From Tokens to Thoughts: How LLMs and Humans Trade Compression for Meaning

Chen Shani
Stanford University
cshani@stanford.edu

Liron Soffer
Tel Aviv University
lironso@mail.tau.ac.il

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Yann LeCun
New York University; Meta - FAIR
yann.lecun@nyu.edu

Ravid Shwartz-Ziv
New York University; Wand.AI
rs8020@nyu.edu

Abstract

Humans organize knowledge into compact categories through *semantic compression* by mapping diverse instances to abstract representations while preserving meaning (e.g., *robin* and *blue jay* are both *birds*; most birds *can fly*). These concepts reflect a trade-off between expressive fidelity and representational simplicity. Large Language Models (LLMs) demonstrate remarkable linguistic abilities, yet whether their internal representations strike a human-like trade-off between compression and semantic fidelity is unclear. We introduce a novel information-theoretic framework, drawing from Rate-Distortion Theory and the Information Bottleneck principle, to quantitatively compare these strategies. Analyzing token embeddings from a diverse suite of LLMs against seminal human categorization benchmarks, we uncover key divergences. While LLMs form broad conceptual categories that align with human judgment, they struggle to capture the fine-grained semantic distinctions crucial for human understanding. More fundamentally, LLMs demonstrate a strong bias towards aggressive statistical compression, whereas human conceptual systems appear to prioritize adaptive nuance and contextual richness, even if this results in lower compressional efficiency by our measures. These findings illuminate critical differences between current AI and human cognitive architectures, guiding pathways toward LLMs with more human-aligned conceptual representations.

1 Introduction: The Enigma of Meaning in Large Language Models

> "The categories defined by constructions in human languages may vary from one language to the next, but they are *mapped onto a common conceptual space*, which represents a common cognitive heritage, indeed the geography of the human mind."
> *–Croft [2001] p. 139*

The human capacity for concept formation is a cornerstone of intelligence, enabling us to manage information overload by deriving meaning from complex signals. We achieve this by identifying essential features and compressing experiences into cognitively tractable summaries [Murphy, 2004]. This conceptual architecture, often hierarchical (e.g., a *robin* is a *bird*, an *animal* [Rosch et al., 1976]), is a powerful semantic compression: diverse instances are mapped to compact representations.

Preprint. Under review.

Crucially, this process balances representational efficiency (compression) with the preservation of essential semantic fidelity (meaning), a trade-off fundamental to learning and understanding.

Large Language Models (LLMs) exhibit striking capabilities in processing and generating human language, performing tasks that often appear to require deep semantic understanding [Singh et al., 2024, Li et al., 2024]. Despite this, a fundamental enigma persists: **Do LLMs truly grasp concepts and meaning analogously to humans, or is their success primarily rooted in sophisticated statistical pattern matching over vast datasets?** This question is particularly salient given the human ability to effortlessly distill extensive input into compact, meaningful concepts, a process governed by the inherent trade-off between informational compression and semantic fidelity [Tversky, 1977, Rosch, 1973b].

As the mental scaffolding of human cognition, concepts enable efficient interpretation, generalization from sparse data, and rich communication. For LLMs to transcend surface-level mimicry and achieve more human-like understanding, it is critical to investigate how their internal representations navigate the crucial trade-off between information *compression* and the *preservation of semantic meaning*. Do LLMs develop conceptual structures mirroring the efficiency and richness of human thought, or do they employ fundamentally different representational strategies?

To address this, we introduce a novel quantitative methodology rooted in information theory. We develop and apply a framework drawing from **Rate-Distortion Theory** [Shannon, 1948] and the **Information Bottleneck** principle [Tishby et al., 2000] to systematically compare how LLMs and human conceptual structures balance representational complexity (compression) with semantic fidelity. As a crucial human baseline, we leverage seminal datasets from cognitive psychology detailing human categorization [Rosch, 1973a, 1975, McCloskey and Glucksberg, 1978]. A contribution of this work is the digitization and public release of these classic datasets, which offer benchmarks of high empirical rigor often exceeding modern crowdsourced alternatives. Our framework is tailored to dissect how these different systems navigate the compression-meaning trade-off.

Our comparative analysis across a diverse suite of LLMs reveals divergent representational strategies. While LLMs generally form broad conceptual categories aligned with human judgment, they often fail to capture the fine-grained semantic distinctions pivotal to human understanding. More critically, we uncover a stark contrast in priorities: LLMs exhibit a strong drive towards aggressive statistical compression, whereas human conceptual systems appear to favor adaptive nuance and contextual richness, even at a potential cost to sheer compressional efficiency by our measures. This divergence underscores fundamental differences and informs pathways for developing AI with more human-aligned conceptual understanding.

2 Research Questions and Scope

Advancing AI beyond pattern matching towards deeper semantic understanding hinges on whether LLMs develop conceptual structures analogous to human cognition. Human concepts efficiently balance semantic richness with cognitive manageability, a trade-off between meaning and informational compression. This paper investigates if and how LLMs replicate this fundamental balance.

Prior work has explored the conceptual landscape of LLMs, including their grasp of relational knowledge [Shani et al., 2023, Misra et al., 2021], methods for extracting interpretable concepts [Hoang-Xuan et al., 2024, Maeda et al., 2024], emergent representations via sparse activations [Li et al., 2024], embedding geometry concerning hierarchies [Park et al., 2024], and autoregressive concept prediction [Barrault et al., 2024]. While insightful, these studies often lack a deep, quantitative comparison of the *compression-meaning trade-off* using an information-theoretic lens benchmarked against rich human cognitive data, or they may not ground concept definitions in established cognitive theory. Consequently, a rigorous comparative evaluation of how LLMs and humans balance representational efficiency with semantic fidelity remains a key open area. Separately, cognitive science has applied information theory to human concept learning [Imel and Zaslavsky, 2024, Tucker et al., 2025, Wolff, 2019, Sorscher et al., 2022], yet typically without connecting to modern AI models.

This work aims to bridge this gap by integrating cognitive psychology, information theory, and modern NLP. We pose three central research questions to guide our investigation:

[RQ1]: To what extent do concepts emergent in LLMs align with human-defined conceptual categories?

[RQ2]: Do LLMs and humans exhibit similar internal geometric structures within these concepts, especially concerning item typicality?

[RQ3]: How do humans and LLMs differ in their strategies for balancing representational compression with the preservation of semantic fidelity when forming concepts?

These three questions steer our investigation, which approaches each through the unifying lens of the information-theoretic framework detailed in Section 4. RQ1 begins by examining the alignment of broad conceptual categories, a key aspect of how information is compressed. RQ2 then delves into the finer-grained internal structures of these categories, probing the preservation of semantic nuances such as item typicality. Building on these analyses, RQ3 employs the full framework to comprehensively compare how LLMs and humans may divergently optimize the overall trade-off between compression and meaning. To ground these comparisons, we consistently utilize seminal human categorization datasets [Rosch, 1973a, 1975, McCloskey and Glucksberg, 1978] as empirical benchmarks. Our overarching aim is to use this comparative, information-theoretic approach not only to evaluate current LLMs but also to advance our understanding of efficient and meaningful representation in both artificial and natural intelligence.

3 Benchmarking Against Human Cognition

Empirically investigating the relationship between LLM representations and human conceptual structures requires two critical components: robust benchmarks of human categorization and a diverse selection of LLMs. This section details these components.

3.1 Human Conceptual Baselines: Empirical Data from Seminal Cognitive Science

Our comparison is anchored by data from seminal studies in cognitive psychology that mapped human categorization processes. These studies offer rich empirical evidence of how humans form concepts, judge category membership, and perceive typicality. Critically, unlike many modern crowdsourced datasets which can be noisy, these classic benchmarks were meticulously curated by cognitive science experts, reflecting deep cognitive patterns rather than superficial associations, and were grounded in then-advancing theories of conceptual structure. We focus on three influential works:

Rosch (1973): This foundational work by Rosch [1973a] explored semantic categories as part of the research program leading to prototype theory [Rosch, 1973c]. This theory posits that categories are organized around "prototypical" members rather than strict, equally shared features. The dataset includes 48 items in eight common semantic categories (e.g., furniture, bird), with prototypicality rankings (e.g., 'robin' as a typical bird, 'bat' as atypical).

Rosch (1975): Building on prototype theory, Rosch [1975] further detailed how semantic categories are cognitively represented. This work provides extensive typicality ratings for a larger set of 552 items across ten categories (e.g., 'orange' as a prototypical fruit, 'squash' as less so).

