Novelty Assessment Report

Paper: Uncovering the computational ingredients that support human-like conceptual representations in large language models
PDF URL: https://openreview.net/pdf?id=g5pYm2OmfA
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

The ability to translate diverse patterns of inputs into structured patterns of behavior has been thought to rest on both humans' and machines' ability to learn robust representations of relevant concepts. The rapid advancement of transformer-based large language models (LLMs) has led to a diversity of computational ingredients — architectures, fine tuning methods, and training datasets among others — but it remains unclear which of these ingredients are most crucial for building models that develop human-like representations. Further, most current LLM benchmarks are not suited to measuring representational alignment between humans and models, making existing benchmark scores unreliable for assessing if current LLMs are making progress towards becoming useful cognitive models. Here, we address these limitations by first evaluating a set of over 70 models that widely vary in their computational ingredients on a triplet similarity task, a method well established in the cognitive sciences for measuring human conceptual representations, using concepts from the THINGS database. Comparing human and model representations, we find that models that undergo instruction-finetuning and which have larger dimensionality of attention heads are among the most human aligned. We also find that factors such as choice of activation function, multimodal pretraining, and parameter size have limited bearing on alignment. Correlations between alignment scores and scores on existing benchmarks reveal that while some benchmarks (e.g., MMLU) are better suited than others (e.g., MUSR) for capturing representational alignment, no existing benchmark is capable of fully accounting for the variance of alignment scores, demonstrating their insufficiency in capturing human-AI alignment. Taken together, our findings help highlight the computational ingredients most essential for advancing LLMs towards models of human conceptual representation and address a key benchmarking gap in LLM evaluation.

Core Task Landscape

This paper addresses: **Representational Alignment Between Language Models and Human Conceptual Knowledge**

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:

- **Conceptual Representation Emergence and Structure**
- **Alignment Assessment Methodologies**
- **Behavioral Alignment and Psycholinguistic Correspondence**
- **Multimodal Conceptual Alignment**
- **Grounding and Embodied Semantics**
- **Neural and Brain-Based Alignment**
- **Value and Preference Alignment**
- **Compression-Meaning Trade-offs and Conceptual Efficiency**
- **Applied Alignment in Domain-Specific Contexts**
- **Theoretical Frameworks and Dual-Level Representations**

Complete Taxonomy Tree

- Representational Alignment Between Language Models and Human Conceptual Knowledge Survey Taxonomy
- Conceptual Representation Emergence and Structure
  - Concept Formation from Language Prediction (3 papers)
  - [4] Revealing emergent human-like conceptual representations from language prediction. (Ningyu Xu, 2025) View paper
  - [11] Implicit representations of meaning in neural language models (Belinda Z. Li, 2021) View paper
  - [12] Human-like conceptual representations emerge from language prediction (Xu, 2025) View paper
  - Semantic Representation Architecture (3 papers)
  - [6] Semantics-aware BERT for language understanding (Zhuosheng Zhang, 2020) View paper
  - [23] Exploring semantics in pretrained language model attention (FrÃ©dÃ©ric Charpentier, 2024) View paper
  - [44] Frame Representation Hypothesis: Multi-Token LLM Interpretability and Concept-Guided Text Generation (Pedro H. V. Valois, 2024) View paper
  - Conceptual Consistency and Understanding (2 papers)
  - [7] Do large language models â□understandâ□□ their knowledge? (Venkat Venkatasubramanian, 2024) View paper
  - [13] Unpacking large language models with conceptual consistency (Sahu, 2022) View paper
- Alignment Assessment Methodologies
  - Similarity-Based Alignment Metrics ★ (4 papers)
  - [0] Uncovering the computational ingredients that support human-like conceptual representations in large language models (Anon et al., 2026) View paper
  - [1] Mapping language models to grounded conceptual spaces (R Patel, 2022) View paper

