Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Cognitive Alignment Between Humans and LLMs Across Multimodal Domains

Yuwei Wang

yuwei.wang@ia.ac.cn

Institute of Automation, Chinese Academy of Sciences, Beijing https://orcid.org/0000-0003-2677-6563

Dongqi Liang

Institute of Automation, Chinese Academy of Sciences, Beijing

Yi Zeng

Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, Beijing https://orcid.org/0000-0002-9595-9091

Article

Keywords: Large Language Models, Cognitive Alignment, Cognitive Boundaries, Cross-modality Consistency, Concept Learning

Posted Date: January 29th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-5736241/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Cognitive Alignment Between Humans and LLMs Across Multimodal Domains

Yuwei Wang[1,3,4,5], Dongqi Liang[1,2,3,4,5], Yi Zeng[1,2,3,4,5*]

[1*] Brain-inspired Cognitive Intelligence Lab, Institute of Automation, Chinese Academy of Sciences, No.95 Zhongguancun East Road Haidian District, Beijing, 100190, China.
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China.
[3] Beijing Institute of AI Safety and Governance, Beijing, China.
[4] Beijing Key Laboratory of Artificial Intelligence Safety and Superalignment, Beijing, China.
[5] Center for Long-term Artificial Intelligence, Beijing, China.

*Corresponding author(s). E-mail(s): yi.zeng@ia.ac.cn;
Contributing authors: yuwei.wang@ia.ac.cn; liangdongqi2023@ia.ac.cn;

Abstract

Large Language Models (LLMs) show remarkable text abilities, prompting investigations into their alignment with human cognition. Using the Brain-Based Componential Semantic Representation (BBSR), a neurobiologically grounded semantic framework, we evaluate nine LLMs, including Qwen2, Llama-3, Llama-3.1, and GPT series. We examine their multimodal cognitive boundaries, representational similarity, cross-modality consistency, abstract/concrete divergences, psycholinguistic factors, and stability across repeated responses. Larger models align more closely with human cognition, particularly for concrete concepts and early-acquired words. Still, discrepancies persist, especially for abstract concepts, spatial cognition, embodied experiences (e.g., olfaction, gustation), and causal reasoning. These findings reveal limitations in LLM cognitive architectures, emphasizing models embodying human cognition.

Keywords: Large Language Models, Cognitive Alignment, Cognitive Boundaries, Cross-modality Consistency, Concept Learning

1 Introduction

Conceptual learning, a cornerstone of human cognition, is deeply rooted in the integration of multimodal sensory and motor experiences. From a cognitive psychology perspective, humans acquire, store, and retrieve conceptual knowledge by synthesizing information derived from diverse sensory modalities—including vision, sound, touch, olfaction, gustation and motor interactions[1][2]. This multisensory integration enables humans to form semantic representations that are grounded in their real-world experiences, reflecting the interplay of perceptual and contextual knowledge [3]. In computational neuroscience, theories such as the embodied cognition framework and dual coding theory emphasize how the human brain encodes concepts through distributed neural systems[1][2][4][5][6], which span sensory-specific and heteromodal cortical areas[7][8] . This multisensory integration, facilitated by dedicated neural circuits and processes within the brain[9], allows humans to form holistic understandings of concepts, where meaning is not only symbolic but also grounded in perceptual and motor simulations[2]. This integrated, multimodal approach to learning and representation underscores the complexity of human conceptual processing.

Unlike human cognition, which is grounded in multimodal sensory experience, the advent of LLMs has revolutionized artificial intelligence by enabling machines to learn and process vast amounts of linguistic data. Models such as GPT-3.5 [10], GPT-4 [11], PaLM[12], GLM[13], and the Llama[14] and Qwen[15] series have demonstrated remarkable proficiency in natural language understanding, generation, and reasoning tasks. These models power a wide range of applications. However, despite their capabilities, these models rely predominantly on textual inputs[16]. This reliance on a single modality (language) raises critical questions about their ability to achieve true cognitive alignment with humans, particularly in domains that heavily rely on multisensory grounding. While LLMs can process and generate text describing sensory experiences, they lack the actual embodied experiences that are crucial for human conceptualization. This absence of direct perceptual and motor interaction with the world may lead to discrepancies in how LLMs and humans represent and process concepts. This fundamental difference raises questions about the extent to which LLMs align with human cognition, particularly in their ability to represent concepts across multimodal domains.

Datasets designed to quantify human conceptual representations have played a crucial role in cognitive science and are increasingly valuable for understanding and evaluating LLMs. These datasets, which aim to capture how humans represent and process concepts, can be broadly categorized based on their data format: property-based and rating-based. Property-based datasets describe concepts as collections of discrete features or attributes(e.g., "has wheels," "is edible"), often gathered through feature listing tasks, providing rich, interpretable descriptions, exemplified by the McRae feature production norms[17] and the CSLB concept property norms[18]. Although offering valuable insights into conceptual structure, property-based approaches face challenges in standardization due to subjective feature selection and struggle to quantify the relative importance of individual features. Rating-based datasets, conversely, employ quantitative scales to measure the strength of various perceptual, semantic, and affective attributes, capturing graded, multidimensional representations. The

Lancaster Sensorimotor Norms[19] offer comprehensive ratings for a wide range of sensorimotor properties, while the Bristol Norms[20] focus on affective dimensions like valence, arousal, and dominance. Similar to the modality exclusivity norms for nouns[21][22], these datasets quantify the association between concepts and specific sensory modalities (e.g., visual, auditory, tactile, gustatory, and olfactory) , enabling a more nuanced understanding of multimodal representation.

Among rating-based approaches, the Brain-Based Componential Semantic Representation (BBSR) dataset [23] provides a distinctive perspective by integrating behavioral ratings with neuroimaging data, thereby bridging neural and psychological levels of analysis. It quantifies the contributions of 65 dimensions, grouped into 14 domains spanning vision (e.g., brightness, color, motion), somatic (touch, temperature, weight, pain), audition (sound, music, speech), gustation (taste), olfaction (smell), motor (head, limb use, practice), spatial (landmark, path, near, away), temporal (time, duration), causal (caused, consequential), social (social, human, communication, self), cognition (cognition), emotion (benefit, harm, pleasant, etc.), drive (drive, needs), and attention (attention, arousal). By assessing these multidimensional aspects across varied concepts, BBSR reflects how human semantics is distributed over sensory, motor, and affective components.

Leveraging the BBSR dataset, we investigate the alignment between human and LLM cognition across multimodal domains, focusing on (1) the cognitive boundaries of multimodal representations in LLMs, (2) their alignment with human conceptual representations across different modalities and categories, and (3) the stability and consistency of their representations , enabling a fine-grained comparison of how LLMs and humans represent grounded concepts. By offering critical insights into the limitations of current LLMs in capturing the richness of human cognition, this study paves the way for future research on developing more embodied and multimodal AI systems, ultimately contributing to a deeper understanding of conceptual representation in both humans and machines and informing the design of more cognitively aligned AI.