McCloskey & Glucksberg (1978): McCloskey and Glucksberg [1978] investigated the "fuzzy" boundaries of natural categories, showing that membership is often graded rather than absolute. Their data covers 449 items in 18 categories, with typicality scores and membership certainty ratings (e.g., 'dress' is typical clothing, 'bandaid' less so).

While originating from different research groups with distinct theoretical emphases, these datasets share rigorous experimental designs and provide data on both category assignments and item typicality. We aggregated data from these studies, creating a unified benchmark of 1,049 items across 34 categories. This aggregated dataset, which we have digitized and make publicly available (see Appendix A.1), offers a crucial, high-fidelity empirical foundation for evaluating the human-likeness of computational models and we encourage its use for future research.

Appendix A.6 complements this benchmark by reporting WordNet-based polysemy statistics for every item. These figures characterize the lexical ambiguity of the stimuli and frame the subsequent

embedding analyses. The analysis shows that polysemy is very rare in our data and thus cannot explain the findings we report later in the paper.

3.2 Large Language Models Under Study

We include a diverse array of LLMs to assess how conceptual representation might vary with computational architecture and scale. This selection covers prevalent architectural paradigms (encoder-only, decoder-only) and a wide spectrum of model sizes, from 300 million to 72 billion parameters.

Our analysis features encoder-only models from the BERT family (e.g., BERT-Large [Devlin et al., 2019, He et al., 2020, Zhuang et al., 2021]). The majority are decoder-only autoregressive models, including: six Llama family models (1B to 70B, e.g., Llama 3.1 70B [Touvron et al., 2023a,b, Grattafiori et al., 2024]); five Gemma family models (2B to 27B [Team et al., 2024, 2025]); thirteen Qwen family models (0.5B to 72B [Bai et al., 2023, Yang et al., 2024]); four Phi family models (e.g., Phi-4 [Javaheripi et al., 2023, Abdin et al., 2024, Abouelenin et al., 2025]); a Mistral 7B model [Karamcheti et al., 2021]; two GPT-2 models of different scales [Radford et al., 2019]; and checkpoints from the OLMo-7B model [Groeneveld et al., 2024]. Other, less common architectures include a single encoder-decoder T5 model [Raffel et al., 2020], two text encoders from CLIP ViT (ViT-B/32 and ViT-B/16) [Radford et al., 2021], and the classic encoder models GloVe [Pennington et al., 2014] and Word2Vec [Mikolov et al., 2013a,b]. Note: GloVe and Word2Vec provide static embeddings only, and thus their analysis is constrained. Appendix A.2 provides a comprehensive list of all model variants, identifiers, and architectural details.

For each LLM, we extract static, token-level embeddings from its input embedding layer (the ‘E‘ matrix). This choice aligns our analysis with the context-free nature of stimuli typical in human categorization experiments, ensuring a comparable representational basis. These embeddings form the foundation for deriving LLM-generated conceptual clusters in our subsequent analyses.

To complement the analysis, we extend our methodology beyond static embeddings to contextual representations. For each LLM, we extract token-level embeddings using a set of carefully controlled prompts (see Appendix A.3), enabling a principled basis for directly comparing static token representations to their contextualized counterparts. This allows for a similar assessment while mitigating confounding issues such as prompt length and token position. We found minimal variation in results across prompt templates and pooling methods. Appendix A.3 provides a detailed exploration.

Finally, tokenizer-overlap analysis (Appendix A.7) shows substantial lexical sharing across model families; neither token count nor tokenization overlap correlates with AMI, confirming that tokenizer effects are negligible and thus cannot explain our findings.

4 A Framework for Comparing Compression and Meaning

To understand how LLMs and human cognition grapple with the fundamental challenge of representing meaning, we introduce an information-theoretic framework. This framework is designed to **analyze the critical trade-off, or tension, between compressing information into efficient representations and preserving the rich semantic fidelity essential for true understanding**. Drawing upon core principles from *Rate-Distortion Theory (RDT)* [Shannon, 1948] and the *Information Bottleneck (IB) principle* [Tishby et al., 2000], our approach provides a cohesive lens for addressing all three of our research questions. Our investigation progresses by first exploring distinct facets of this trade-off related to representational compactness and semantic preservation, before synthesizing these insights to evaluate the overall efficiency of conceptual representation. Our research questions, viewed through this progressive information-theoretic perspective, are approached as follows:

> **[RQ1] Probing Representational Compactness via Categorical Alignment:** We begin by examining how information is condensed into categorical structures. Both human categorization and LLM-derived clustering simplify diverse items X into structured groups C. For RQ1, we assess alignment between model-based clusters (C_{LLM}) and human categories (C_{Human}) by quantifying shared information (e.g., via Adjusted Mutual Information), offering an initial view on how similarly compactness is achieved. The principles of efficient input representation here relate to the “Complexity” aspect of our framework.

[RQ2] Probing Semantic Preservation via Internal Structure: Next, we assess how well meaning is preserved *within* these compressed representations. An effective system must retain crucial semantic nuances. For RQ2, we investigate this by correlating LLM-internal measures of item centrality with human typicality judgments, probing how faithfully fine-grained semantic information is represented, that is, can LLMs capture the internal structure of C_{Human}? This relates to the "Distortion" (or fidelity) aspect of our framework.

[RQ3] Evaluating the Integrated Trade-off for Total Representational Efficiency: Finally, having explored compactness and preservation, we leverage our full framework. RQ3 employs a unified objective function, $\mathcal{L}$ (detailed below), to quantitatively assess the total efficiency with which LLMs and human systems navigate this fundamental trade-off.

The following subsections detail the theoretical underpinnings of this framework.

4.1 Theoretical Underpinnings: Rate-Distortion Theory and the Information Bottleneck

To rigorously formalize the balance between representational compactness and preserved meaning, we draw upon information theory. **Rate-Distortion Theory (RDT)** [Shannon, 1948] provides the foundational language. RDT quantifies the minimal "rate" R (representational complexity) needed to represent a source X as C, subject to a maximum "distortion" D (fidelity loss). The goal is often to optimize $R + \lambda D$, offering a principled evaluation of representational efficiency.

The **Information Bottleneck (IB) principle** [Tishby et al., 2000] is a related approach. IB seeks a compressed representation C of an input X that maximizes information about a relevant variable Y while minimizing $I(X;C)$, the mutual information C retains about X (the bottleneck's "cost"). This is typically framed as minimizing $I(X;C) - \beta I(C;Y)$.

Our analytical framework directly applies RDT's core idea of balancing rate and distortion. We formulate an objective function, $\mathcal{L}$, designed to explicitly balance a **complexity term** (analogous to RDT's rate), which quantifies the informational cost of representing items X through their conceptual clusters C, and a **distortion term** (analogous to RDT's D), which measures semantic information lost or obscured within these clusters. Our complexity term, incorporating $I(X;C)$, resonates with the IB principle. However, our distortion term directly measures intra-cluster semantic fidelity loss (specifically, the variance of item embeddings relative to their cluster centroids), differing from canonical IB formulations where distortion is often implicitly tied to an external relevance variable Y. This direct approach allows us to evaluate how any given clustering C, whether derived from human cognitive data or LLM embeddings, intrinsically balances its own structural compactness and the meaningfulness of its components with respect to the original data X.

4.2 The $\mathcal{L}$ Objective: Balancing Representational Complexity and Semantic Distortion

Building on these information-theoretic foundations, this section formally defines the two key components of our framework–*Complexity* and *Distortion*. These components allow us to quantitatively address the aspects of representational compactness (core to [RQ1]) and semantic preservation (central to [RQ2]) that were introduced earlier. We then combine these into a unified objective function, $\mathcal{L}$, designed to evaluate the overall efficiency of the compression-meaning trade-off, which is the primary focus of [RQ3]. The $\mathcal{L}$ function evaluates the efficiency of the conceptual clusters C derived from items X (e.g., token embeddings):

$$\mathcal{L}(X, C; \beta) = \text{Complexity}(X, C) + \beta \cdot \text{Distortion}(X, C). \tag{1}$$

Here, $\beta \geq 0$ is a hyperparameter that balances the relative importance of the two terms.

The Complexity (Rate) Term: The first component, **Complexity**(X, C), measures the informational cost or intricacy of representing the original items X through their assignments to clusters C. It is quantified by the mutual information $I(X;C)$ between the items and their cluster labels. A lower $I(X;C)$ signifies greater compression, meaning the cluster assignments C make the specific items X more predictable (i.e., require less information to specify beyond the cluster label). Defining $I(X;C) = H(X) - H(X|C)$, and assuming $|X|$ equiprobable unique items for the initial entropy calculation ($H(X) = \log_2 |X|$), the conditional entropy is $H(X|C) = \frac{1}{|X|} \sum_{c \in C} |C_c| \log_2 |C_c|$. This assumes that for this complexity calculation, items within each cluster C_c (of size $|C_c|$) are

indistinguishable beyond their shared label c. Thus:

$$\text{Complexity}(X, C) = \log_2 |X| - \frac{1}{|X|} \sum_{c \in C} |C_c| \log_2 |C_c|. \tag{2}$$

This term formalizes the representational compactness aspect central to [RQ1].

