Novelty Assessment Report

Paper: Evaluating steering techniques using human similarity judgments
PDF URL: https://openreview.net/pdf?id=kkNr2niHtT
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Current evaluations of Large Language Model (LLM) steering techniques focus on task-specific performance, overlooking how well steered representations align with human cognition. Using a well-established triadic similarity judgment task, we assessed steered LLMs on their ability to flexibly judge similarity between concepts based on size or kind. We found that prompt-based steering methods outperformed other methods both in terms of steering accuracy and model-to-human alignment. We also found LLMs were biased towards "kind" similarity and struggled with "size" alignment. This evaluation approach, grounded in human cognition, adds further support to the efficacy of prompt-based steering and reveals privileged representational axes in LLMs prior to steering.

Core Task Landscape

This paper addresses: **Evaluating LLM Steering Techniques Using Triadic Similarity Judgments**

A total of **4 papers** were analyzed and organized into a taxonomy with **5 categories**.

Taxonomy Overview

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

- **Steering and Alignment Evaluation Methods**
- **Application Domains and Task-Specific Implementations**

Complete Taxonomy Tree

- Evaluating LLM Steering Techniques Using Triadic Similarity Judgments Survey Taxonomy
- Steering and Alignment Evaluation Methods
  - Cognitive Alignment Through Similarity Judgments ★ (1 papers)
  - [0] Evaluating steering techniques using human similarity judgments (Anon et al., 2026) View paper
  - Mechanistic Interpretability of LLM Behavior (1 papers)
  - [4] LLM Assertiveness can be Mechanistically Decomposed into Emotional and Logical Components (Tagade, 2025) View paper
- Application Domains and Task-Specific Implementations
  - Knowledge-Augmented Reasoning Systems (1 papers)
  - [1] Reflection on knowledge graph for large language models reasoning (Yiming Zhou, 2025) View paper
  - Recommender Systems with LLM Integration (1 papers)
  - [3] Large Language Models for Recommender Systems: A Problem-Driven Survey (Ziyuan Guan, 2025) View paper
  - Security and Adversarial Robustness in Web Agents (1 papers)
  - [2] Mind the Web: The Security of Web Use Agents (Shapira, 2025) View paper

Narrative

Core task: Evaluating LLM steering techniques using triadic similarity judgments. The field of LLM steering and alignment has grown into a diverse landscape, organized here around two main branches. The first branch, Steering and Alignment Evaluation Methods, encompasses approaches that assess how well models can be guided toward desired behaviors, including cognitive alignment strategies that probe whether models internalize human-like conceptual structures. The second branch, Application Domains and Task-Specific Implementations, focuses on deploying steering techniques in concrete settings such as recommendation systems, web agents, and knowledge-intensive tasks. Works like LLMs for Recommenders[3] illustrate how steering principles translate into domain-specific challenges, while Knowledge Graph Reflection[1] and Web Agent Security[2] demonstrate the breadth of contexts where alignment and control matter. Together, these branches reflect a field balancing foundational evaluation questions with practical deployment concerns.

Within the Steering and Alignment Evaluation Methods branch, a particularly active line of inquiry examines how to measure alignment beyond surface-level performance metrics, exploring whether models exhibit human-like reasoning patterns or merely mimic outputs. Steering Techniques Evaluation[0] sits squarely in this cognitive alignment cluster, using triadic similarity judgments—a psychologically grounded method—to assess whether steering interventions genuinely shift internal representations in interpretable ways. This contrasts with more application-driven works like LLMs for Recommenders[3], which prioritize task success over cognitive fidelity, and complements efforts like LLM Assertiveness Decomposition[4], which dissects model behavior into interpretable components. The central tension across these directions is whether evaluation should emphasize human-aligned internal structure or downstream utility, with Steering Techniques Evaluation[0] leaning toward the former by grounding its metrics in human similarity perception.

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

Both subtopics address how to understand and evaluate LLM internal representations and behaviors, but from complementary perspectives. Cognitive Alignment focuses on external validation through human similarity judgments to assess whether LLMs represent

concepts like humans do. Mechanistic Interpretability takes an internal approach, decomposing model representations to understand specific behavioral traits.