2 Results

2.1 Cognitive Representations in LLMs

2.1.1 Comparative Multimodal Cognitive Boundaries Between LLMs and Humans

The BBSR dataset encompasses 535 human-labeled concepts, each annotated with cross-modal semantic associations. To explore the cognitive boundaries between humans and LLMs, we computed the mean representation vectors for both human-labeled concepts and concept representations generated by LLMs. These vectors were then employed to construct cognitive space profiles for both humans and LLMs, which were visualized for comparative analysis.

The radar charts presented in Figure 1 vividly illustrate the cognitive boundaries between human and LLM representations across 65 distinct dimensions of multimodal conceptual profiles. These profiles capture the extent to which each LLM approximates human cognitive boundaries, which are defined by the intricate interrelations between

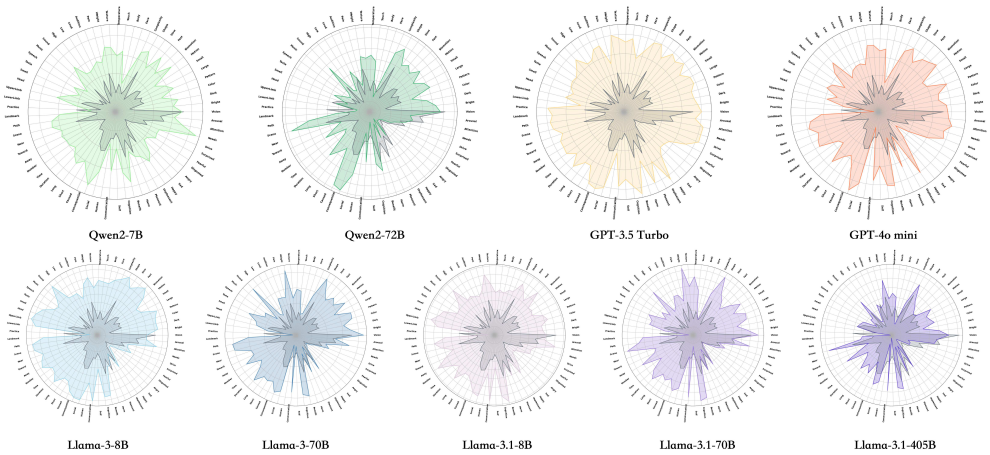


Fig. 1 Comparison of Cognitive Boundaries Between Humans and LLMs Across Multimodal Conceptual Profiles

sensory experiences and conceptual cognition. By comparing these cognitive boundaries, we reveal both the convergences and divergences in how LLMs and humans perceive and integrate multimodal information.

A clear pattern emerges from the data: larger, more sophisticated models exhibit a broader, more consistent alignment with human cognitive boundaries. For instance, models like **Llama-3.1-405B** and **Qwen2-72B** demonstrate a more expansive cognitive boundary that aligns more closely with the human profile across various conceptual dimensions. In contrast, smaller models, such as **Llama-3.1-8B** or **Llama-3-8B**, show a narrower alignment, limited to fewer dimensions. This trend suggests that increasing model size and complexity enables a more comprehensive representation of the nuanced and interconnected boundaries between sensory modalities and conceptual constructs.

However, even the largest models, such as **Llama-3.1-405B**, fail to fully encapsulate the full range and consistency of human cognitive boundaries. This indicates fundamental limitations in current neural architectures, particularly in their ability to replicate the embodied nature of human cognition, where sensory experiences are dynamically integrated. Despite their increased parameter sizes, these models do not capture the rich, cross-modal integration that forms the essence of human cognition, thereby leaving gaps in the understanding of cognitive boundaries.

Moreover, within model families, there is a clear and consistent trend: as the number of parameters increases, the cognitive representations generated by LLMs gradually shift closer to the human cognitive boundaries. This suggests that, as model capacity expands, LLMs are converging toward a more human-like understanding of multimodal conceptual semantics. Nevertheless, these boundaries remain asymptotic, highlighting that even the most advanced models fail to fully approximate the human cognitive space.

The observed discrepancies between LLM and human multimodal cognitive boundaries underline the limitations of unimodal training paradigms. While human cognition

integrates sensory inputs from modalities such as vision, touch, and sound into a unified conceptual framework, most LLMs are trained on textual data alone, inherently lacking direct sensory grounding. This restricts the models' ability to capture the full spectrum of human multisensory integration, making them insufficient to represent the complexities of human multimodal conceptualization.

2.1.2 Semantic Neighborhoods and Conceptual Proximity in LLMs

The representation of semantic neighborhoods and conceptual proximity plays a crucial role in understanding how LLMs approximate human cognitive structures. In this study, we evaluate the ability of various LLMs to replicate human conceptual similarity judgments by using the mean Precision at k (mP@k) as a metric for comparison.

Table 1 Performance metrics (mP@k) for different LLMs across varying k.

LLM	**mP@1**	**mP@3**	**mP@5**	**mP@10**	**mP@15**	**mP@20**
Qwen2-7B	0.1276	0.1764	0.2079	0.2606	0.3009	0.3239
Qwen2-72B	0.1839	0.2395	0.2702	0.3242	0.3601	0.3826
Llama-3-8B	0.0788	0.1201	0.1538	0.1949	0.2225	0.2480
Llama-3-70B	0.1614	0.1945	0.2304	0.2773	0.3041	0.3257
Llama-3.1-8B	0.0450	0.0688	0.0904	0.1366	0.1686	0.1932
Llama-3.1-70B	0.1257	0.1839	0.2195	0.2687	0.3062	0.3345
Llama-3.1-405B	0.1801	0.2514	0.2814	0.3383	0.3714	0.3934
GPT-3.5 Turbo	0.1051	0.1326	0.1670	0.2150	0.2417	0.2649
GPT-4o mini	0.1463	0.1976	0.2218	0.2679	0.2963	0.3162

To compute mP@k, we first establish the top-k most similar concepts in the human cognitive space for each target concept. These serve as the ground truth for measuring semantic proximity. Next, we retrieve the top-k most similar concepts from the cognitive space of each LLM, focusing on the number of overlaps between the human-labeled and model-generated top-k lists. The Precision at k (P@k) is calculated for each concept as follows:

$$P@k = \frac{|Overlap(Human_k, LLM_k)|}{k}$$

Consider the concept "*book*" with k = 5. Assume the top-5 most similar concepts in the human cognitive space ($Human_5$) are *computer, magazine, newspaper, camera, cash*, and a particular LLM returns *newspaper, magazine, camera, flower, bridge* as its top-5 similar concepts (LLM_5). The intersection of these two sets is *newspaper, magazine, camera*, so $P@5 = \frac{3}{5} = 0.6$. The mean Precision at k (mP@k) is then computed by averaging the P@k values across all concepts:

$$mP@k = \frac{1}{C}\sum_{j=1}^{C} P@k(j)$$

where C is the total number of concepts evaluated. This metric provides a measure of how well each model captures human-like conceptual proximity within the top-k retrieved concepts.