The Distortion Term: The second component, **Distortion**(X, C), quantifies the loss of semantic fidelity incurred by grouping items into clusters. It is measured as the average intra-cluster variance of the item embeddings, reflecting how tightly items are bound to their cluster's central tendency and thus the cluster's semantic coherence. This directly relates to the preservation of fine-grained semantic information, an idea explored in [RQ2]. For each cluster $c \in C$, its centroid is $x_c = \frac{1}{|C_c|} \sum_{x \in c} x$ (the mean embedding of its items). Its internal variance is $\sigma_c^2 = \frac{1}{|C_c|} \sum_{x \in c} \|x - x_c\|^2$. The total distortion for the clustering C is the weighted average of these variances:

$$\text{Distortion}(X, C) = \frac{1}{|X|} \sum_{c \in C} |C_c| \cdot \sigma_c^2. \tag{3}$$

A lower distortion value implies that, on average, items are close to their respective cluster centroids, suggesting better preservation of shared semantic features within each cluster.

The Unified Objective Function: Substituting the formal definitions of Complexity (Equation 2) and Distortion (Equation 3) into our general formulation for $\mathcal{L}$ (Equation 1) yields the complete objective function that underpins our comparative analysis:

$$\mathcal{L}(X, C; \beta) = \left(\log_2 |X| - \frac{1}{|X|} \sum_{c \in C} |C_c| \log_2 |C_c| \right) + \beta \cdot \left(\frac{1}{|X|} \sum_{c \in C} |C_c| \cdot \sigma_c^2 \right). \tag{4}$$

This $\mathcal{L}$ function provides a single, principled measure for evaluating how effectively a given clustering C balances the need for informational compression against the imperative to preserve semantic meaning, serving as the direct quantitative tool for addressing [RQ3].

With the $\mathcal{L}$ objective now fully specified, our information-theoretic framework provides a comprehensive toolkit. The Complexity term (Equation 2) allows us to quantify aspects of representational compactness pertinent to [RQ1], while the Distortion term (Equation 3) enables the assessment of semantic preservation, crucial for [RQ2]. The overall $\mathcal{L}$ function (Equation 4) then directly facilitates the evaluation of the integrated compression-meaning trade-off, central to [RQ3]. Thus, this framework equips us to systematically and quantitatively investigate how LLMs and human cognition manage the balance between informational efficiency and semantic richness. We apply this framework in our empirical investigation detailed in Section 5.

5 Unpacking Representational Strategies: An Empirical Investigation

Building on our information-theoretic framework (Section 4) and established benchmarks (Section 3), we now empirically investigate our research questions. This section details the specific methodologies employed to compare LLM and human conceptual strategies across the key dimensions of conceptual alignment, internal semantic structure, and overall representational efficiency.

[RQ1] Assessing Conceptual Alignment To investigate how LLM-derived conceptual categories align with human-defined ones (RQ1), probing representational compactness, we cluster LLM token embeddings using k-means (K set by human category counts per dataset). Alignment with human categories is quantified using Adjusted Mutual Information (AMI), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI), against a random clustering baseline.

[RQ2] Examining Internal Cluster Geometry and Semantic Preservation To assess how LLM representations capture human-like typicality (RQ2), examining internal category geometry, we calculate the cosine similarity of each item's token embedding to the token embedding of its human-assigned category name (e.g., 'robin' to 'bird'). These LLM-derived similarities are then correlated (Spearman's ρ) with human typicality ratings from our cognitive science datasets.

[RQ3] Evaluating the Efficiency of the Compression-Meaning Trade-off To evaluate the overall balance of compression and meaning (RQ3), we apply our framework by computing the $\mathcal{L}$ objective (Equation 4, $\beta = 1$) for both human and LLM-derived conceptual structures (the latter from k-means over a range of K). This compares how each system balances Complexity $I(X;C)$ against Distortion. Cluster entropy is an ancillary measure of compactness.

For robustness, all k-means clustering involves one hundred random initializations with averaged results. Appendix A.8 provides details on supplementary metrics like Silhouette scores.

5.1 [RQ1] The Big Picture: Alignment of Conceptual Categories

We first investigate whether LLMs form conceptual categories aligned with human judgment.

Key Finding: Broad Alignment with Human Categories

LLM-derived clusters significantly align with human-defined conceptual categories, suggesting they capture key aspects of human conceptual organization. Notably, certain encoder models exhibit surprisingly strong alignment, sometimes outperforming much larger models, highlighting that factors beyond sheer scale influence human-like categorical abstraction.

Experimental Recap: LLM token embeddings from our benchmark datasets [Rosch, 1973a, 1975, McCloskey and Glucksberg, 1978] were clustered (k-means; K matching human category counts). Alignment with human categories was measured using AMI, NMI, and ARI (AMI shown in Figure 1; see Appendices A.8, A.9 for full details). Both the static and the contextual embeddings were extracted. The contextual embeddings were measured layer-wise, and reported using Peak AMI across the layers. Note that GloVe and Word2Vec only have static embeddings.

Results and Observations: Across all tested LLMs, derived conceptual clusters aligned with human categories significantly above random chance (Figure 1, showing averaged peak AMI scores). This indicates their semantic spaces encode information supporting human-like grouping at a macro level. Notably, the BERT family (especially BERT-large-uncased) demonstrated robust alignment, often comparable to or exceeding that of much larger decoder-only models. This suggests that architectural or pre-training factors, not just scale, influence the formation of human-like categorical structures.

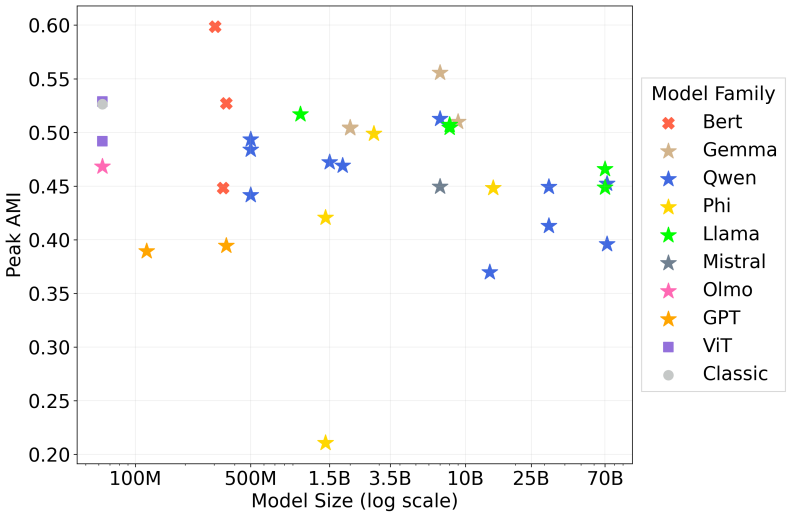


Figure 1: **LLM-derived Clusters Show Above-Chance Alignment with Human Conceptual Categories.** Adjusted Mutual Information (AMI) between human categories and LLM-embedding clusters versus model size. Results are averaged over three psychological datasets. All models perform significantly better than random clustering. Encoder architectures achieve consistently high performance across modalities and scales; notably, the classic Word2Vec and GloVe models, despite offering only static, outdated embeddings, reach static AMI scores that match the peak AMI of modern LLMs. See Appendix A.4 for a comparison of static, mean, and peak AMI.

Interpretation: These findings confirm that LLMs can recover broad, human-like categories from their embeddings, validating deeper comparative analyses. This macro-level agreement necessitates examining the finer-grained internal geometry of these categories, which we address next.

Influence of Training Dynamics on Alignment: To explore how semantic structure develops during training, we applied both the static embedding and the contextual (layer-wise) analysis on 10% of OLMo's checkpoints (spaced evenly across training steps). We find that concept representations improve steadily during training, with a rapid gain in structure early on, followed by a slower, more gradual refinement. Interestingly, the model exhibits a two-phase dynamic: initial optimization moves semantic processing from deep layers to mid-network layers during rapid learning, followed by continued architectural reorganization that optimizes efficiency while maintaining performance (Appendix A.5 Figure 8). This suggests that concepts emerge early in training but are progressively consolidated and reallocated across layers, reflecting a shift from brute-force memorization toward more efficient, distributed representations. Such dynamics highlight that alignment is not just a final-state property of the trained model, but is deeply shaped by the trajectory of learning itself.