    - [31] A Flexible Method for Behaviorally Measuring Alignment Between Human and Artificial Intelligence Using Representational Similarity Analysis (Bose, 2024) View paper
    - [35] Exploring Human and Language Model Alignment in Perceived Design Similarity Using Ordinal Embeddings (Matthew Keeler, 2025) View paper
    - Abstraction and Relational Alignment (2 papers)
    - [5] Abstraction Alignment: Comparing Model and Human Conceptual Relationships (Boggust, 2024) View paper
    - [8] Abstraction Alignment: Comparing Model-Learned and Human-Encoded Conceptual Relationships (Angie Boggust, 2025) View paper
    - Cross-Linguistic and Cross-Cultural Alignment (4 papers)
    - [34] Language Model Alignment in Multilingual Trolley Problems (Jin, 2024) View paper
    - [46] Exploring Multilingual Concepts of Human Values in Large Language Models: Is Value Alignment Consistent, Transferable and Controllable across Languages? (Shaoyang Xu, 2024) View paper
    - [47] Exploring Multilingual Concepts of Human Value in Large Language Models: Is Value Alignment Consistent, Transferable and Controllable across Languages? (Dong Wei-Long, 2024) View paper
    - Non-Aligned and Divergent Representations (2 papers)
    - [29] Identifying and interpreting non-aligned human conceptual representations using language modeling (Wanqian Bao, 2024) View paper
    - [45] Divergences between Language Models and Human Brains (Emmy Liu, 2024) View paper
- Behavioral Alignment and Psycholinguistic Correspondence
    - Psycholinguistic Task Replication (1 papers)
    - [10] Do large language models resemble humans in language use? (Zhenguang Cai, 2024) View paper
    - Semantic Disambiguation and Ambiguity (2 papers)
    - [37] Do large language models resolve semantic ambiguities in the same way as humans? The case of word segmentation in Chinese sentence reading (Liao, 2024) View paper
    - [39] A study on semantic understanding of large language models from the perspective of ambiguity resolution (Shuguang Yang, 2023) View paper
    - Meaning and Semantic Depth (5 papers)
    - [3] Dissociating language and thought in large language models (Kyle Mahowald, 2024) View paper
    - [19] Meaning and understanding in large language models (VladimÃr HavlÃk, 2024) View paper
    - [25] From form (s) to meaning: Probing the semantic depths of language models using multisense consistency (Xenia Ohmer, 2024) View paper
    - [26] Do language models have semantics? on the five standard positions (Anders SÃ¸gaard, 2025) View paper
    - [28] Against AI understanding and sentience: large language models, meaning, and the patterns of human language use (Durt Christoph, 2023) View paper
- Multimodal Conceptual Alignment
    - Vision-Language Representational Convergence (1 papers)
    - [16] Seeing Through Words, Speaking Through Pixels: Deep Representational Alignment Between Vision and Language Models (Trott, 2025) View paper
    - Object and Physical Concept Understanding (2 papers)
    - [17] Can language models understand physical concepts? (Li Lei, 2023) View paper
    - [32] Human-like object concept representations emerge naturally in multimodal large language models (Changde Du, 2024) View paper
    - Multimodal Fine-Tuning and Representation Shift (1 papers)
    - [33] Analyzing Fine-tuning Representation Shift for Multimodal LLMs Steering alignment (Shukor, 2025) View paper
- Neural and Brain-Based Alignment
    - Brain-Model Representational Similarity (3 papers)
    - [9] Decoding Semantic Representations in the Brain Under Language Stimuli with Large Language Models (A Sato, 2025) View paper
    - [22] Distinct social-linguistic processing between humans and large audio-language models: Evidence from model-brain alignment (Hanlin Wu, 2025) View paper
    - [27] Language models align with brain regions that represent concepts across modalities (Ryskina, 2025) View paper
    - Brain-Tuning and Neural Optimization (1 papers)
    - [18] Improving semantic understanding in speech language models via brain-tuning (Klakow, 2024) View paper
    - Concreteness and Lexical Effects (1 papers)
    - [14] The Representational Alignment between Humans and Language Models is implicitly driven by a Concreteness Effect (Choksi, 2025) View paper
- Value and Preference Alignment
    - Human Value and Norm Alignment (3 papers)
    - [2] Aligning large language models with human: A survey (Wang Yu-Fei, 2023) View paper
    - [24] In conversation with artificial intelligence: aligning language models with human values (Atoosa Kasirzadeh, 2023) View paper
    - [40] The problem of alignment (Tsvetelina Hristova, 2025) View paper
    - Conversational and Interaction Alignment (1 papers)
    - [38] Human Preferences for Constructive Interactions in Language Model Alignment (Yara Kyrychenko, 2025) View paper
- Compression-Meaning Trade-offs and Conceptual Efficiency (1 papers)
    - [49] From Tokens to Thoughts: How LLMs and Humans Trade Compression for Meaning (Shani, 2025) View paper
- Applied Alignment in Domain-Specific Contexts
    - Semantic Communication Systems (3 papers)
    - [15] Large language model enabled semantic communication systems (Wang Zhen-yi, 2024) View paper
    - [21] Semantic Importance-Aware Communications with Semantic Correction Using Large Language Models (Shuaishuai Guo, 2025) View paper
    - [30] Large-Language-Model-Enabled Text Semantic Communication Systems (Zhenyi Wang, 2025) View paper
    - Knowledge Graph and Skill Graph Construction (1 papers)
    - [20] Skill Graph Construction From Semantic Understanding (Shi-yong Lin, 2023) View paper
    - Interactive Machine Teaching and Co-Adaptation (1 papers)