A clear trend emerges from the results, showing that larger models generally achieve higher mP@k scores, indicating a stronger alignment with human conceptual similarity judgments within the top-k retrieved concepts. For instance, **Llama-3.1-405B** achieves the highest performance (e.g., mP@20 = 0.3934), while smaller models such as **Llama-3.1-8B** and **Llama-3-8B** exhibit significantly lower mP@k values across all thresholds (e.g., mP@20 = 0.1932 and 0.2480, respectively). This suggests that larger model sizes, often correlated with expanded training datasets and greater representational capacity, facilitate a more accurate encoding of semantic relationships as they are perceived by humans, within the top-k retrieved concepts. This finding is consistent with principles in cognitive informatics, which posit that more complex and expansive neural architectures are better suited to capturing the intricate, high-dimensional nature of human conceptual spaces, within a limited scope.

While larger models tend to outperform smaller ones, the rate of improvement diminishes as model size increases. For example, the performance gap between **Llama-3.1-70B** (mP@20 = 0.3345) and **Llama-3.1-405B** (mP@20 = 0.3934) is smaller than the gap between **Llama-3.1-8B** (mP@20 = 0.1932) and **Llama-3.1-70B**. Similarly, **Qwen2-72B** shows only a modest improvement over **Qwen2-7B** despite being an order of magnitude larger. This diminishing return could reflect the inherent limits of statistical learning approaches in fully capturing the complexity of human cognition, which not only involves statistical associations but also context-sensitive reasoning and experiential grounding, even when considering only the top-k most similar concepts.

Moreover, the relative rankings of models vary across different thresholds (e.g., mP@1 vs. mP@20), revealing that some models excel at capturing fine-grained conceptual similarities within a very narrow scope (e.g., mP@1), while others are more adept at handling broader comparisons within a larger scope (e.g., mP@20). For instance, **GPT-3.5 Turbo** demonstrates relatively low performance across all thresholds (e.g., mP@20 = 0.2649), suggesting that it lacks the necessary granularity for human-like conceptual reasoning, even when considering only the top retrieved results. In contrast, **GPT-4o mini** exhibits consistent performance across all thresholds, reflecting a more well-rounded semantic space that aligns with human cognition both locally and globally, within the top-k retrieved concepts.

These findings provide valuable insights into the comparative strengths and limitations of different LLMs in approximating human conceptual structures within a limited scope. The results underscore the importance of model size and complexity in capturing the multidimensional nature of human cognition within the top-k retrieved concepts while also highlighting the challenges posed by the limitations of current models in fully replicating the rich, context-dependent, and embodied nature of human semantic processing.

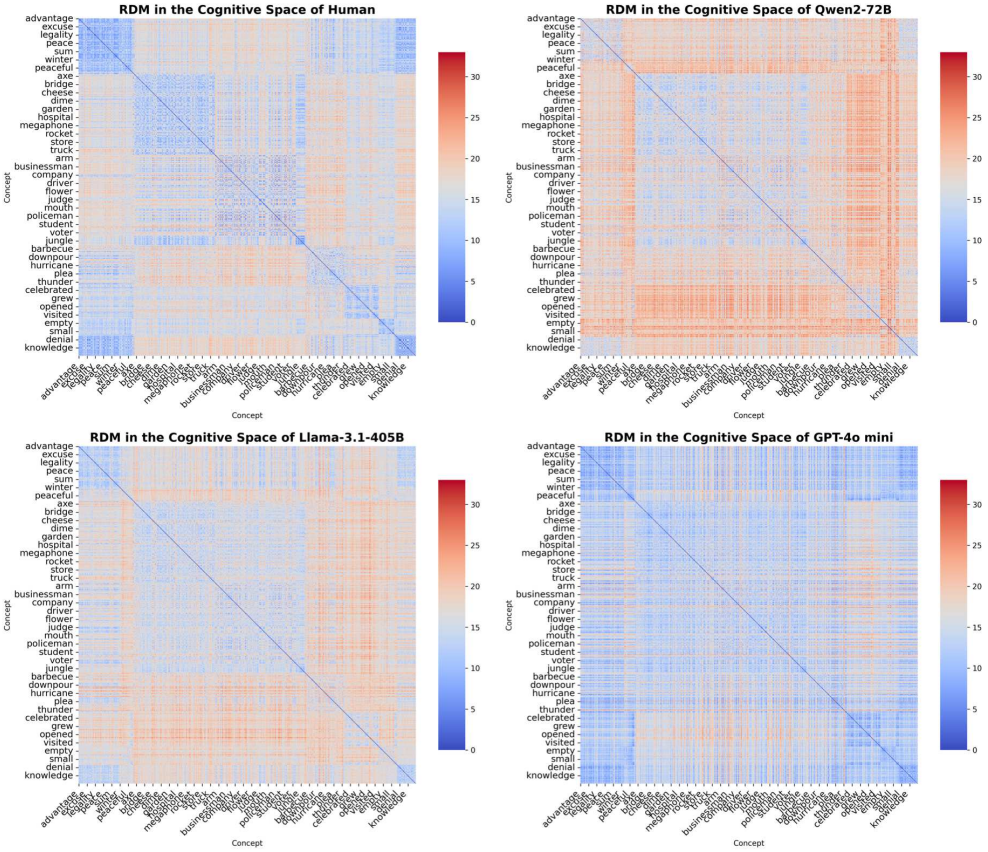


Fig. 2 Representational Dissimilarity Matrices (RDMs) Organized by Concept Categories

2.2 Alignment Between Human and LLM Cognition

2.2.1 Cross-Space Representational Similarity: Human vs. LLM Cognition

In this section, we explore the alignment between human and LLM cognitive spaces using Representational Similarity Analysis (RSA). Specifically, we compute the Representational Dissimilarity Matrices (RDMs) based on Euclidean distance for both humans and LLMs and visualize the alignment effects in Figure 2.

To represent the cognitive spaces of each model, we selected the largest models from each series: **Qwen2-72B**, **Llama-3.1-405B**, and **GPT-4o mini** for RDM visualization. The concepts in the RDM heatmap are organized according to the categories that reflects the semantic richness and complexity of human cognition: **abstract entity**, **abstract property**, **abstract action**, **artifact**, **living object**, **natural object**, **event**, **physical action**, **physical property**, **physical state**, **mental state**, and **mental entity**.

The RDMs provide key insights into the extent of alignment between human and LLM cognitive structures across various concept categories. Notably, all three LLMs demonstrated some degree of alignment with human cognitive representations, though this alignment varied significantly depending on the nature of the concepts being evaluated.