5.2 [RQ2] Delving Deeper: Fidelity to Fine-Grained Semantics

Having established that LLMs broadly align with human conceptual categories (Section 5.1), we next investigate a more nuanced question: Do LLMs also capture the internal semantic structure of these categories, particularly human-like item typicality?

Key Finding: Limited Capture of Semantic Nuance

While LLMs effectively form broad conceptual categories, **their internal representations demonstrate only modest alignment with human-perceived fine-grained semantic distinctions**, such as item typicality or psychological distance to category prototypes. This suggests a divergence in how LLMs and humans structure information within concepts.

Experimental Recap: For this RQ, as detailed in this section's introduction, we compared human typicality judgments from the cognitive science datasets [Rosch, 1973a, 1975, McCloskey and Glucksberg, 1978] with an LLM-based measure. Specifically, we calculated the cosine similarity between each item's token embedding and the token embedding of its **human-assigned category name** (e.g., 'robin' vs. 'bird'). These item-to-category-label similarities were then correlated (Spearman's ρ [Wissler, 1905]) with human-rated typicality scores.

We employed two analysis approaches: static-layer analysis utilized static items/category embeddings from the embedding layer, while peak AMI layer analysis utilized items/category contextual embeddings from the layer that maximized AMI as measured in RQ1. To extract category embeddings in the peak AMI approach, we replaced the item with the category name in the same prompt template (see template and template- and pooling robustness analysis in Appendix A.4).

Results and Observations: Spearman correlations between LLM-derived item-to-category-label similarities and human typicality judgments are generally modest across most models and datasets, with peak AMI layers showing particularly weak correlations (Tables 3, 4 in Appendix A.10; Figure 15). Even when correlations reach statistical significance ($p < 0.05$), their magnitudes typically indicate a limited correspondence.

Several architectural patterns emerge: representation-focused models (Word2Vec, GloVe) and some encoder models (T5-large, the two ViT encoders, and BERT large) demonstrate stronger static-layer performance than decoder-only models (Llama, Gemma, etc.), suggesting that models trained specifically for representation learning may be more effective for capturing semantic category relationships in their embeddings. In contrast, modern decoder-only models show consistently lower static-layer correlations, potentially due to their training objectives being more focused on next-token prediction rather than semantic representation.

Notably, peak AMI layer analysis shows that correlations weaken substantially relative to static-layer analysis across most architectural variants, suggesting that **cross-cluster relations (RQ1) and within-cluster organization (RQ2) are encoded at different layers**. This implies that downstream applications should select layers based on the resolution they require; whether capturing broad cross-cluster structure or fine-grained within-cluster organization. Furthermore, while some architectural

approaches capture certain aspects of human typicality judgments, their internal representations exhibit only modest alignment with human-perceived fine-grained semantic distinctions, such as item typicality or psychological distance to category prototypes. As a result, no single architectural approach robustly reproduces the full spectrum of human typicality gradients using this measure. Additional visualizations supporting these findings are provided in Appendix A.11.

Interpretation: These findings suggest that while LLMs can identify features for broad categorization, their organization of semantic space around explicit category labels does not fully mirror the nuanced prototype structures evident in human typicality judgments. The factors driving an item's embedding similarity to its category label's embedding in LLMs may differ from the rich, multifaceted criteria (e.g., perceptual attributes, functional roles) underpinning human typicality. LLMs might instead capture a more statistically uniform association to category labels, thereby under-representing the graded, prototype-centric nature of human concepts. The tension between representational compactness and semantic fidelity becomes particularly evident when comparing peak AMI layers to static layers: while peak AMI layers optimize the Information Bottleneck trade-off for efficient clustering, they may prioritize compression over preserving the nuanced semantic information needed for human-like typicality judgments. This divergence in capturing fine-grained semantics leads to our subsequent inquiry into overall information processing efficiency.

5.3 [RQ3] The Efficiency Angle: The Compression-Meaning Trade-off

Having explored categorical alignment (RQ1) and internal semantic structure (RQ2), we now address our central question: How do LLM and human representational strategies compare in overall efficiency when balancing informational compression against semantic meaning preservation? Our information-theoretic framework directly probes this trade-off.

Key Finding: Divergent Efficiency Strategies

LLMs demonstrate markedly superior information-theoretic efficiency in their conceptual representations compared to human conceptual structures. Evaluated via our $\mathcal{L}$ objective, LLM-derived clusters consistently achieve a more "optimal" balance (by this measure) between representational complexity (compression) and semantic distortion. Human conceptualizations, while richer, appear less statistically compact, suggesting optimization for pressures beyond pure statistical compressibility.

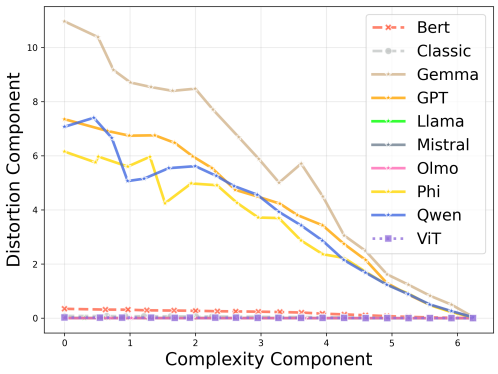
Experimental Recap: As detailed in this section's introduction, we analyzed human-defined categories and LLM-derived clusters (from k-means across various K) using two primary information-theoretic measures: mean cluster entropy (S_α) [Giraldo et al., 2014, Wei et al., 2025] and our $\mathcal{L}$ objective function (Equation 4, with $\beta = 1$).

Results and Observations: Illustrative results from one dataset (Rosch, 1975) are shown in Figure 2; trends were consistent across all datasets (full results in Appendix A.13).

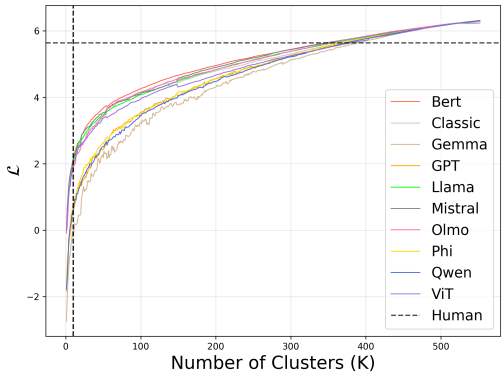
Cluster Entropy Insights: Human concepts consistently exhibit higher mean entropy than LLM-derived clusters, even at similar K values (Figure 2, left). This suggests that, by this measure, human categories are less statistically "compact" and encompass greater internal diversity than LLM clusters.

Information-Theoretic Objective ($\mathcal{L}$) Insights: The $\mathcal{L}$ objective reveals an even starker divergence (Figure 2, right). LLM-derived clusters consistently achieve significantly lower $\mathcal{L}$ values than human conceptual categories across most tested K. Since a lower $\mathcal{L}$ signifies a more statistically "optimal" trade-off between minimizing complexity and distortion within our framework, this implies LLMs are more "efficient" by this specific information-theoretic benchmark.

Interpretation: The combined results from entropy and the $\mathcal{L}$ objective strongly indicate a fundamental difference in representational strategy. LLMs appear highly optimized for statistical compactness, achieving information-theoretically "efficient" representations by minimizing redundancy and internal variance. Human conceptual systems, in contrast, while appearing "suboptimal" by these statistical measures, are likely shaped by a broader array of functional imperatives. These include the demands of adaptive generalization, rich causal and functional inference, the constraints of neural embodi-



(a) **Distortion as a function of Complexity demonstrates the balance of the Information Bottleneck in our information-theoretic objective ($\mathcal{L}$).** The plot illustrates how the IB balances compression (complexity) against information preservation (distortion). Encoder models (BERT, ViT, classic models) show lower distortion rates compared to all the decoder models across all complexity values.



(b) **Contextual embeddings achieve better-than-human $\mathcal{L}$ trade-off.** Our information-theoretic objective ($\mathcal{L}$) vs. K. Lower $\mathcal{L}$ indicates a more statistically optimal compression-meaning balance compared to human baseline.