Similarities: - Both aim to understand LLM representational structures and how they relate to interpretable concepts - Both address alignment questions—whether LLM representations match desired or expected patterns - Both can inform model evaluation and improvement strategies

Differences: - Cognitive Alignment uses external human judgment tasks (triadic similarity) as ground truth; Mechanistic Interpretability analyzes internal model components directly - Cognitive Alignment focuses on representational alignment with human cognition broadly; Mechanistic Interpretability targets specific behavioral traits like assertiveness or confidence - Cognitive Alignment evaluates outputs through behavioral tasks; Mechanistic Interpretability decomposes intermediate representations and activations - The original leaf explicitly excludes mechanistic decomposition methods, while the sibling excludes external human judgment evaluation

Suggested Search Directions: - Hybrid approaches combining similarity judgments with mechanistic analysis to validate internal decompositions - Studies comparing human-aligned representations (from similarity tasks) with mechanistically identified features - Research on whether mechanistically interpretable features correspond to human-judged similarity structures

Sibling Subtopics

- **Mechanistic Interpretability of LLM Behavior** (leaves: 1, papers: 1)
- Scope: Studies decomposing internal LLM representations to understand behavioral traits like assertiveness or confidence.
- Exclude: Excludes external alignment evaluation via human judgment tasks; see Cognitive Alignment Through Similarity Judgments.

Contributions Analysis

Overall novelty summary. The paper introduces an evaluation framework for LLM steering techniques grounded in triadic similarity judgments, a task drawn from cognitive psychology. Within the taxonomy, it occupies the 'Cognitive Alignment Through Similarity Judgments' leaf under 'Steering and Alignment Evaluation Methods,' where it is currently the sole representative. This positioning reflects a sparse research direction: the broader 'Steering and Alignment Evaluation Methods' branch contains only two leaves, with the sibling leaf ('Mechanistic Interpretability of LLM Behavior') focusing on internal representation decomposition rather than external human-alignment tasks. The taxonomy reveals that cognitive-alignment evaluation via similarity tasks is an underexplored niche within the larger steering literature.

The taxonomy structure shows that most related work clusters in the 'Application Domains' branch, emphasizing task-specific implementations (knowledge reasoning, recommender systems, web agent security) rather than foundational evaluation methods. The neighboring 'Mechanistic Interpretability' leaf examines internal model behavior through decomposition techniques, offering a complementary but distinct approach to understanding steering effects. The paper's cognitive-psychology grounding distinguishes it from these application-oriented and mechanistic directions, bridging human cognition research with LLM steering evaluation in a way that the taxonomy suggests is relatively novel within this literature sample.

Across three identified contributions, the literature search examined 30 candidates total, with 10 candidates per contribution. None of the contributions were clearly refuted by prior work in this limited sample. The evaluation framework using triadic similarity judgments, the dual-axis competence-alignment measurement, and the discovery of privileged representational axes each showed no overlapping prior work among the examined candidates. This suggests that, within the scope of the top-30 semantic matches and their citations, the specific combination of triadic similarity tasks, dual-axis evaluation, and representational bias analysis appears distinctive, though the search scale leaves open the possibility of relevant work beyond this sample.

Based on the limited search scope, the work appears to occupy a relatively unexplored intersection of cognitive psychology and LLM steering evaluation. The taxonomy's sparse population in this direction and the absence of refuting candidates among 30 examined papers suggest novelty, though this assessment is constrained by the search methodology. A more exhaustive review of cognitive science applications to LLM evaluation or broader steering literature might reveal additional relevant precedents not captured in this top-K semantic search.

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

Contribution 1: Evaluation framework using triadic similarity judgments for LLM steering

Description: The authors introduce an evaluation approach for LLM steering techniques grounded in cognitive science methods. They apply triadic similarity judgment tasks—where agents judge which of two items is most similar to a reference item along specified dimensions (size or kind)—to assess both steering accuracy and alignment with human mental representations.

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. Harnessing the Power of Semi-Structured Knowledge and LLMs with Triplet-Based Prefiltering for Question Answering

URL: View paper

Brief Assessment

Triplet Prefiltering QA[21] uses triplets for knowledge graph queries in question answering, not for evaluating LLM steering methods or assessing alignment with human cognition through similarity judgments.