Table 2 Cognitive Alignment Index (CAI) Across Concept Categories for Different LLMs

LLM	ALL	Abstract Entity	Abstract Property	Artifact	Event	Living Object	Mental Entity	Natural Object	Physical Action	Physical Property
Qwen2-7B	0.4401	0.3774	0.4120	0.5099	0.6060	0.6475	0.5247	0.6607	0.5078	0.2433
Qwen2-72B	0.4122	0.3305	0.6835	0.6160	0.4762	0.6742	0.5131	0.1778	0.5536	0.4301
Llama-3-8B	0.2726	0.2653	0.2927	0.3207	0.3589	0.5377	0.3049	0.7925	0.3545	0.2637
Llama-3-70B	0.3932	0.4049	0.4624	0.5740	0.5099	0.6727	0.6329	0.4018	0.4467	-0.0187
Llama-3.1-8B	0.3150	0.2286	0.3087	0.3549	0.3577	0.5969	0.3933	0.5225	0.2330	0.0027
Llama-3.1-70B	0.4057	0.4167	0.2314	0.5709	0.5934	0.6901	0.7148	0.3988	0.4560	0.0696
Llama-3.1-405B	0.5337	0.5534	0.3250	0.6629	0.7018	0.7445	0.6355	0.6012	0.6094	0.4115
GPT-3.5 Turbo	0.3461	0.3775	0.6017	0.2799	0.4209	0.5128	0.5918	0.7129	0.3768	0.0956
GPT-4o mini	0.4222	0.4290	0.4676	0.4347	0.4679	0.5884	0.6775	0.7150	0.5925	0.1999

Furthermore, we introduce the Cognitive Alignment Index (CAI) as a quantitative measure to assess the alignment between the RDMs of LLMs and humans. The CAI is derived by computing the Spearman correlation between the lower triangular elements of the human and LLM RDMs. This approach captures the degree of similarity in the representational structures of both systems. Specifically, the lower triangular elements of the RDMs are defined as:

$$\text{RDM}_{\text{lower}} = \{R_{ij} \mid i < j\}$$

where R_{ij} represents the dissimilarity between the i-th and j-th concepts in the RDM. The Spearman correlation between these elements quantifies the degree of alignment between the human and LLM cognitive spaces, providing a robust metric for comparing their representational structures. The CAI is computed both at a macro level, reflecting the overall alignment across all concept categories, and at the category-specific level to reveal how well LLMs align with human cognitive structures within distinct conceptual domains.

The CAI for each concept category is presented in Table 2, offering a detailed view of how well each LLM approximates human conceptual representation across various semantic dimensions.

$$\rho_{\text{CAI}} = \text{Spearman}(\text{RDM}_{\text{human, lower}}, \text{RDM}_{\text{LLM, lower}})$$

This analysis allows for a nuanced understanding of the strengths and limitations of different LLMs in capturing human-like conceptual structures.

An important consideration in our analysis is the concept of "used", which appears twice in the BBSR dataset as a polyseme. To avoid the confounding effects of this

ambiguity, we excluded the concept "used" from RSA. Furthermore, for the presentation of the CAI, we restricted our focus to concept categories containing more than five individual concepts, ensuring robust statistical representation for each category.

For abstract concepts such as "truth," "hope," and "irony," human RDMs revealed high dissimilarity, reflecting the intricate, multidimensional nature of human cognition in these domains. This is consistent with the view that abstract concepts are grounded in a complex web of contextual and affective associations. Among the LLMs, **Qwen2-72B** exhibited the strongest alignment with human representations in this domain, suggesting that it is better equipped to capture the nuanced structure of abstract reasoning compared to its counterparts. This finding highlights the model's potential to better approximate human abstract cognitive processing, though it still falls short of fully replicating the depth of human conceptualization.

In contrast, for concrete concept categories such as **artifacts** (e.g., "chair," "car") and **living objects** (e.g., "dog," "tree"), all models demonstrated a higher degree of alignment with human RDMs, reflecting the relative ease with which LLMs encode physical, tangible concepts. Among these models, **Llama-3.1-405B** performed slightly better, suggesting a robust capacity to model categorical hierarchies and visual-like properties that are often more straightforward for machines to represent.

When considering **events** (e.g., "celebrated," "storm") and **physical actions** (e.g., "walked," "threw"), the alignment was moderate across all LLMs, with some slight deviations observed in the **GPT-4o mini** model. This suggests that while LLMs can capture the basic structure of dynamic, physical phenomena, there are subtle discrepancies in how these models represent temporal and causally complex events, especially in comparison to human cognition.

Interestingly, all models exhibited notable difficulty in representing **mental states** and **affective entities** (e.g., "hope," "grief"), which are deeply rooted in subjective and emotional experiences. This observation reinforces the hypothesis that LLMs are better suited to processing external, objective phenomena, which are more easily grounded in sensory or factual data, rather than the internal, subjective states that are central to human consciousness. The failure of LLMs to fully capture these dimensions highlights the limitations of current neural architectures in encoding the complexities of human affective and introspective cognition.

From a cognitive informatics perspective, these findings align with existing theories that posit human cognition as being hierarchically organized, with abstract concepts being more multidimensional and context-dependent than concrete ones. The differential performance of the LLMs across concept categories supports the notion that while LLMs can approximate certain aspects of this cognitive hierarchy, they lack the depth and flexibility required for higher-level abstraction. This observation resonates with findings from neuroscience, which have shown that abstract concepts are often represented by distributed neural networks spanning multiple brain regions, reflecting their inherent complexity and the integration of diverse cognitive processes.

2.2.2 Cross-Modality Consistency in Human and LLM Cognition

Human cognition is inherently multimodal, integrating information from various sensory modalities. To assess how well LLMs capture this crucial aspect of human

cognition, we now analyze cross-modality consistency in human and LLM representations.

Figure 3 presents an evaluation of the consistency between various LLMs and human judgments across multiple cognitive dimensions, using the mean squared error (MSE) as the primary metric (where lower MSE values indicate better performance). The cognitive dimensions analyzed encompass perceptual modalities (vision, somatic, audition), as well as temporal, spatial, and higher-order cognitive functions (emotion, social reasoning, and causal reasoning). We provide an overarching interpretation of the results, followed by detailed insights for each model.

Auditory concepts consistently achieve the lowest MSE values across all models, suggesting that auditory-related representations (e.g., speech, sound) are the easiest for LLMs to align with human cognition. This alignment likely stems from the strong linguistic association of auditory descriptors within the training corpora, where text-based LLMs are frequently exposed to auditory experiences described in verbal terms.

In contrast, **Vision** and **Somatic** reasoning exhibit relatively higher MSE values, reflecting the inherent difficulty of modeling purely perceptual and embodied experiences. These cognitive dimensions are fundamentally multimodal, requiring sensory grounding that is not inherently available to text-based LLMs. The higher MSE in these domains underscores the challenge of capturing the richness of visual and bodily experiences that humans naturally integrate through direct sensory input.

Temporal reasoning demonstrates superior performance relative to **Spatial** reasoning, as indicated by lower MSE values in the temporal domain. This suggests that LLMs are better equipped to handle sequential and time-related concepts (e.g., "before," "after") than spatial relationships (e.g., "above," "beside"). The relatively higher proficiency in temporal reasoning may reflect the frequent presence of temporal expressions in textual corpora, particularly in narrative structures that emphasize time progression and event sequencing.