Figure 2: **LLMs Show More Statistically "Optimal" Compression Than Humans in the $\mathcal{L}$ Measure.** (a) Distortion as a function of Complexity showing the trade-off relationship between the two terms in the IB-RDT objective ($\mathcal{L}$). Encoder models (BERT, ViT, classic models) show lower distortion rates compared to all the decoder models across all complexity values. (b) IB-RDT objective ($\mathcal{L}$) as a function of K used for k-means clustering using contextual embeddings. Lower $\mathcal{L}$ indicates a more statistically optimal compression-meaning balance. Human categories consistently show higher $\mathcal{L}$ values. Results shown for Rosch [1975] dataset; full results in Appendix A.13.

ment, and the requirements of nuanced communication–pressures that may favor representations less statistically "tidy" but ultimately more flexible and powerful for navigating a complex world.

The Objective $\mathcal{L}$ vs. Downstream Task Performance: The relationship between concept compression, as measured by the $\mathcal{L}$ objective function, and downstream task performance shows no correlation, indicating that concept compression captures an orthogonal property to benchmark accuracy. See Appendix A.14 for detailed analysis.

6 Discussion and Conclusion

Our information-theoretic investigation reveals a fundamental divergence: LLMs and humans employ starkly different strategies in balancing informational compression with semantic meaning. While LLMs achieve broad categorical alignment with human judgment (RQ1; Section 5.1), they falter in capturing fine-grained semantic nuances such as typicality (RQ2; Section 5.2) and, critically, exhibit vastly different representational efficiency profiles (RQ3; Section 5.3). This pattern strongly suggests that LLMs and humans are optimizing for different objectives.

LLMs appear aggressively optimized for statistical compactness. They form information-theoretic efficient representations, as evidenced by their lower cluster entropy and more "optimal" $\mathcal{L}$ scores. This hints they minimize redundancy and maximize statistical regularity, likely a consequence of their training on immense text corpora. This intense focus on compression, however, limits their capacity to fully encode the rich, prototype-based semantic details vital for deep, human-like understanding.

Human cognition prioritizes adaptive richness, contextual flexibility, and broad functional utility, even if this incurs a cost in statistical compactness as measured by our framework. The higher entropy and $\mathcal{L}$ scores observed for human concepts likely reflect an optimization for a wider array of complex cognitive demands. These include nuanced representations for robust generalization, supporting potent inferential capabilities (causal, functional, goal-oriented), enabling effective communication through learnable and shareable structures, and grounding concepts in rich, multimodal experiences.

The brain's neural architecture itself may inherently favor distributed, context-sensitive, and adaptable representations over statically optimal compression. Human cognition, therefore, appears to "invest" in what our statistical measures register as inefficiency for better adaptability and versatility.

The noteworthy performance of smaller encoder models like BERT in specific alignment tasks (Section 5.1) also underscores that architectural design and pre-training objectives significantly influence a model's ability to abstract human-like conceptual information. This observation highlights important avenues for future AI development focused on enhancing human-AI alignment.

These divergent representational strategies carry significant implications. **For AI development**, achieving more human-like understanding demands moving beyond current paradigms often centered on scaling and statistical pattern matching. Future efforts should explore principles that explicitly foster richer, more nuanced conceptual structures; our information-theoretic framework and $\mathcal{L}$ objective (Section 4) offer a potential class of tools for guiding and evaluating models toward this more human-like balance. **For cognitive science**, LLMs, with their distinct optimization biases, serve as valuable computational foils. Comparing their operational strategies against human performance can illuminate the unique constraints and multifaceted objectives that have shaped human concept formation, providing a powerful testbed for cognitive theories.

In essence, LLMs excel at statistical compressibility, treading a representational path fundamentally distinct from human cognition, which champions adaptive richness and functional utility, often above sheer statistical efficiency. This core difference is critical: it highlights current limitations in AI's pursuit of human-like understanding and charts vital directions for future research. Progressing AI "from tokens to thoughts", towards systems that genuinely comprehend and reason, will necessitate embracing principles that cultivate this richer, contextually-aware conceptual structure. Our framework offers a quantitative step in this direction, encouraging further exploration of how apparent "inefficiencies" might, in fact, be hallmarks of robust, human-like intelligence.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R Costa-jussà, David Dale, et al. Large concept models: Language modeling in a sentence representation space. *arXiv preprint arXiv:2412.08821*, 2024.

William Croft. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press, USA, 2001.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Luis Gonzalo Sanchez Giraldo, Murali Rao, and Jose C Principe. Measures of entropy from data using infinitely divisible kernels. *IEEE Transactions on Information Theory*, 61(1):535–548, 2014.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dirk Groeneveld, William Merrill, Luca Soldaini, Bill Yuchen Lin, Madeleine Lee, Siddharth Vashishtha, David Wadden, Hailey Schoelkopf, Oyvind Tafjord, Peter West, Nathan Lambert, Jonathan Ernst, Jeffry Kahn, Raj Khanna, Kaixin Ma, Potsawee Manakul, Maxwell Nye, Kyle Richardson, Dustin Schwenk, Zewei Shen, Zhengxuan Shen, Xinyi Song, Mark Wadden, Luke Zettlemoyer, Matthew E. Peters, and Noah A. Smith. Olmo: Accelerating the science of language models, 2024. URL https://arxiv.org/abs/2402.00838.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

Nhat Hoang-Xuan, Minh Vu, and My T Thai. Llm-assisted concept discovery: Automatically identifying and explaining neuron functions. *arXiv preprint arXiv:2406.08572*, 2024.

Nathaniel Imel and Noga Zaslavsky. Optimal compression in human concept learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 1(3):3, 2023.

Siddharth Karamcheti, Laurel Orr, Jason Bolton, Tianyi Zhang, Karan Goel, Avanika Narayan, Rishi Bommasani, Deepak Narayanan, Tatsunori Hashimoto, Dan Jurafsky, et al. Mistral–a journey towards reproducible language model training, 2021.

Yuxiao Li, Eric J Michaud, David D Baek, Joshua Engels, Xiaoqing Sun, and Max Tegmark. The geometry of concepts: Sparse autoencoder feature structure. *arXiv preprint arXiv:2410.19750*, 2024.

Akihiro Maeda, Takuma Torii, and Shohei Hidaka. Decomposing co-occurrence matrices into interpretable components as formal concepts. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4683–4700, 2024.

Michael E McCloskey and Sam Glucksberg. Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, 6(4):462–472, 1978.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at International Conference on Learning Representations (ICLR)*, 2013a. URL https://arxiv.org/abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119, 2013b.

Kanishka Misra, Allyson Ettinger, and Julia Taylor Rayz. Do language models learn typicality judgments from text? *arXiv preprint arXiv:2105.02987*, 2021.

Gregory Murphy. *The big book of concepts*. MIT press, 2004.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. *arXiv preprint arXiv:2406.01506*, 2024.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162/.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

E Rosch. On the internal structure of perceptual and semantic categories. *Cognitive development and the acquisition of language/New York: Academic Press*, 1973a.

Eleanor Rosch. Prototype theory. *Cognitive development and the acquisition of language*, pages 111–144, 1973b.

Eleanor Rosch. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192, 1975.

Eleanor Rosch, Carol Simpson, and R Scott Miller. Structural bases of typicality effects. *Journal of Experimental Psychology: Human perception and performance*, 2(4):491, 1976.

Eleanor H Rosch. Natural categories. *Cognitive psychology*, 4(3):328–350, 1973c.

Chen Shani, Jilles Vreeken, and Dafna Shahaf. Towards concept-aware large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13158–13170, 2023.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*, 2024.

Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, 119(43): e2200800119, 2022.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Mycal Tucker, Julie Shah, Roger Levy, and Noga Zaslavsky. Towards human-like emergent communication via utility, informativeness, and complexity. *Open Mind*, 9:418–451, 2025.

Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

Lan Wei, Dong Wang, and Yu Wang. Generalized relative entropy: New look at rényi entropy and its exploration from complexity measures to sparsity measures with applications in machine condition monitoring. *Mechanical Systems and Signal Processing*, 223:111917, 2025.

Clark Wissler. The spearman correlation formula. *Science*, 22(558):309–311, 1905.

J Gerard Wolff. Information compression as a unifying principle in human learning, perception, and cognition. *Complexity*, 2019(1):1879746, 2019.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized BERT pre-training approach with post-training. In Sheng Li, Maosong Sun, Yang Liu, Hua Wu, Kang Liu, Wanxiang Che, Shizhu He, and Gaoqi Rao, editors, *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China, August 2021. Chinese Information Processing Society of China. URL `https://aclanthology.org/2021.ccl-1.108/`.

A Limitations

While this study offers valuable insights, several limitations should be considered.