- [43] Supporting Co-Adaptive Machine Teaching through Human Concept Learning and Cognitive Theories (Simret Araya Gebreegziabher, 2024) View paper
- Conceptual Modeling and Design Applications (1 papers)
- [41] Conceptual Modeling and Large Language Models: Impressions From First Experiments With ChatGPT (Fill, 2023) View paper

• Theoretical Frameworks and Dual-Level Representations
- Concept Alignment Frameworks (1 papers)
- [36] Concept alignment (Rane, 2024) View paper
- Dual Embodied-Symbolic Representations (1 papers)
- [48] Dual Embodied-Symbolic Concept Representations for Deep Learning (Chang, 2022) View paper
- Structural Alignment and Discourse Organization (1 papers)
- [42] Align to Structure: Aligning Large Language Models with Structural Information (Kim, 2025) View paper

Narrative

Core task: representational alignment between language models and human conceptual knowledge. The field examines how internal representations in language models correspond to human conceptual structures, spanning multiple methodological and theoretical branches. Alignment Assessment Methodologies develop techniques to measure correspondence through similarity metrics, behavioral probing, and psycholinguistic benchmarks, while Conceptual Representation Emergence and Structure investigates how abstract concepts arise during training. Multimodal Conceptual Alignment extends these questions to vision-language systems, and Neural and Brain-Based Alignment compares model activations directly to neural recordings. Grounding and Embodied Semantics explores whether models capture perceptual and physical aspects of meaning, contrasting with purely distributional approaches. Value and Preference Alignment addresses normative concepts and ethical reasoning, and Theoretical Frameworks propose dual-level or hybrid architectures to explain representational capacities. Applied branches examine domain-specific contexts such as medical or legal reasoning, while Compression-Meaning Trade-offs study how efficiency constraints shape conceptual fidelity.

Within Alignment Assessment Methodologies, a particularly active line of work develops similarity-based metrics that quantify structural correspondence between model and human representations. Computational Ingredients Conceptual[0] contributes to this effort by proposing new computational measures for assessing conceptual alignment, situated among studies that use representational similarity analysis and geometric comparisons. Nearby works such as Behavioral Alignment Measurement[31] emphasize task-based probing to validate alignment claims, while Design Similarity Alignment[35] explores how alignment metrics can inform model design choices. A central tension across these branches concerns whether similarity in representational geometry suffices to demonstrate genuine conceptual understanding, or whether behavioral and grounding criteria are necessary. Some studies like Dissociating Language Thought[3] argue for dissociations between linguistic competence and deeper conceptual knowledge, raising questions about what alignment metrics actually capture and how they relate to human-like reasoning.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. Mapping language models to grounded conceptual spaces

Authors: R Patel, E Pavlick | **Year/Venue**: 2022 | **URL**: View paper

Abstract

â□¦ In Extended Abstracts of the 2021 CHI Conference on Human Factors â□¦ concepts we aim to teach language models. The first column shows the category of related concepts that models â□¦

Relationship Analysis

Both papers belong to the Similarity-Based Alignment Metrics category, employing similarity judgments to assess representational alignment between language models and human conceptual knowledge. The original paper uses triplet similarity tasks to systematically evaluate 77 models across computational ingredients (architecture, fine-tuning, dimensionality) to identify factors driving human-model alignment, while the candidate paper focuses on grounding language models to conceptual spaces through few-shot learning in specific domains (spatial terms, cardinal directions, colors) to test whether models can generalize grounded meanings. The key difference is that the original paper conducts a broad comparative analysis of model ingredients using established human embeddings, whereas the candidate paper investigates grounding mechanisms and generalization capabilities within constrained conceptual domains.