The higher-order cognitive dimensions—**Emotion**, **Social** reasoning, and **Causal** reasoning—reveal the most significant discrepancies across models. Overall, the **Emotion** dimension exhibits relatively low MSE values across all evaluated LLMs. This suggests that emotional cognition is one of the more accessible domains for current LLMs to model effectively. The relatively strong performance in this dimension can likely be attributed to the high prevalence of emotional expressions and affective contexts in the training corpora, enabling models to capture and infer emotional states through linguistic cues. While differences in performance persist—such as **GPT-4o mini** achieving the lowest MSE, reflecting its superior ability to approximate human affective cognition—all models demonstrate a commendable capacity to align with human emotional reasoning. This suggests that despite the inherent complexity of emotional cognition, which is deeply rooted in subjective experiences and cultural context, LLMs have made considerable strides in capturing and simulating these characteristics through linguistic patterns. **Social** reasoning, in contrast, shows relatively consistent performance across the models, with moderate MSE values observed for all. This suggests that while LLMs are capable of approximating certain aspects of social cognition, their understanding remains rudimentary compared to the complexity of human social interactions, which involve a range of subtle cues, context, and cultural

knowledge. **Causal** reasoning emerges as the most challenging domain for all models, although **GPT-4o mini** exhibits a marginally lower MSE than its counterparts. This underscores the difficulty LLMs face in modeling causal relationships, which often require a deep understanding of temporal sequences, underlying mechanisms, and context-dependent inferences.

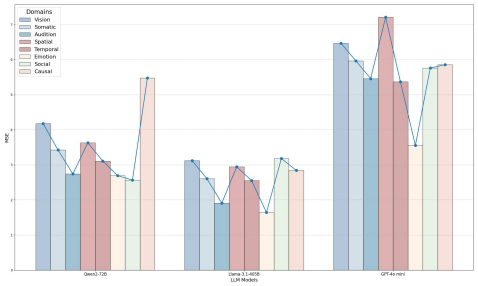


Fig. 3 Comparison of Cross-Modality Consistency Between Human and LLM Cognition

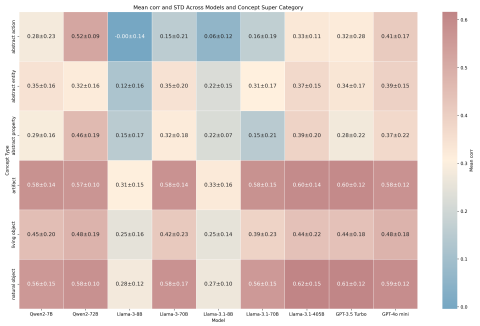


Fig. 4 Human-LLM Alignment by Abstract and Concrete Categories

2.2.3 Divergences in Abstract and Concrete Conceptual Representations

To further investigate the alignment between LLMs and human cognition across different conceptual domains, we focus on a comparative analysis of abstract and concrete concept categories. Specifically, we selected three abstract categories—**abstract actions**, **abstract entities**, and **abstract properties**—and three concrete categories—**artifacts**, **living objects**, and **natural objects**. For each category, we computed the mean Spearman correlation coefficients between the conceptual representations of humans and various LLMs, followed by a detailed visualization of the results.

Figure 4 presents a comprehensive comparison of the alignment between different LLMs and human conceptual understanding across these diverse semantic categories. The findings reveal complex patterns of alignment that vary depending on both the conceptual nature (abstract vs. concrete) and the architecture or scale of the LLM.

Across the examined categories, LLMs generally exhibit stronger alignment with human cognition in concrete concept categories than in abstract ones. For instance, the correlation coefficients for "artifact" and "natural object" consistently exceed 0.5 across most models, with **LLaMA-3.1-405B** and **GPT-3.5 Turbo** achieving the highest scores (0.60 ± 0.14 for **artifact** and 0.62 ± 0.15 for **natural object**). This pattern is in line with cognitive informatics research, which posits that concrete concepts, being directly grounded in sensory and perceptual experiences, are more readily encoded and understood by both humans and computational systems. The higher alignment in these categories likely reflects the models' capacity to leverage extensive

corpora of visual and textual data related to everyday physical objects, which are more straightforwardly represented in both human and machine cognition.

In contrast, abstract concepts—particularly **abstract action** and **abstract property**—show significantly lower alignment with human judgments across all LLMs. For example, **LLaMA-3-70B** demonstrates near-zero alignment for **abstract action** (0.00 ± 0.14), and even the best-performing models, such as **GPT-3.5 Turbo**, achieve only moderate correlations (e.g., 0.33 ± 0.11 for **abstract property**). This marked disparity underscores the inherent challenge in modeling abstract categories, which often lack direct sensory grounding and are more reliant on higher-order cognitive functions, including reasoning, cultural context, and domain-specific knowledge. The difficulty of encoding these more complex, intangible concepts reflects the limits of current LLM architectures, which are typically optimized for more concrete, perceptually grounded representations.

Furthermore, larger models, such as **LLaMA-3.1-405B** and **GPT-3.5 Turbo**, consistently outperform smaller models across both abstract and concrete categories, suggesting that model scale plays a significant role in enhancing conceptual understanding. This trend is particularly evident in abstract categories like **abstract entity** and **abstract property**, where smaller models, such as **LLaMA-3-70B**, demonstrate a consistent performance gap. The increased scale of larger models likely facilitates a more robust encoding of high-dimensional relationships and abstract reasoning, supporting the notion that larger parameter spaces enable more nuanced and comprehensive cognitive representations.

2.2.4 Psycholinguistic Determinants of Conceptual Alignment: Concreteness and AoA

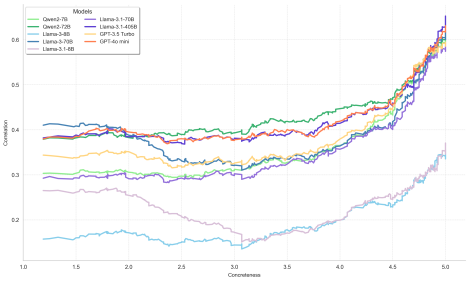


Fig. 5 Correlation Between LLM-Human Alignment and Concreteness

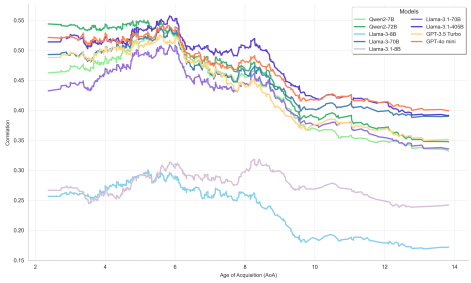


Fig. 6 Correlation Between LLM-Human Alignment and Age of Acquisition

In this section, we explore the psycholinguistic determinants, **Concreteness** and **Age of Acquisition (AoA)**, that underpin the alignment between LLMs and human semantic cognition. These two metrics serve as fundamental psycholinguistic indicators for quantifying the properties of concepts and provide critical insights into the extent to which LLMs capture human-like conceptual representations. While Concreteness reflects the sensory and perceptual grounding of concepts, AoA offers a window into the

developmental and experiential trajectory of word acquisition. Together, these metrics allow us to investigate the cognitive processes that influence conceptual alignment in both humans and LLMs.

The Concreteness ratings utilized in this study were sourced from the comprehensive dataset provided by Brysbaert et al.[24], which offers scores for over 40,000 English concepts. These ratings reflect the degree to which a concept denotes a tangible object or perceptual experience, as determined by crowd-sourced judgments. Similarly, AoA ratings, drawn from Kuperman et al.[25], provide age-of-acquisition norms for over 30,000 English words. AoA reflects the typical age at which a word is learned, capturing key developmental factors in lexical acquisition and offering valuable insights into the cognitive mechanisms underlying concept learning.