- Our analysis primarily focuses on English; generalizability across languages with different structures is an open question.
- Human categorization data as a benchmark may not fully capture cognitive complexity and could introduce biases.
- Our IB-RDT objective is applied to specific LLMs; other models or representations might behave differently.
- Our analysis is limited to textual input and does not explore image-based representations.

Future work could address these by expanding to other languages, exploring alternative cognitive models, and testing these principles on different architectures or in real-world applications.

A.1 Dataset Access Details

The aggregated and digitized human categorization datasets from Rosch [1973a, 1975], McCloskey and Glucksberg [1978] are made available in CSV format at: `https://huggingface.co/datasets/CShani/human-concepts`.

A.2 LLM Details

- **BERT family:** deberta-large, bert-large-uncased, roberta-large [Devlin et al., 2019, He et al., 2020, Zhuang et al., 2021].
- **QWEN family:** qwen2-0.5b, qwen2.5-0.5b, qwen1.5-0.5b, qwen2.5-1.5b, qwen2-1.5b, qwen1.5-1.5b, qwen1.5-4b, qwen2.5-4b, qwen2-7b, qwen1.5-14b, qwen1.5-32b, qwen1.5-72b, qwen2.5-72b [Bai et al., 2023, Yang et al., 2024].
- **Llama family:** llama-3.2-1b, llama-3.1-8b, llama-3-8b, llama-3-70b, llama-3.1-70b [Touvron et al., 2023a,b, Grattafiori et al., 2024].
- **Phi family:** phi-1.5, phi-1, phi-2, phi-4 [Javaheripi et al., 2023, Abdin et al., 2024, Abouelenin et al., 2025].
- **Gemma family:** gemma-2b, gemma-2-2b, gemma-7b, gemma-2-9b [Team et al., 2024, 2025].
- **Mistral family:** mistral-7b-v0.3 [Karamcheti et al., 2021].
- **GPT family:** gpt2, gpt2-medium [Radford et al., 2019].
- **OLMo family:** olmo-7b [Groeneveld et al., 2024].
- **ViT family:** Clip ViT-B/32, Clip ViT-B/16 [Radford et al., 2021].
- **Classic static embeddings:** GloVe, Word2Vec [Pennington et al., 2014, Mikolov et al., 2013a,b].

A.3 Contextual Prompts and Pooling Strategies

Contextual embeddings of LLMs require feeding words into the model through a prompt. Because tokenizers often split a word into multiple tokens, and since some items in our datasets consist of two or more words, we face a design choice regarding how to aggregate token representations. In our methodology, we adopt *average pooling* over the actual tokens, ensuring that all subword pieces contribute equally. Figures 3 and 4 reveal that the average pooling strategy achieves consistent performance and demonstrates the tightest distribution, making it the most reliable choice for our research.

For prompts, we selected a neutral template, `"This is a {word}. "` (with a trailing space), designed to minimize any additional semantic bias on the target item. Figures 5 and 6 show this prompt to balance performance and consistency, making it ideal for baseline comparisons.

In this section, we explore alternative pooling strategies and evaluate a diverse set of prompt templates across multiple models.

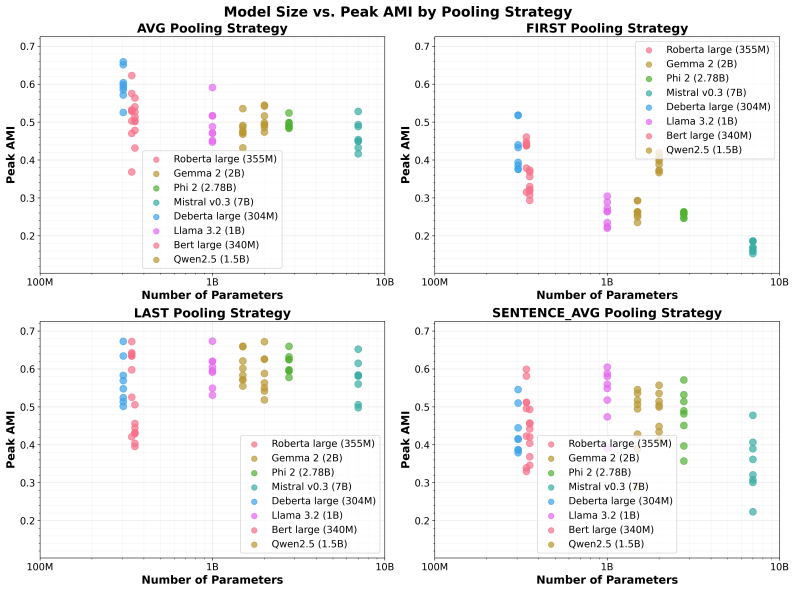


Figure 3: **Average pooling demonstrates consistent performance across different models, making it the most reliable choice for future research.** Each point corresponds to a prompt template applied to a model.

Pooling Strategies. We compare four common approaches:

- **Avg** - mean over all tokens representing the word
- **First** - representation of the first token
- **Last** - representation of the last token
- **Sentence Avg** - mean over the full sentence embedding

Prompt Templates. To test robustness, we design eight templates spanning different linguistic framings:

- `"This is a {word}."`
- `"This is a {word}. "` (with trailing space)
- `"The concept of {word} is"`
- `"When we think of {word}, we consider"`
- `"A typical {word} would be"`
- `"Examples of {word} include"`
- `"The category {word} contains"`
- `"One kind of {word} is"`

Models. We evaluate eight representative LLMs covering major architectures:

- bert-large-uncased (BERT family)
- deberta-large (DeBERTa family)
- gemma-2-2b (Gemma family)
- Llama-3.2-1B (Llama family)
- Mistral-7B-v0.3 (Mistral family)
- phi-2 (Phi family)
- Qwen2.5-1.5B (Qwen family)
- roberta-large (RoBERTa family)

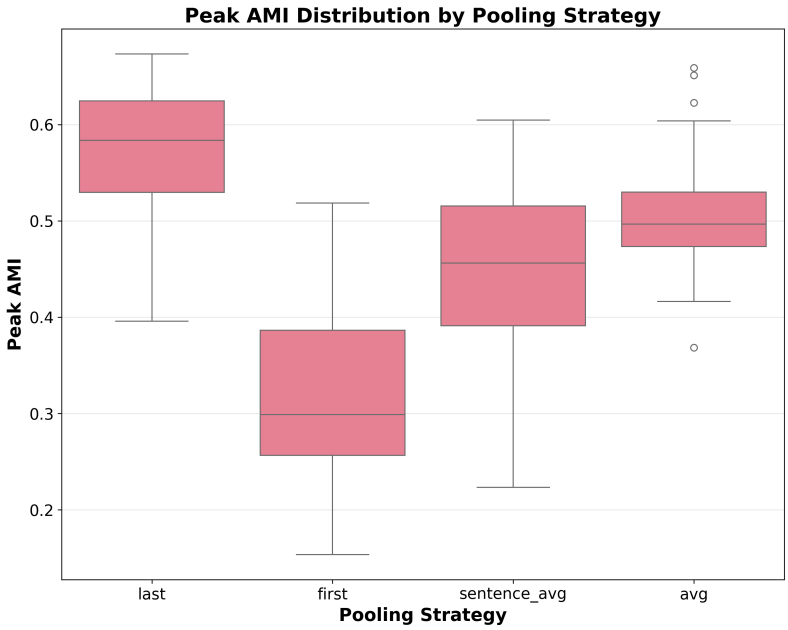


Figure 4: **Average pooling demonstrates the tightest distribution, indicating the highest consistency and reliability across different conditions.** Performance Distribution by Pooling Strategy - Box plots showing AMI distribution for each pooling strategy

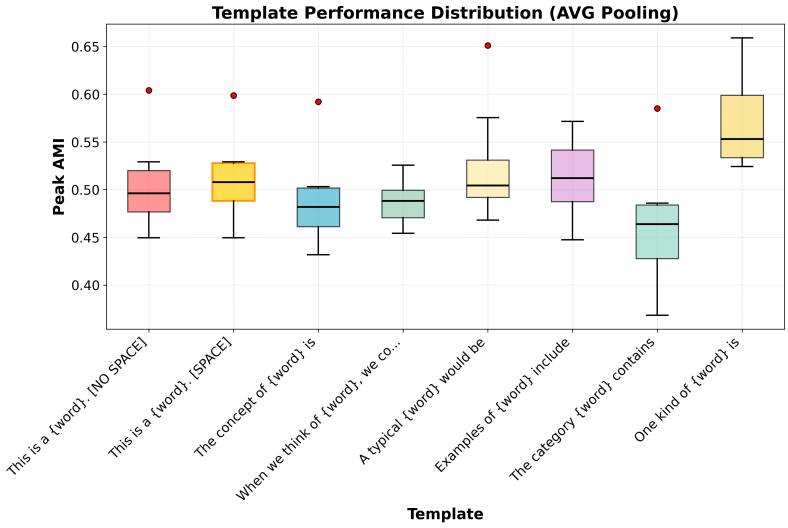


Figure 5: **The neutral prompt template "This is a word. " demonstrates balanced performance and moderate consistency, making it ideal for baseline comparisons.** Performance Distribution across various prompts.