---

2. A Flexible Method for Behaviorally Measuring Alignment Between Human and Artificial Intelligence Using Representational Similarity Analysis

Authors: Bose, Ritwik, Mattson Ogg, Ritwik Bose, Ratto, et al. (11 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

As we consider entrusting Large Language Models (LLMs) with key societal and decision-making roles, measuring their alignment with human cognition becomes critical. This requires methods that can assess how these systems represent information and facilitate comparisons with human understanding across diverse tasks. To meet this need, we adapted Representational Similarity Analysis (RSA), a method that uses pairwise similarity ratings to quantify alignment between AIs and humans. We tested this a...

Relationship Analysis

Both papers employ similarity-based alignment metrics within the same taxonomy category, using pairwise similarity judgments to quantify representational alignment between language models and human conceptual knowledge. They overlap in their use of triplet/similarity tasks to derive semantic embeddings and compare model-human representational geometry using methods like Procrustes alignment and RSA. The key difference is that the original paper systematically investigates which computational ingredients (instruction-tuning, dimensionality, architecture) drive alignment across 77 models using the THINGS dataset, while the candidate paper focuses on adapting RSA methodology to measure alignment flexibility across modalities (text and images) and individual-level variability with fewer models (primarily GPT-4o and VLMs).

---

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

Authors: Matthew Keeler, Mark D. Fuge, Aoran Peng, Mark Fuge, Scarlett Miller, et al. (6 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Well-studied techniques that enhance diversity in early design concept generation require effective metrics for evaluating human-perceived similarity between ideas. Recent work suggests collecting triplet comparisons between designs directly from human raters and using those triplets to form an embedding where similarity is expressed as a Euclidean distance. While effective at modeling human-perceived similarity judgments, these methods are expensive and require a large number of triplets to b...

Relationship Analysis

Both papers employ triplet-based similarity judgment tasks to assess representational alignment between language models and human conceptual knowledge, placing them in the same Similarity-Based Alignment Metrics category. They overlap in using ordinal embedding methods to derive semantic representations from triplet comparisons and measuring human-LLM alignment through distance-based metrics. However, the original paper systematically evaluates 77+ LLMs across diverse computational ingredients using the THINGS object concept dataset to identify architectural factors driving alignment, while the candidate paper focuses specifically on engineering design sketches and explores how different description templates (function, behavior, structure) affect LLM-human similarity agreement in a design context.

Contributions Analysis

Overall novelty summary. The paper systematically evaluates over 70 language models on triplet similarity judgments using concepts from the THINGS database, examining which computational ingredients (architecture, instruction-finetuning, training data) predict human-LLM representational alignment. It resides in the Similarity-Based Alignment Metrics leaf, which contains four papers total. This leaf sits within the broader Alignment Assessment Methodologies branch, indicating a moderately populated research direction focused on quantifying human-model correspondence through distance-based and similarity measures rather than behavioral or neural approaches.

The taxonomy reveals several neighboring methodological branches: Abstraction and Relational Alignment (2 papers) uses graph-based structural representations, while Cross-Linguistic and Cross-Cultural Alignment (4 papers) examines consistency across languages. The parent branch Alignment Assessment Methodologies excludes studies of emergent representations without measurement (those belong under Conceptual Representation Emergence) and behavioral comparisons (Behavioral Alignment). The paper's focus on triplet tasks positions it squarely within similarity-based methods, distinct from the brain-based approaches in Neural and Brain-Based Alignment (5 papers) or the multimodal studies in Multimodal Conceptual Alignment (3 papers).

Among 27 candidates examined across three contributions, none were identified as clearly refuting the work. The systematic evaluation of computational ingredients examined 10 candidates with 0 refutable; the model-fair comparison methodology examined 7 candidates with 0 refutable; and the benchmark-alignment relationship analysis examined 10 candidates with 0 refutable. This limited search scope suggests the specific combination of large-scale model comparison (70+ models), triplet similarity methodology, and computational ingredient analysis may represent a relatively underexplored configuration within the similarity-based alignment literature, though the small candidate pool prevents definitive conclusions about novelty.

The analysis covers top-K semantic matches and citation expansion within a 27-paper scope, not an exhaustive field survey. The absence of refutable candidates may reflect either genuine novelty in the specific methodological combination or limitations in search coverage. The taxonomy context suggests the paper contributes to an active but not overcrowded research direction, with the Similarity-Based Alignment Metrics leaf representing one of several complementary approaches to measuring human-model representational correspondence.