To assess the alignment between LLMs and human cognition, we evaluated 9 LLMs with varying architectures and model sizes. For each concept in the BBSR dataset, we calculated the Spearman correlation between the LLM-generated vectors and the human-derived Concreteness and AoA ratings. These correlations were computed across both the Concreteness and AoA scales, ensuring robustness by applying a smoothing window size of 100. The resulting correlation values were then plotted as functions of Concreteness and AoA, generating two distinct line plots for further analysis.

Figure 5 explores the relationship between LLM outputs and Concreteness ratings. The results show a clear trend: as Concreteness increases, the alignment between LLMs and human judgments strengthens, with the highest correlations observed for highly concrete words (Concreteness > 4.0). This suggests that LLMs excel in modeling the semantics of tangible, imageable concepts—those that are easily grounded in perceptual experiences. On the other hand, abstract words (Concreteness < 2.0) exhibit markedly lower correlations, indicating the inherent challenge in representing concepts that lack clear sensory grounding.

This pattern of stronger alignment for concrete words is consistent with findings in cognitive neuroscience, which have demonstrated that concrete words tend to elicit stronger activation in sensorimotor brain regions. This neural activation underscores the perceptual grounding of concrete concepts, which may be more easily encoded in LLMs that are trained on large corpora rich in sensory-based language. The relatively weaker alignment for abstract words suggests that LLMs may overly rely on co-occurrence patterns in textual data, which fail to fully capture the complex relational and context-dependent semantics of abstract concepts.

Figure 6 examines the correlation between LLM outputs and AoA ratings, focusing on the developmental timeline of concept acquisition. Notably, performance across all models peaks in the middle AoA range (5–8 years), corresponding to words typically acquired during early childhood education. This peak aligns with the "critical period hypothesis" in cognitive psychology, which posits that vocabulary acquisition during the early years is characterized by frequent exposure and the formation of robust semantic networks. It appears that LLMs are particularly adept at modeling this early-acquired vocabulary, effectively capturing the linguistic richness that emerges during this critical developmental phase.

However, the correlation decreases for words acquired later in life (AoA > 10 years), reflecting a decline in performance for terms that are typically learned in adolescence or adulthood. This drop may be due to the reduced frequency and contextual variability of these words in human experience, which in turn limits their representation in the training data of LLMs. Consequently, the models' performance appears to be influenced by the distribution of words in their training corpora, with less frequent, late-acquired words showing weaker alignment with human semantic judgments.

The results of these analyses highlight key psycholinguistic factors that drive the alignment between LLMs and human cognition: Concreteness and AoA. LLMs exhibit strong alignment with human semantic understanding for concepts that are grounded in sensory experience (concrete concepts) and those that are typically acquired early in life (concepts with a low AoA). This pattern is consistent with cognitive theories that emphasize the role of perceptual grounding and developmental experience in shaping human conceptual knowledge. However, the models' performance on abstract concepts and late-acquired words suggests that LLMs still face significant challenges in capturing the full complexity of human cognition, particularly for concepts that lack sensory grounding or are learned later in life.

2.3 Stability in LLM Cognitive Representations

In this section, we introduce the *Intraclass Correlation Coefficient (ICC)* as a metric for evaluating the stability of LLMs across repeated trials, specifically for various cognitive dimensions. Stability in LLM responses is crucial for assessing their reliability in modeling consistent human-aligned semantic representations. By computing the ICC, we assess the consistency of LLM-generated ratings over five independent trials across diverse cognitive dimensions.

Table 3 Intraclass Correlation Coefficients (ICCs) for LLMs Across Cognitive Domains

LLM_name	Audition	Vision	Somatic	Olfaction	Gustation	Spatial	Temporal	Emotion	Social	Causal
Qwen2-7B	0.8571	0.7288	0.5642	0.5916	0.6759	0.5403	0.5510	0.7526	0.8579	0.4236
Qwen2-72B	0.9888	0.9776	0.9760	0.9602	0.9652	0.9648	0.9810	0.9869	0.9838	0.8902
Llama-3-8B	0.7998	0.6007	0.5518	0.4377	0.5191	0.3687	0.4217	0.8506	0.6493	0.2419
Llama-3-70B	0.9433	0.8795	0.8512	0.8180	0.7749	0.9004	0.8964	0.9415	0.9105	0.6759
Llama-3.1-8B	0.4444	0.4235	0.3083	0.1761	0.1334	0.2146	0.1935	0.5912	0.4869	0.2011
Llama-3.1-70B	0.7668	0.7649	0.6615	0.4621	0.4137	0.4877	0.6863	0.8115	0.7002	0.3636
Llama-3.1-405B	0.8815	0.8449	0.8225	0.7469	0.6921	0.7605	0.8707	0.7457	0.8140	0.5108
GPT-3.5 Turbo	0.8011	0.6052	0.6025	0.6342	0.6884	0.4480	0.5836	0.7814	0.8358	0.2977
GPT-4o mini	0.9252	0.8237	0.8475	0.7901	0.8187	0.8045	0.9021	0.9149	0.9309	0.6605

To examine the domain-level stability, the 65 cognitive dimensions in BBSR dataset were grouped into broader conceptual domains, encompassing **Audition**, **Vision**, **Somatic**, **Olfaction**, **Gustation**, **Spatial**, **Temporal**, **Emotion**, **Social**, and **Causal**. Each of these domains contains a subset of related dimensions, and the ratings for all dimensions within each domain were aggregated for analysis. For each

domain m (where $m = 1, 2, ..., M$), we calculated the ICC, which quantifies the consistency of LLM responses across $T = 5$ independent trials. Mathematically, the ICC for a given domain m is defined as:

$$\text{ICC}_m = \frac{\sigma^2_{\text{between},m}}{\sigma^2_{\text{between},m} + \sigma^2_{\text{within},m}}$$

Where:

- ICC_m is the Intraclass Correlation Coefficient for domain m,
- $\sigma^2_{\text{between},m}$ is the variance of the mean ratings across all concepts and dimensions in domain m,
- $\sigma^2_{\text{within},m}$ is the variance of ratings across the 5 independent trials within domain m.

To compute these variances, let $y_{c,d,t}$ represent the rating produced by the LLM for concept c, dimension d, and trial t, and define:

$$\bar{y}_{c,d} = \frac{1}{T}\sum_{t=1}^{T} y_{c,d,t}, \quad \bar{y}_{c,m} = \frac{1}{|D_m|}\sum_{d \in D_m} \bar{y}_{c,d}, \quad \bar{y}_m = \frac{1}{C}\sum_{c=1}^{C} \bar{y}_{c,m}.$$

The between-domain variance is computed as:

$$\sigma^2_{\text{between},m} = \frac{1}{C}\sum_{c=1}^{C} \left(\bar{y}_{c,m} - \bar{y}_m\right)^2,$$

and the within-domain variance is given by:

$$\sigma^2_{\text{within},m} = \frac{1}{C|D_m|(T-1)}\sum_{c=1}^{C}\sum_{d \in D_m}\sum_{t=1}^{T} \left(y_{c,d,t} - \bar{y}_{c,d}\right)^2.$$

The ICC value ranges from 0 to 1, where a value closer to 1 indicates higher stability and consistency in the model's response across trials. The results, summarized in Table 3, reveal key insights into the stability of LLM-generated cognitive representations.