---

2. Deep metric learning-based semi-supervised regression with alternate learning

URL: View paper

Brief Assessment

Semi Supervised Regression[23] focuses on parameter estimation using triplet-based metric learning for regression tasks with scarce labeled data, not on evaluating language model steering methods or cognitive alignment.

---

3. Triplets better than pairs: Towards stable and effective self-play fine-tuning for LLMs

URL: View paper

Brief Assessment

Triplet Self Play[19] focuses on self-play fine-tuning methods for LLM adaptation using triplet-based optimization, not on evaluation frameworks using triadic similarity judgments for assessing steering techniques.

---

4. F2rl: Factuality and faithfulness reinforcement learning framework for claim-guided evidence-supported counterspeech generation

URL: View paper

Brief Assessment

Factuality Faithfulness Reinforcement[18] focuses on counterspeech generation using reinforcement learning with factuality and faithfulness reward models. It does not address triadic similarity judgment tasks or evaluation methods for LLM steering techniques.

5. A Metric-Based Detection System for Large Language Model Texts

URL: View paper

Brief Assessment

Metric Based Detection[15] focuses on detecting LLM-generated texts using metric learning and triplet-based training for classification, not on evaluating LLM steering techniques or assessing alignment with human mental representations through similarity judgments.

6. Exploring Human and Language Model Alignment in Perceived Design Similarity Using Ordinal Embeddings

URL: View paper

Brief Assessment

Design Similarity Alignment[20] uses triadic comparisons to align LLM embeddings with human design similarity judgments, not to evaluate LLM steering techniques. The candidate focuses on engineering design perception rather than cognitive control mechanisms.

7. Does a Large Language Model Really Speak in Humanâ□□Like Language?

URL: View paper

Brief Assessment

Human Like Language[16] focuses on comparing latent community structures between human-written and LLM-generated text through paraphrasing analysis, not on evaluating LLM steering techniques using triadic similarity judgments.

8. MKFGO: integrating multi-source knowledge fusion with pretrained language model for high-accuracy protein function prediction

URL: View paper

Brief Assessment

Multi Source Protein[24] focuses on protein function prediction using deep learning and does not address LLM steering evaluation or triadic similarity judgment tasks for language models.

9. Triplet-based contrastive method enhances the reasoning ability of large language models

URL: View paper

Brief Assessment

Triplet Contrastive Reasoning[17] focuses on contrastive methods for enhancing reasoning ability, not on evaluating LLM steering techniques using triadic similarity judgment tasks from cognitive science.

10. A classified feature representation three-way decision model for sentiment analysis

URL: View paper

Brief Assessment

Three Way Sentiment[22] focuses on sentiment analysis using three-way decision models for feature representation, not on evaluating LLM steering techniques or triadic similarity judgment tasks for language models.

Contribution 2: Dual-axis evaluation measuring competence and alignment

Description: The authors propose evaluating steering methods along two distinct dimensions: competence (task accuracy) and alignment (how well steered model representations match human representational geometry). This dual evaluation framework distinguishes between performance and cognitive similarity, addressing the gap between what systems do and how they do it.

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. Large language model alignment: A survey

URL: View paper

Brief Assessment

LLM Alignment Survey[27] provides a broad overview of alignment evaluation methods but does not specifically propose or discuss dual-axis evaluation frameworks that simultaneously measure task accuracy (competence) and human representational alignment (cognitive similarity) as distinct evaluation dimensions.

2. Assessment of multimodal large language models in alignment with human values

URL: View paper

Brief Assessment

Multimodal Human Values[28] focuses on evaluating multimodal LLMs across three levels (semantic, logic, human values) using the HHH framework, not on dual-axis evaluation of steering methods measuring task accuracy versus human representational geometry alignment.

3. Aligning large multimodal models with factually augmented rlhf

URL: View paper

Brief Assessment

Factually Augmented RLHF[29] focuses on multimodal alignment through RLHF for vision-language models, evaluating helpfulness and hallucination reduction. It does not propose a dual-axis framework measuring both task accuracy (competence) and human representational alignment as distinct evaluation dimensions for steering methods.

4. Rrhf: Rank responses to align language models with human feedback

URL: View paper

Brief Assessment

RRHF Rank Responses[32] focuses on aligning language models with human preferences through ranking loss on response quality, not on dual-axis evaluation of competence versus representational alignment with human cognitive similarity.