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

Contribution 1: Systematic evaluation of computational ingredients predicting human-LLM representational alignment

Description: The authors systematically evaluate 77+ language models varying in architecture, fine-tuning methods, training data, and other computational ingredients using a triplet similarity task with concepts from the THINGS database. They identify which ingredients (e.g., instruction fine-tuning, attention head dimensionality) most strongly predict alignment between model and human conceptual 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. Blackbox meets blackbox: Representational similarity and stability analysis of neural language models and brains

URL: View paper

Brief Assessment

Blackbox Representational Similarity[77] focuses on comparing neural language model representations to brain fMRI data during story reading, not on systematically evaluating computational ingredients (architecture, fine-tuning, training data) that predict human-LLM alignment on conceptual similarity tasks.

---

2. Brains and language models converge on a shared conceptual space across different languages

URL: View paper

Brief Assessment

Shared Conceptual Space[72] focuses on cross-language neural alignment using fMRI and multilingual language models, not on systematically evaluating computational ingredients (architecture, fine-tuning, training data) that predict human-LLM representational alignment on conceptual similarity tasks.

---

3. Optimizing human-controlled preference alignment in large language models via dense token masking: A methodological approach

URL: View paper

Brief Assessment

Dense Token Masking[76] focuses on preference alignment optimization through token masking techniques, not on systematic evaluation of computational ingredients (architecture, fine-tuning methods, training data) that predict human-LLM representational alignment using triplet similarity tasks.

---

4. Do Large Language Models Think Like the Brain? Sentence-Level Evidences from Layer-Wise Embeddings and fMRI

URL: View paper

Brief Assessment

LLMs Think Brain[78] focuses on layer-wise neural alignment between LLM embeddings and fMRI brain activity during sentence comprehension, not on systematically evaluating computational ingredients (architecture, fine-tuning methods, training data) that predict representational alignment.

---

5. Seal: Systematic error analysis for value alignment

URL: View paper

Brief Assessment

Seal Error Analysis[69] focuses on evaluating reward models in RLHF pipelines for value alignment (helpfulness/harmlessness), not on systematic evaluation of computational ingredients (architecture, fine-tuning methods, training data) predicting human-LLM representational alignment on conceptual similarity tasks.

---

6. From representation to response: assessing the alignment of large language models with human judgment patterns

URL: View paper

Brief Assessment

Representation to Response[73] focuses on evaluating LLMs using odd-one-out triplet tasks with image captioning strategies, not on systematically varying computational ingredients (architecture, fine-tuning methods, training data) to identify which predict alignment.

---

7. The neural architecture of language

URL: View paper

Brief Assessment

Neural Architecture Language[70] appears to focus on neural representations in language processing broadly, but the provided candidate text is too fragmentary to assess whether it systematically evaluates computational ingredients of LLMs for human alignment using methods like triplet similarity tasks.

---

8. Adaptive Token Boundaries: Integrating Human Chunking Mechanisms into Multimodal LLMs

URL: View paper

Brief Assessment

Adaptive Token Boundaries[71] focuses on dynamic cross-modal tokenization for multimodal LLMs and human-model alignment in visual-linguistic tasks, not on systematic evaluation of computational ingredients (architecture, fine-tuning methods, training data) across 77+ language models using triplet similarity tasks.

---

9. Uncovering the Computational Ingredients of Human-Like Representations in LLMs

URL: View paper

Brief Assessment

Computational Ingredients Representations[75] is the same paper as the original submission. The candidate paper IS the original paper itself, making comparison impossible. This appears to be a retrieval error where the paper was matched against itself.

---

10. Analyzing encoded concepts in transformer language models

URL: View paper

Brief Assessment

Encoded Concepts Transformers[74] analyzes how latent concepts are encoded in transformer representations using clustering and alignment with human-defined linguistic concepts, not systematic evaluation of computational ingredients (architecture, fine-tuning, training data) predicting human-LLM representational alignment.

Contribution 2: Model-fair comparison methodology using triadic similarity judgments

Description: The authors develop a species-fair comparison approach by administering the same triadic similarity judgment task to both models and humans, then deriving semantic embeddings using analogous methods. This ensures that discrepancies in alignment are not attributable to different embedding methods or unfair comparisons across model families.