Perceptual Modalities: Within this category, we observe a clear hierarchy of stability. **Audition** consistently exhibits the highest ICC values across most models, echoing the findings on cross-modal consistency. This reinforces the notion that LLMs are particularly adept at capturing auditory information, likely due to its strong linguistic grounding in text. **Vision** and **Somatic** modalities demonstrate moderate stability, though generally lower than audition, reflecting the challenges of representing purely perceptual experiences. Notably, **Olfaction** and **Gustation** consistently show the lowest ICC values across all models, indicating significantly lower stability in these sensory domains. This suggests that LLMs struggle to consistently model these less perceptually grounded or less frequently represented sensory experiences.

Spatiotemporal Reasoning: Comparing **Spatial** and **Temporal** reasoning reveals a nuanced picture. While **Temporal** reasoning generally exhibits slightly higher stability than **Spatial** reasoning, particularly in larger models, this difference is less pronounced than the variations observed within the perceptual modalities. This suggests that while LLMs may be slightly better at handling sequential information, both spatial and temporal reasoning still pose challenges for consistent representation.

Higher-Order Cognitive Functions: Among these complex cognitive processes, **Emotion** demonstrates relatively high stability across most models, aligning with the findings on cross-modal consistency. This suggests that LLMs can effectively capture emotional content conveyed through language. **Social** reasoning exhibits moderate stability, indicating a reasonable but not perfect capacity to model social concepts. Finally, **Causal** reasoning presents the greatest challenge, consistently showing the lowest ICC values within this category, highlighting the difficulties LLMs face in capturing causal relationships. This mirrors the challenges observed in cross-modal consistency, where causal reasoning also proved difficult.

3 Discussion

This study offers a comprehensive investigation into the cognitive alignment between humans and LLMs across multimodal domains, using the neurobiologically grounded BBSR framework to evaluate nine LLMs on three fronts. First, we examined *Cognitive Representations in LLMs*, showing that larger models exhibit more human-like semantic neighborhoods but remain challenged by abstract and embodied concepts. Second, we assessed *Alignment Between Human and LLM Cognition* through representational similarity and cross-modality analyses, revealing stronger alignment for concrete or early-acquired concepts while uncovering notable challenges in spatial cognition and higher-order functions like causal reasoning. Third, we analyzed *Stability in LLM Cognitive Representations*, finding that scale correlates positively with consistency, though stability varies by domain—audition and emotion fare well, while olfaction, gustation, and causal reasoning remain problematic.

Despite advances in scale, training data, and architecture, limitations persist, highlighting the need to bridge gaps between artificial and human intelligence. Several promising research directions emerge:

- **Multimodal Training:** Current reliance on text-based data restricts LLMs from fully capturing human cognition's multisensory depth. Future endeavors should integrate vision, audition, haptics, olfaction, gustation, proprioception, and interoception, potentially through images, audio, video, and simulation-based data, to enhance embodied concept representations and address issues in spatial and temporal reasoning.
- **Embodied AI:** Allowing LLMs to interact with physical or simulated environments can deepen their conceptual grounding, particularly for tasks involving spatial reasoning, object affordances, and motor actions that purely text-based methods fail to capture.
- **Developmental Modeling:** Gradual exposure to increasingly complex concepts, mirroring human developmental stages, may strengthen models' ability to replicate human-like conceptual growth and acquisition patterns.
- **Improved Evaluation Metrics:** While BBSR proves invaluable, additional benchmarks that encompass the full spectrum of human cognition are needed. Combining quantitative and qualitative approaches, including behavioral data and neuroscientific insights, would better reflect real-world complexity.

By tackling these challenges, we can advance LLMs that not only excel at linguistic tasks but also approximate the deeper, multimodal intricacies of human cognition, with significant implications for human-computer interaction, robotics, and artificial general intelligence.

4 Methods

This study investigates the alignment between LLMs and human cognition by examining how these models represent semantic dimensions associated with concepts, as defined by the BBSR dataset proposed by Binder et al. [23]. By administering the same semantic judgment tasks to both humans (as represented in the BBSR dataset) and a selection of LLMs, we aim to quantify the degree to which these models capture human-like conceptual understanding. This comparative approach is crucial for bridging the gap between computational models of language and human cognition, ultimately contributing to the development of LLMs that exhibit not only linguistic proficiency but also cognitive and semantic coherence with human experience. This section details the methodological foundations, prompt design, evaluation domains and dimensions, testing procedure, and data processing employed in this study.

4.1 Theoretical Foundations and Framework

The BBSR dataset [23] provides a robust and empirically validated foundation for our investigation. Building upon decades of research in cognitive psychology and neuroscience, Binder et al. developed a comprehensive model that maps semantic content to underlying neural processes. This framework identifies 14 distinct semantic domains and 65 experiential dimensions, grounded in neurobiological evidence, that collectively capture the multifaceted nature of human conceptual understanding. These dimensions span a wide range of human experience, encompassing sensory (e.g., vision, audition), motor, affective, spatial, temporal, causal, and social aspects.

A key strength of the BBSR framework lies in its grounding in neurobiological plausibility. By incorporating sensory modalities, the model accounts for the embodied nature of conceptual representation, aligning with theories of embodied cognition. Furthermore, the inclusion of both concrete and abstract concepts ensures a broad coverage of human semantic experience. The rigorous methodology employed by Binder et al., including normative data collection, clustering analyses, and validation through cross-domain comparisons, establishes the BBSR dataset as a robust benchmark for studying conceptual representation.

Adopting the BBSR framework for evaluating LLMs serves two critical purposes. First, it enables a direct comparison between human and artificial semantic representations within a well-defined and empirically validated structure. Second, by grounding the evaluation in a model that reflects the functional organization of the human brain, we establish a biologically plausible basis for understanding the strengths and limitations of LLMs in replicating human cognition. This approach allows us to move beyond superficial comparisons of linguistic performance and delve into the deeper question of cognitive alignment.

4.2 Prompt Engineering and Structure

To ensure methodological consistency with the BBSR framework, we meticulously designed prompts that mirrored the structure used by Binder et al. in their normative data collection. All queries followed the general format: {*head*} {*relation*} {*content*}, where {*head*} is a concept-specific lead-in phrase, {*relation*} indicates the type of relationship being assessed (i.e., the degree to which a concept is associated with a given dimension), and {*content*} specifies the semantic dimension under consideration. The {*head*} phrases were adapted based on the concept's part of speech:

- **Nouns:** *"To what degree do you think of this thing as..."*
- **Verbs:** *"To what degree do you think of this as..."*
- **Adjectives:** *"To what degree do you think of this property as..."*

This structure directly parallels the approach used in the BBSR study, facilitating a direct comparison of results. To streamline responses and enable quantitative analysis, we appended the instruction *"Just provide the number, without any other output."* to each prompt. This instruction effectively constrained the LLMs to return only numerical ratings, simplifying data processing and minimizing output variability.