5. Direct Language Model Alignment from Online AI Feedback

URL: View paper

Brief Assessment

Online AI Feedback[30] focuses on online feedback mechanisms for preference-based alignment methods (DPO, RLHF), not on dual-axis evaluation frameworks measuring both task accuracy and representational alignment with human cognition.

6. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback

URL: View paper

Brief Assessment

DRESS Vision Language[34] focuses on vision-language models using natural language feedback for alignment with human preferences (helpfulness, honesty, harmlessness), not on evaluating steering techniques with dual metrics of task accuracy and representational alignment as in the original paper.

---

7. Aligning large language models with human: A survey

URL: View paper

Brief Assessment

Aligning with Human[26] is a survey paper focused on alignment training methodologies (RLHF, SFT) and data collection for LLMs, not on dual-axis evaluation frameworks measuring both task accuracy and representational similarity to human cognition.

---

8. Principle-driven self-alignment of language models from scratch with minimal human supervision

URL: View paper

Brief Assessment

Principle Driven Self Alignment[31] focuses on aligning language models using principle-driven self-alignment techniques with minimal human supervision, not on dual-axis evaluation frameworks measuring both task accuracy and human representational alignment. The candidate does not address evaluation methodologies that assess cognitive similarity between model and human representations.

---

9. Pretraining language models with human preferences

URL: View paper

Brief Assessment

Pretraining Human Preferences[25] evaluates language models using task performance metrics and alignment with human preferences, but focuses on pretraining objectives rather than steering techniques. The dual-axis framework in the original paper specifically evaluates steering methods using human representational geometry from triadic similarity judgments, which is methodologically distinct from the pretraining-focused evaluation in the candidate paper.

---

10. Decoding-Time Language Model Alignment with Multiple Objectives

URL: View paper

Brief Assessment

Decoding Time Alignment[33] focuses on multi-objective optimization at decoding time for different reward functions (safety, coding, user preference), not on evaluating cognitive alignment with human representational geometry through similarity judgments.

Contribution 3: Discovery of privileged representational axes in LLMs

Description: The authors identify that LLMs exhibit inherent biases in their representational structure, specifically showing stronger alignment with kind-based similarity over size-based similarity even without explicit steering. This finding reveals systematic differences in how LLMs organize semantic knowledge compared to humans.

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. Semantic-Aware Methods for the Analysis of Bias and Underrepresentation in Language Resources

URL: View paper

Brief Assessment

Semantic Bias Analysis[12] focuses on detecting representational bias in datasets (particularly Wikipedia biographies) through biographical event extraction, not on analyzing inherent biases in LLM representational structures or privileged axes in semantic organization.

---

2. Mitigating political bias in language models through reinforced calibration

URL: View paper

Brief Assessment

Political Bias Mitigation[9] focuses on mitigating political bias in language model generation through reinforcement learning calibration, not on discovering privileged representational axes or inherent biases in semantic organization. The candidate addresses ideological bias in generated text rather than systematic differences in how LLMs organize semantic knowledge.

---

3. Evaluating biased attitude associations of language models in an intersectional context

URL: View paper

Brief Assessment

Intersectional Bias Evaluation[10] focuses on measuring valence associations and social biases in language models, not on identifying privileged representational axes in semantic similarity tasks or comparing kind-based versus size-based similarity organization.

---

4. Bias and fairness in large language models: A survey

URL: View paper

Brief Assessment

Bias Fairness Survey[5] focuses on social biases and fairness in LLMs (gender, race, etc.), not on cognitive representational structures like kind vs. size similarity axes.

---

5. Tokens, the oft-overlooked appetizer: Large language models, the distributional hypothesis, and meaning

URL: View paper

Brief Assessment

Tokens and Meaning[14] focuses on tokenization's impact on LLM cognition and the distributional hypothesis, not on privileged representational axes or inherent biases in semantic organization. The paper does not address steering techniques or similarity judgments that would challenge the original's novelty claim about discovering systematic biases in representational structure.

---

6. Semantic and structural analysis of implicit biases in large language models: An interpretable approach

URL: View paper

Brief Assessment

Implicit Bias Analysis[7] focuses on detecting social stereotypes (gender, profession, religion, race) in LLM outputs through semantic embedding and attention perturbation, not on analyzing privileged representational axes in semantic similarity tasks like kind vs. size dimensions.