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. Generalized Conditional Similarity Learning via Semantic Matching

URL: View paper

Brief Assessment

Conditional Similarity Learning[54] focuses on learning multiple feature spaces for conditional image similarity under different semantic conditions (e.g., supervised/weakly-supervised CSL), not on comparing human vs. model representations using triadic judgments.

---

2. Mianet: Aggregating unbiased instance and general information for few-shot semantic segmentation

URL: View paper

Brief Assessment

Mianet Few Shot[51] focuses on few-shot semantic segmentation using visual features and word embeddings, not on comparing model representations through triadic similarity judgments or semantic embedding alignment methodologies.

---

3. Teleological Vectors: A Mathematical Framework for Semantic Goal Alignment

URL: View paper

Brief Assessment

Teleological Vectors Alignment[58] focuses on goal-action alignment in cybernetic systems using reward models, not on triadic similarity judgments for comparing semantic embeddings across models and humans.

---

4. : Temporal Heterogeneous Information Network Embedding in Hyperbolic Spaces

URL: View paper

Brief Assessment

Hyperbolic Network Embedding[55] focuses on temporal heterogeneous information network embedding in hyperbolic spaces for graph-structured data, not on comparing semantic representations between humans and language models using triadic similarity tasks.

---

5. Correcting the triplet selection bias for triplet loss

URL: View paper

Brief Assessment

Triplet Selection Bias[53] focuses on correcting selection bias in triplet loss for metric learning in computer vision tasks, not on model-fair comparison methodologies for semantic embeddings using triadic similarity judgments.

---

6. Identifying ambiguous similarity conditions via semantic matching

URL: View paper

Brief Assessment

Ambiguous Similarity Conditions[52] focuses on weakly supervised conditional similarity learning for image retrieval tasks, not on comparing human and model semantic representations through triadic similarity judgments.

---

7. TriCon-Fair: Triplet Contrastive Learning for Mitigating Social Bias in Pre-trained Language Models

URL: View paper

Brief Assessment

TriCon Fair Bias[57] focuses on mitigating social bias in language models through triplet contrastive learning, not on comparing semantic embeddings between models and humans using triadic similarity judgments.

Contribution 3: Analysis of alignment-benchmark relationships revealing benchmarking gaps

Description: The authors demonstrate that existing LLM benchmarks (e.g., BigBenchHard, MMLU) correlate with representational alignment to varying degrees, but none fully captures alignment variance. This reveals a key gap in current LLM evaluation practices and highlights the insufficiency of standard benchmarks for measuring human-AI alignment.

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 vision-language model alignment and misalignment: A survey through the lens of explainability

URL: View paper

Brief Assessment

Vision Language Misalignment[65] focuses on vision-language model alignment and misalignment phenomena, not on LLM benchmark relationships with representational alignment or evaluation gaps in language-only models.

---

2. MT-RAIG: Novel Benchmark and Evaluation Framework for Retrieval-Augmented Insight Generation over Multiple Tables

URL: View paper

Brief Assessment

MT RAIG Benchmark[64] focuses on retrieval-augmented insight generation over multiple tables and proposes an evaluation framework for table-based reasoning. This is a completely different domain from the original paper's analysis of LLM representational alignment with human conceptual representations using triplet similarity tasks.

---

3. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges

URL: View paper

Brief Assessment

Vision Language Alignment[62] focuses on vision-language model benchmarks and evaluation practices, not on LLM representational alignment or the specific correlation analysis between alignment metrics and benchmarks like BigBenchHard and MMLU that the original paper investigates.

---

4. Multimodal Cultural Safety: Evaluation Frameworks and Alignment Strategies

URL: View paper

Brief Assessment

Multimodal Cultural Safety[68] focuses on cultural safety evaluation in vision-language models, not on representational alignment or the relationship between LLM benchmarks and human conceptual representations.

---

5. BRACE: A Benchmark for Robust Audio Caption Quality Evaluation

URL: View paper

Brief Assessment

BRACE Audio Caption[59] focuses on evaluating audio caption quality metrics in reference-free settings, not on analyzing relationships between LLM benchmarks and representational alignment with human cognition.

---

6. Resi: A comprehensive benchmark for representational similarity measures

URL: View paper

Brief Assessment

Resi Similarity Benchmark[60] focuses on evaluating representational similarity measures across neural network architectures using grounded similarity tests. It does not examine relationships between LLM benchmarks and representational alignment to humans, which is the core novelty claim of the original paper.