4.3 Evaluation Domains and Dimensions

Our evaluation encompassed the 14 semantic domains and 65 experiential dimensions defined by Binder et al. These dimensions were selected for their demonstrated neurobiological relevance, reflecting distinct neural processors or cognitive systems identified through neuroimaging and behavioral studies. For instance, the *Vision* domain includes dimensions such as *Brightness*, *Color*, and *Shape*, corresponding to distinct processing pathways in the visual cortex. Similarly, the *Emotion* domain incorporates affective responses like *Happiness*, *Fear*, and *Anger*, which are associated with activity in the limbic system. This comprehensive set of dimensions allows for a fine-grained analysis of semantic representation across diverse aspects of human experience.

4.4 Experimental Protocol

Nine LLMs were evaluated in this study: **Qwen2-7B**, **Qwen2-72B**, **Llama-3-8B**, **Llama-3-70B**, **Llama-3.1-8B**, **Llama-3.1-70B**, **Llama-3.1-405B**, **GPT-3.5 Turbo**, and **GPT-4o mini**. Each model was tested independently to assess its ability to rate the salience of each semantic dimension for a given concept. The experimental protocol consisted of the following steps:

1. **Isolated Prompting:** Each query was presented as a single, self-contained prompt, with the chat history cleared before each test to eliminate potential context effects and ensure independent responses.

2. **Response Replication:** To mitigate the inherent stochasticity of LLM outputs and enhance the reliability of our measurements, each concept was evaluated five times independently per LLM. The mean of these five trials was then used as the final representation vector for the concept.

3. **Output Normalization:** Responses were expected to fall within the [0, 6] range, consistent with the scale used by Binder et al[23]. Responses falling outside this range were capped at the respective boundary (0 or 6). This normalization step ensured comparability with the human data.

4. **Handling Non-Numerical Responses:** In rare instances (<0.1%), LLMs produced non-numerical or incoherent outputs. In such cases, the query was repeated. If multiple attempts failed to yield a valid numerical response, a score of 0 was assigned for that dimension. These instances were predominantly observed in smaller models.

5. **Composite Dimension Handling:** For composite dimensions such as *Temperature*, *Texture*, and *Weight*, which encompass multiple sub-dimensions (e.g., *Hot* and *Cold* for *Temperature*), scores were determined by taking the maximum rating across the relevant sub-dimensions. This approach maintained consistency with the human testing methodology employed by Binder et al[23].

4.5 Data Aggregation and Analysis

Following the testing procedure, the LLMs' responses were aggregated across all dimensions for each concept, generating high-dimensional semantic representation vectors. These vectors were then used to compare the alignment between LLMs and human normative data from the BBSR dataset. Subsequent analyses focused on evaluating the internal consistency and variability of the models' semantic judgments, providing a comprehensive assessment of their capacity to emulate human-like conceptual understanding.

Acknowledgements. This work was supported in part by the Beijing Major Science and Technology Project under Contract No.Z241100001324005.

Declarations

The authors declare that they have no competing interests.

Data Availability and Code Availability

The datasets used in this study, including the BBSR dataset and corresponding task-specific query files, are publicly available on GitHub: https://github.com/wywgreat/CognitiveAlignment. The repository also contains processed results and intermediate outputs for reproducibility.

Inclusion & Ethics Statement

This research was conducted with a commitment to inclusion, diversity, and ethical practices. The study emphasizes the alignment between LLMs and human cognition across multimodal domains.

Ethical considerations were paramount in designing and executing the study. All analyses adhered to rigorous ethical guidelines for computational and cognitive research. The dataset used does not involve personally identifiable information or data derived from vulnerable populations. Instead, it represents aggregated normative human ratings from previously established and publicly accessible datasets. Additionally, this study seeks to address global inclusivity in artificial intelligence research by evaluating LLMs across a wide range of cognitive domains. By exploring the alignment between artificial and human cognition, this work contributes to creating AI systems that are more equitable and representative of diverse human experiences.

References

[1] Paivio, A. *Mental Representations: A dual coding approach* (Oxford University Press, North Kettering Business Park Hipwell Road Kettering Northants NN14 1UA, 1990).

[2] Barsalou, L. W., Santos, A., Simmons, W. K. & Wilson, C. D. Language and simulation in conceptual processing. *Symbols, embodiment, and meaning* 245–283 (2008).

[3] Fernandino, L., Humphries, C. J., Conant, L. L., Seidenberg, M. S. & Binder, J. R. Heteromodal cortical areas encode sensory-motor features of word meaning. *Journal of Neuroscience* **36**, 9763–9769 (2016).

[4] Xu, Y., Yong, H. & Bi, Y. A tri-network model of human semantic processing. *Frontiers in Psychology* **8**, 1538 (2017).

[5] Bi, Y. Dual coding of knowledge in the human brain. *Trends in Cognitive Sciences* **25**, 883–895 (2021).

[6] Wang, X., Men, W., Gao, J., Caramazza, A. & Bi, Y. Two forms of knowledge representations in the human brain. *Neuron* **107**, 383–393 (2020).

[7] Binder, J. R., Desai, R. H., Graves, W. W. & Conant, L. L. Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. *Cerebral Cortex* **19**, 2767–2796 (2009).

[8] Handjaras, G. *et al.* How concepts are encoded in the human brain: A modality independent, category-based cortical organization of semantic knowledge. *NeuroImage* **135**, 232–242 (2016).

[9] Cappe, C., Rouiller, E. M. & Barone, P. Multisensory anatomical pathways. *Hear Res* **258**, 28–36 (2009).

[10] Brown, T. *et al.* Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020).

[11] Achiam, J. *et al.* Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[12] Chowdhery, A. *et al.* Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* **24**, 1–113 (2023).

[13] Du, Z. *et al.* Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360* (2021).

[14] Dubey, A. *et al.* The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[15] Yang, A. *et al.* Qwen2 technical report. *arXiv preprint arXiv:2407.10671* (2024).

[16] Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[17] McRae, K., Cree, G. S., Seidenberg, M. S. & McNorgan, C. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods* **37**, 547–559 (2005).

[18] Devereux, B. J., Tyler, L. K., Geertzen, J. & Randall, B. The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods* **46**, 1119–1127 (2014).

[19] Lynott, D., Connell, L., Brysbaert, M., Brand, J. & Carney, J. The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods* 1–21 (2019).

[20] Stadthagen-Gonzalez, H. & Davis, C. J. The bristol norms for age of acquisition, imageability, and familiarity. *Behavior research methods* **38**, 598–605 (2006).

[21] Lynott, D. & Connell, L. Modality exclusivity norms for 423 object properties. *Behavior research methods* **41**, 558–564 (2009).

[22] Lynott, D. & Connell, L. Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. *Behavior research methods* **45**, 516–526 (2013).

[23] Binder, J. R. *et al.* Toward a brain-based componential semantic representation. *Cognitive neuropsychology* **33**, 130–174 (2016).

[24] Brysbaert, M., Warriner, A. B. & Kuperman, V. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods* **46**, 904–911 (2014).

[25] Kuperman, V., Stadthagen-Gonzalez, H. & Brysbaert, M. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods* **44**, 978–990 (2012).