---

7. The large language model (LLM) bias evaluation (age bias)

URL: View paper

Brief Assessment

Age Bias Evaluation[13] focuses on age bias evaluation in LLMs, not on privileged representational axes or semantic organization structures like kind vs. size similarity.

---

8. Large language models portray socially subordinate groups as more homogeneous, consistent with a bias observed in humans

URL: View paper

Brief Assessment

Homogeneity Bias LLMs[11] investigates social bias in text generation (portraying subordinate groups as more homogeneous), not privileged representational axes in semantic organization or similarity judgments.

---

9. Reducing sentiment bias in language models via counterfactual evaluation

URL: View paper

Brief Assessment

Counterfactual Sentiment Bias[6] focuses on sentiment bias in language model text generation and proposes fairness metrics based on counterfactual evaluation. It does not investigate privileged representational axes in semantic organization or compare kind-based versus size-based similarity judgments as the original paper does.

---

10. Language model behavior: A comprehensive survey

URL: View paper

Brief Assessment

Language Model Behavior[8] is a broad survey of LLM capabilities across syntax, semantics, and reasoning, but does not specifically investigate privileged representational axes or systematic biases in semantic organization (e.g., kind vs. size dimensions).

Appendix: Text Similarity Detection

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

References

- [0] Evaluating steering techniques using human similarity judgments View paper
- [1] Reflection on knowledge graph for large language models reasoning View paper
- [2] Mind the Web: The Security of Web Use Agents View paper
- [3] Large Language Models for Recommender Systems: A Problem-Driven Survey View paper
- [4] LLM Assertiveness can be Mechanistically Decomposed into Emotional and Logical Components View paper
- [5] Bias and fairness in large language models: A survey View paper
- [6] Reducing sentiment bias in language models via counterfactual evaluation View paper
- [7] Semantic and structural analysis of implicit biases in large language models: An interpretable approach View paper
- [8] Language model behavior: A comprehensive survey View paper
- [9] Mitigating political bias in language models through reinforced calibration View paper
- [10] Evaluating biased attitude associations of language models in an intersectional context View paper
- [11] Large language models portray socially subordinate groups as more homogeneous, consistent with a bias observed in humans View paper
- [12] Semantic-Aware Methods for the Analysis of Bias and Underrepresentation in Language Resources View paper
- [13] The large language model (LLM) bias evaluation (age bias) View paper
- [14] Tokens, the oft-overlooked appetizer: Large language models, the distributional hypothesis, and meaning View paper
- [15] A Metric-Based Detection System for Large Language Model Texts View paper
- [16] Does a Large Language Model Really Speak in Humanâ□□Like Language? View paper
- [17] Triplet-based contrastive method enhances the reasoning ability of large language models View paper
- [18] F2rl: Factuality and faithfulness reinforcement learning framework for claim-guided evidence-supported counterspeech generation View paper
- [19] Triplets better than pairs: Towards stable and effective self-play fine-tuning for LLMs View paper
- [20] Exploring Human and Language Model Alignment in Perceived Design Similarity Using Ordinal Embeddings View paper
- [21] Harnessing the Power of Semi-Structured Knowledge and LLMs with Triplet-Based Prefiltering for Question Answering View paper
- [22] A classified feature representation three-way decision model for sentiment analysis View paper
- [23] Deep metric learning-based semi-supervised regression with alternate learning View paper
- [24] MKFGO: integrating multi-source knowledge fusion with pretrained language model for high-accuracy protein function prediction View paper
- [25] Pretraining language models with human preferences View paper
- [26] Aligning large language models with human: A survey View paper
- [27] Large language model alignment: A survey View paper
- [28] Assessment of multimodal large language models in alignment with human values View paper
- [29] Aligning large multimodal models with factually augmented rlhf View paper
- [30] Direct Language Model Alignment from Online AI Feedback View paper
- [31] Principle-driven self-alignment of language models from scratch with minimal human supervision View paper
- [32] Rrhf: Rank responses to align language models with human feedback View paper
- [33] Decoding-Time Language Model Alignment with Multiple Objectives View paper
- [34] Dress: Instructing large vision-language models to align and interact with humans via natural language feedback View paper