---

7. Visual Representation Alignment for Multimodal Large Language Models

URL: View paper

Brief Assessment

Visual Representation Alignment[66] focuses on aligning visual representations in multimodal models with vision foundation models to improve vision-centric tasks, not on analyzing relationships between representational alignment and LLM benchmarks or revealing evaluation gaps in benchmark design.

---

8. Agreements 'in the wild': Standards and alignment in machine learning benchmark dataset construction

URL: View paper

Brief Assessment

Benchmark Dataset Construction[63] focuses on ethnographic analysis of dataset construction practices and alignment work in creating benchmark datasets, not on evaluating correlations between LLM benchmarks and representational alignment metrics.

---

9. A benchmarking study of embedding-based entity alignment for knowledge graphs

URL: View paper

Brief Assessment

Entity Alignment Benchmark[61] focuses on entity alignment between knowledge graphs using embedding techniques, not on LLM representational alignment or benchmark evaluation gaps. The candidate addresses a completely different domain (KG entity matching) than the original's focus on human-AI representational alignment and LLM benchmark insufficiency.

---

10. How aligned are different alignment metrics?

URL: View paper

Brief Assessment

Alignment Metrics Aligned[67] focuses on comparing different alignment metrics (neural vs. behavioral, attention maps, odd-one-out tasks) to assess their internal consistency and aggregation methods. The original paper examines how standard LLM benchmarks (MMLU, BigBenchHard) correlate with representational alignment measured via triplet similarity tasks on conceptual embeddings. These are fundamentally different evaluation paradigms—one compares alignment measurement methods themselves, the other relates task performance benchmarks to semantic representation alignment.

Appendix: Text Similarity Detection

Textual similarity detection checked 31 papers and found 1 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. Uncovering the Computational Ingredients of Human-Like Representations in LLMs

Detected in: Contribution: contribution_1

⚠ **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] Uncovering the computational ingredients that support human-like conceptual representations in large language models View paper
- [1] Mapping language models to grounded conceptual spaces View paper
- [2] Aligning large language models with human: A survey View paper
- [3] Dissociating language and thought in large language models View paper
- [4] Revealing emergent human-like conceptual representations from language prediction. View paper
- [5] Abstraction Alignment: Comparing Model and Human Conceptual Relationships View paper
- [6] Semantics-aware BERT for language understanding View paper
- [7] Do large language models â□□understandâ□□ their knowledge? View paper
- [8] Abstraction Alignment: Comparing Model-Learned and Human-Encoded Conceptual Relationships View paper
- [9] Decoding Semantic Representations in the Brain Under Language Stimuli with Large Language Models View paper
- [10] Do large language models resemble humans in language use? View paper
- [11] Implicit representations of meaning in neural language models View paper
- [12] Human-like conceptual representations emerge from language prediction View paper
- [13] Unpacking large language models with conceptual consistency View paper
- [14] The Representational Alignment between Humans and Language Models is implicitly driven by a Concreteness Effect View paper
- [15] Large language model enabled semantic communication systems View paper
- [16] Seeing Through Words, Speaking Through Pixels: Deep Representational Alignment Between Vision and Language Models View paper
- [17] Can language models understand physical concepts? View paper
- [18] Improving semantic understanding in speech language models via brain-tuning View paper
- [19] Meaning and understanding in large language models View paper
- [20] Skill Graph Construction From Semantic Understanding View paper
- [21] Semantic Importance-Aware Communications with Semantic Correction Using Large Language Models View paper
- [22] Distinct social-linguistic processing between humans and large audio-language models: Evidence from model-brain alignment View paper
- [23] Exploring semantics in pretrained language model attention View paper
- [24] In conversation with artificial intelligence: aligning language models with human values View paper
- [25] From form (s) to meaning: Probing the semantic depths of language models using multisense consistency View paper
- [26] Do language models have semantics? on the five standard positions View paper
- [27] Language models align with brain regions that represent concepts across modalities View paper
- [28] Against AI understanding and sentience: large language models, meaning, and the patterns of human language use View paper
- [29] Identifying and interpreting non-aligned human conceptual representations using language modeling View paper
- [30] Large-Language-Model-Enabled Text Semantic Communication Systems View paper
- [31] A Flexible Method for Behaviorally Measuring Alignment Between Human and Artificial Intelligence Using Representational Similarity Analysis View paper
- [32] Human-like object concept representations emerge naturally in multimodal large language models View paper
- [33] Analyzing Fine-tuning Representation Shift for Multimodal LLMs Steering alignment View paper
- [34] Language Model Alignment in Multilingual Trolley Problems View paper
- [35] Exploring Human and Language Model Alignment in Perceived Design Similarity Using Ordinal Embeddings View paper
- [36] Concept alignment View paper
- [37] Do large language models resolve semantic ambiguities in the same way as humans? The case of word segmentation in Chinese sentence reading View paper
- [38] Human Preferences for Constructive Interactions in Language Model Alignment View paper
- [39] A study on semantic understanding of large language models from the perspective of ambiguity resolution View paper
- [40] The problem of alignment View paper
- [41] Conceptual Modeling and Large Language Models: Impressions From First Experiments With ChatGPT View paper
- [42] Align to Structure: Aligning Large Language Models with Structural Information View paper
- [43] Supporting Co-Adaptive Machine Teaching through Human Concept Learning and Cognitive Theories View paper
- [44] Frame Representation Hypothesis: Multi-Token LLM Interpretability and Concept-Guided Text Generation View paper
- [45] Divergences between Language Models and Human Brains View paper