A.4 Static vs. Contextual AMI Exploration

To understand how LLMs develop conceptual alignment with human categories, we examine the progression from static to contextual embeddings. Figure 7 presents three complementary views of this progression across different model scales and architectures.

The left subplot shows **Static AMI** scores, which represent the conceptual alignment achieved by models' input embeddings before any contextual processing (i.e., the E matrix embeddings of the target word). These scores reveal that even at the most basic level, LLMs encode semantic information that supports human-like categorical grouping. Remarkably, classic static models like Word2Vec

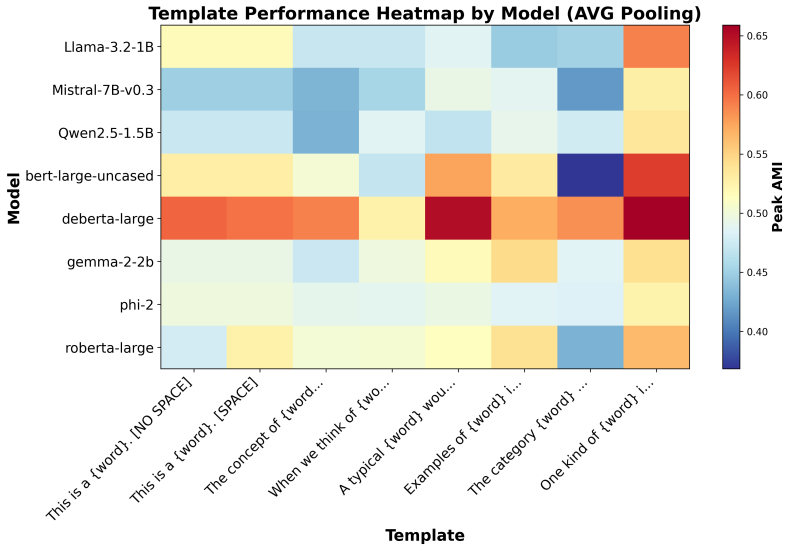


Figure 6: **The neutral template `"This is a {word}. "` remains a stable choice across families.** Heatmap showing template performance per model.

and GloVe achieve static AMI scores that rival the peak contextual performance of modern LLMs, suggesting that fundamental conceptual structure is captured early in the learning process.

The middle subplot displays **Average AMI** across all layers, providing a measure of overall semantic representation quality throughout the network. This metric shows the typical performance a model achieves across its entire depth, offering insight into how consistently different layers maintain conceptual alignment. The improvement from static to average AMI demonstrates that contextual processing generally enhances rather than diminishes semantic understanding.

The right subplot reveals **Peak AMI**, representing the optimal conceptual alignment achieved by any single layer. This metric identifies where in the network conceptual understanding is maximized, typically occurring in middle-to-late layers before declining in the final layers. The progression from static to peak AMI shows that contextual processing not only preserves but significantly enhances the conceptual alignment present in static embeddings.

Several key insights emerge from this multi-metric analysis. First, all models demonstrate above-chance alignment even in their static embeddings, confirming that basic semantic structure is a fundamental property of learned representations. Second, the consistent improvement from static to peak AMI across all model types suggests that contextual processing universally enhances conceptual understanding rather than creating it de novo. Third, encoder architectures of different types (BERT, ViT encoders, and classic models) achieve comparable or superior performance to much larger decoder models, highlighting that architectural factors and pre-training objectives significantly influence conceptual alignment quality beyond mere model scale.

This analysis complements the main text findings by showing that LLMs do not simply achieve above-chance alignment with human categories but rather so through a systematic progression from basic to sophisticated conceptual representations, with contextual processing serving as an amplifier rather than a generator of semantic understanding.

A.5 Training Dynamics

The OLMo analysis examines how semantic structure develops during training by analyzing 57 intermediate checkpoints from the OLMo-7B model, representing evenly spaced sampling (every 10K training steps) spanning from 1K to 557K steps (covering approximately 4B to 2.5T tokens).

The analysis employs two complementary sampling strategies: **representative sampling (6 checkpoints)** captures major developmental phases at 1K, 101K, 201K, 301K, 401K, and 501K steps, while **high-resolution sampling (57 checkpoints)** reveals the inherent noise and fluctuations in training.

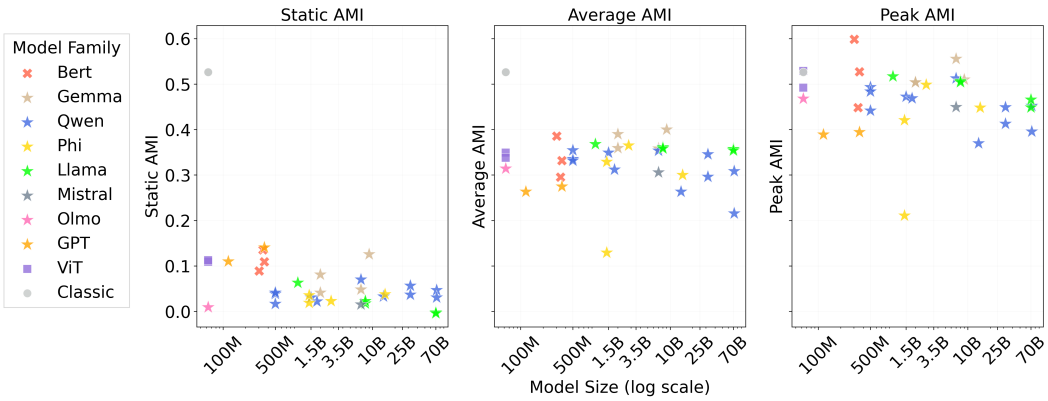


Figure 7: **LLMs begin with basic conceptual alignment in static embeddings and achieve progressively stronger alignment through contextual processing.** Three subplots showing model size (log scale) versus different AMI metrics. **Left:** Static AMI reveals baseline categorical structure. **Middle:** Average AMI across layers shows overall semantic quality. **Right:** Peak AMI demonstrates high conceptual alignment. The consistent improvement from static to peak AMI across all model types reveals that contextual processing enhances rather than creates conceptual understanding, with encoders (BERT, ViT encoders and classic models) achieving comparable or superior performance to much larger decoder models.

Despite significant training noise, the overall semantic development follows a stable, predictable pattern captured by the representative sampling, as shown in Figure 8. The complete training trajectory with all 57 checkpoints is presented in Figure 9.

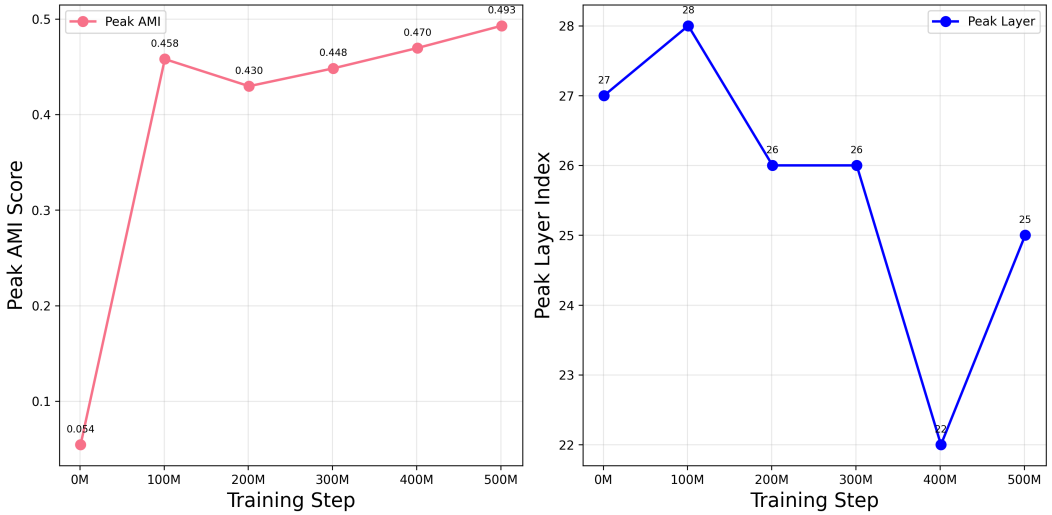


Figure 8: **Left: OLMo-7B representations steadily strengthen during training**: Concept representations develop rapidly at early steps, then refine more gradually over time. **Right: Semantic processing shifts from deep to mid-network layers**: The model undergoes a two-phase dynamic - initially moving semantic processing upward during rapid learning, then reorganizing to optimize efficiency while preserving performance. To improve readability, we present six representative checkpoints that capture the trend.

