper
- [43] Plurality of value pluralism and AI value alignment View paper
- [44] CVC: A Large-Scale Chinese Value Rule Corpus for Value Alignment of Large Language Models View paper
- [45] Prism: Perspective reasoning for integrated synthesis and mediation as a multi-perspective framework for ai alignment View paper
- [46] " There are no solutions, only trade-offs.'"Taking A Closer Look At Safety Data Annotations. View paper
- [47] LLMs as mediators: Can they diagnose conflicts accurately? View paper
- [48] Training A Small Emotional Vision Language Model for Visual Art Comprehension View paper
- [49] Prosocial behavior in Large Language Models: Value alignment and affective mechanisms View paper
- [50] WaterPool: A Watermark Mitigating Trade-offs among Imperceptibility, Efficacy and Robustness View paper
- [51] LLM Social Simulations Are a Promising Research Method View paper
- [52] Social Behavioral Theory View paper
- [53] SLEAP: A deep learning system for multi-animal pose tracking View paper
- [54] Collective Constitutional AI: Aligning a Language Model with Public Input View paper
- [55] Human behavior atlas: Benchmarking unified psychological and social behavior understanding View paper
- [56] Machine-assisted social psychology hypothesis generation. View paper
- [57] VLM-Social-Nav: Socially Aware Robot Navigation Through Scoring Using Vision-Language Models View paper
- [58] Simple Behavioral Analysis (SimBA) â an open source toolkit for computer classification of complex social behaviors in experimental animals View paper
- [59] Self Control Analysis of Adolescent Prosocial Behavior Based on Optimized Random Forest Algorithm View paper
- [60] Align on the Fly: Adapting Chatbot Behavior to Established Norms View paper
- [61] Llm post-training: A deep dive into reasoning large language models View paper
- [62] ThinkDial: An Open Recipe for Controlling Reasoning Effort in Large Language Models View paper
- [63] AMFT: Aligning LLM Reasoners by Meta-Learning the Optimal Imitation-Exploration Balance View paper
- [64] Learning When to Think: Shaping Adaptive Reasoning in R1-Style Models via Multi-Stage RL View paper
- [65] AdaCtrl: Towards Adaptive and Controllable Reasoning via Difficulty-Aware Budgeting View paper
- [66] Training and Inference Time Dynamics of Artificial Neural Networks View paper
- [67] Thinking Sparks!: Emergent Attention Heads in Reasoning Models During Post Training View paper
- [68] Adaptive Dual Reasoner: Large Reasoning Models Can Think Efficiently by Hybrid Reasoning View paper
- [69] Scalable Graph Neural Networks for Global Knowledge Representation and Reasoning View paper
- [70] Computational analysis of 100 K choice dilemmas: Decision attributes, trade-off structures, and model-based prediction View paper
- [71] Parallel trade-offs in human cognition and neural networks: The dynamic interplay between in-context and in-weight learning View paper
- [72] How do large language models navigate conflicts between honesty and helpfulness? View paper
- [73] CognAlign: A Multi-Agent Cognitive-Alignment Framework for Transparent, Bias-Aware Medical Triage Using Small Language Models View paper
- [74] Stability-Plasticity Trade-Off in Large Language Models for Health Chatbot Applications View paper
- [75] Analogies versus rules in cognitive architecture View paper
- [76] Machine Reasoning Framework for Large Language Models View paper
- [77] Neuro-symbolic models of human moral judgment: LLMs as automatic feature extractors View paper