- [46] Exploring Multilingual Concepts of Human Values in Large Language Models: Is Value Alignment Consistent, Transferable and Controllable across Languages? View paper
- [47] Exploring Multilingual Concepts of Human Value in Large Language Models: Is Value Alignment Consistent, Transferable and Controllable across Languages? View paper
- [48] Dual Embodied-Symbolic Concept Representations for Deep Learning View paper
- [49] From Tokens to Thoughts: How LLMs and Humans Trade Compression for Meaning View paper
- [50] Exploring Multilingual Human Value Concepts in Large Language Models: Is Value Alignment Consistent, Transferable and Controllable across Languages? View paper
- [51] Mianet: Aggregating unbiased instance and general information for few-shot semantic segmentation View paper
- [52] Identifying ambiguous similarity conditions via semantic matching View paper
- [53] Correcting the triplet selection bias for triplet loss View paper
- [54] Generalized Conditional Similarity Learning via Semantic Matching View paper
- [55] : Temporal Heterogeneous Information Network Embedding in Hyperbolic Spaces View paper
- [56] Unbiased Video Scene Graph Generation via Visual and Semantic Dual Debiasing View paper
- [57] TriCon-Fair: Triplet Contrastive Learning for Mitigating Social Bias in Pre-trained Language Models View paper
- [58] Teleological Vectors: A Mathematical Framework for Semantic Goal Alignment View paper
- [59] BRACE: A Benchmark for Robust Audio Caption Quality Evaluation View paper
- [60] Resi: A comprehensive benchmark for representational similarity measures View paper
- [61] A benchmarking study of embedding-based entity alignment for knowledge graphs View paper
- [62] A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges View paper
- [63] Agreements 'in the wild': Standards and alignment in machine learning benchmark dataset construction View paper
- [64] MT-RAIG: Novel Benchmark and Evaluation Framework for Retrieval-Augmented Insight Generation over Multiple Tables View paper
- [65] Large vision-language model alignment and misalignment: A survey through the lens of explainability View paper
- [66] Visual Representation Alignment for Multimodal Large Language Models View paper
- [67] How aligned are different alignment metrics? View paper
- [68] Multimodal Cultural Safety: Evaluation Frameworks and Alignment Strategies View paper
- [69] Seal: Systematic error analysis for value alignment View paper
- [70] The neural architecture of language View paper
- [71] Adaptive Token Boundaries: Integrating Human Chunking Mechanisms into Multimodal LLMs View paper
- [72] Brains and language models converge on a shared conceptual space across different languages View paper
- [73] From representation to response: assessing the alignment of large language models with human judgment patterns View paper
- [74] Analyzing encoded concepts in transformer language models View paper
- [75] Uncovering the Computational Ingredients of Human-Like Representations in LLMs View paper
- [76] Optimizing human-controlled preference alignment in large language models via dense token masking: A methodological approach View paper
- [77] Blackbox meets blackbox: Representational similarity and stability analysis of neural language models and brains View paper
- [78] Do Large Language Models Think Like the Brain? Sentence-Level Evidences from Layer-Wise Embeddings and fMRI View paper