A.6 Datasets Polysemy

Scope. We quantify lexical ambiguity in our psycholinguistic stimuli by counting the distinct WordNet synsets associated with each lemma. This *polysemy score* lets us estimate how many alternative senses a model must implicitly conflate when it produces a single embedding for a word.

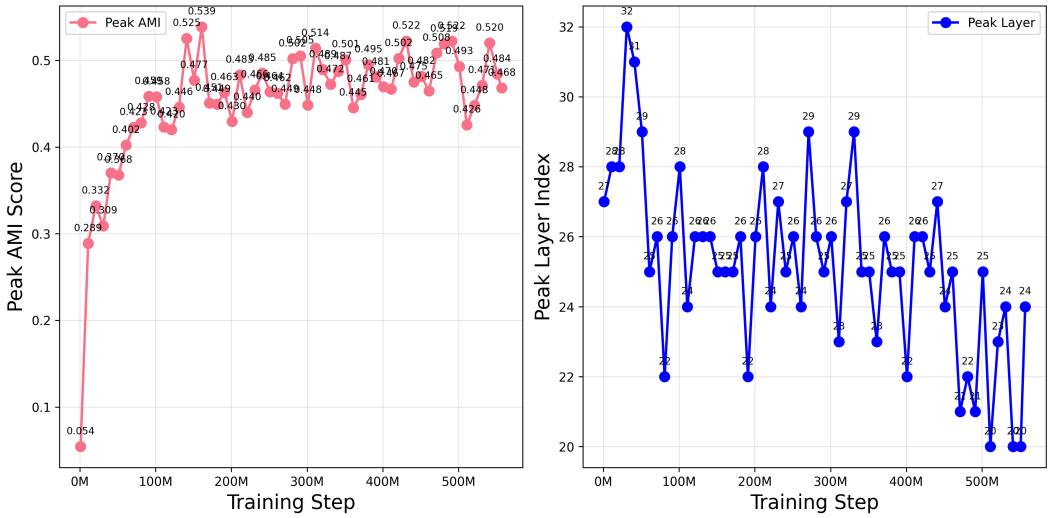
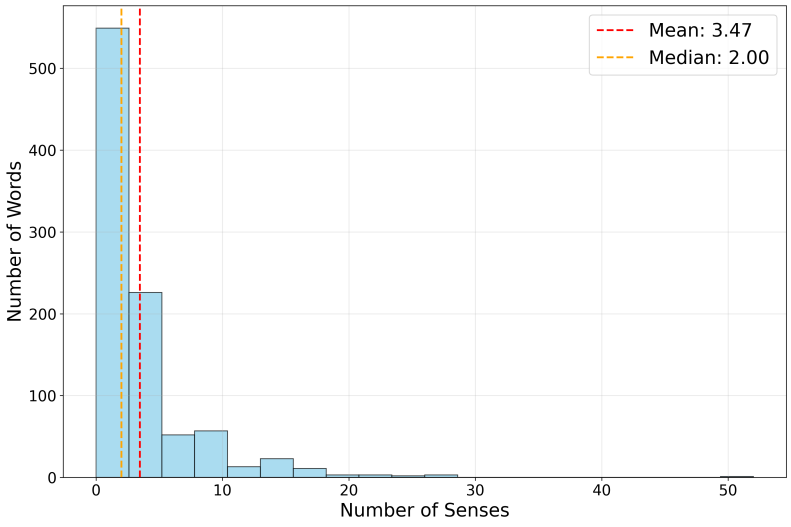


Figure 9: **Complete OLMo-7B training trajectory across 57 checkpoints**: This high-resolution view reveals the inherent noise and fluctuations in training, with individual checkpoint measurements varying throughout the process. Despite this variability, the overall trend aligns with the stable pattern shown in Figure 8, demonstrating that representative sampling effectively captures the underlying semantic development trajectory while filtering out training noise.



][t]

Figure 10: **Polysemy is not likely to influence our results as most of the words in our dataset are concrete nouns, which tend to be less polysemous.** Histogram of WordNet sense counts for the 943 lemmas in our benchmark. The dashed lines indicate the median and mean (2 senses, 3.47 senses, respectively).

Why it matters. Consider *bat*, which can denote either a flying mammal or a piece of sports equipment. The same vector must account for both senses. Aggregating semantically distant senses can blur the representation and thus confound model-human comparisons, especially in tasks that rely on fine-grained semantic similarity. Explicitly tracking polysemy allows us to verify that any performance effects we observe are not artefacts of lexical ambiguity.

Results Figure 10 shows the distribution of polysemy scores. The majority of items are unambiguous (1-2 senses), but a heavy-tailed minority (e.g. *Running* (52 senses), *Saw* (28), *Block* (28)) is highly polysemous. This suggests that our findings are due to real differences between the models rather than polysemy-related artifacts. An additional 141 lemmas that are not in WordNet were omitted.

A.7 Tokenizer Analysis

Rational. The tokenizer of a model has a significant influence over the representations: segmentation rules (WordPiece vs. BPE), vocabulary size and special control tokens can inflate sequence length, skew frequency statistics, and shape error patterns. To ensure fair cross-model comparisons, we therefore (i) cluster checkpoints by the *tokenizer they use* and (ii) quantify how much those tokenizers overlap when applied to our datasets.

Procedure. Before computing overlap, we normalize the vocabulary, stripping tokenizer-specific characters; SentencePiece prefixes (_), GPT-style BPE space prefixes (`Ġ`/`ġ`) and newline markers (`Ċ`), WordPiece continuations markers (##), and related block characters. After this cleanup, tokens differing only by such prefixes collapse to a shared canonical form (e.g. `_house`, `Ġhouse`, and `##house` all become `house`). We then compute pair-wise Jaccard similarity on these cleaned vocabularies.

Table 1 summarizes the core statistics and information regarding the tokenizer types, while Figure 11 visualizes the resulting pairwise vocabulary overlap.

Findings. We find that most tokenizer families share substantial lexical overlap, often exceeding 60%, suggesting a de-facto common token inventory across recent open-source models. First-generation BERT WordPiece (bert-large-uncased and bert-base-uncased) are an outlier, sharing under 16% of tokens with any other group.

Model Family	Mean Tokens / Item	Vocabulary Size	Tokenizer Type
BERT	1.65	30K	WordPiece
DeBERTa & RoBERTa	2.30	50K	WordPiece
GPT	2.30	50K	BPE
Gemma	1.65	256K	SentencePiece (subword)
Llama	2.19	128K	BPE (SentencePiece/Tiktoken)
Mistral	2.35	32K	BPE + control tokens
Phi	2.30	32K	BPE (SentencePiece/Tiktoken)
Qwen	2.19	151.6K	BPE

Table 1: **Tokenizer statistics by tokenizer family.** Mean Tokens/Items refer to the average of the tokens per item in our datasets. The columns Vocabulary Size and Tokenizer Type are properties of the tokenizer.

Model Clustering By Tokenizer Family

- **Llama:** Llama-3.2-1B (representative), Llama-3.1-8B, Meta-Llama-3-8B, Llama-3.1-70B, Meta-Llama-3-70B
- **Gemma:** gemma-2-9b (representative), gemma-7b, gemma-2b, gemma-2-2b
- **Mistral:** Mistral-7B-v0.3 (unique)
- **Phi:** phi-2 (representative), phi-1, phi-1.5
- **RoBERTa:** roberta-large (unique)
- **DeBERTa:** deberta-large (unique)
- **GPT**: gpt2-medium (representative), gpt2
- **BERT:** bert-large-uncased (representative), bert-base-uncased
- **Qwen:** Qwen1.5-0.5B (representative), Qwen1.5-1.8B, Qwen1.5-14B, Qwen2-0.5B, Qwen2-7B, Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen2.5-32B, Qwen1.5-32B, phi-4[1]

A.8 Additional Clustering Metrics

To further validate our cluster alignment findings (Section 5.1), in addition to Adjusted Mutual Information (AMI) and the Normalized Mutual Information (NMI), we also computed the Adjusted

[1] phi-4 has a different tokenizer than the rest of Phi family. The results of its tokenizer match the tokenizer of the Qwen family.

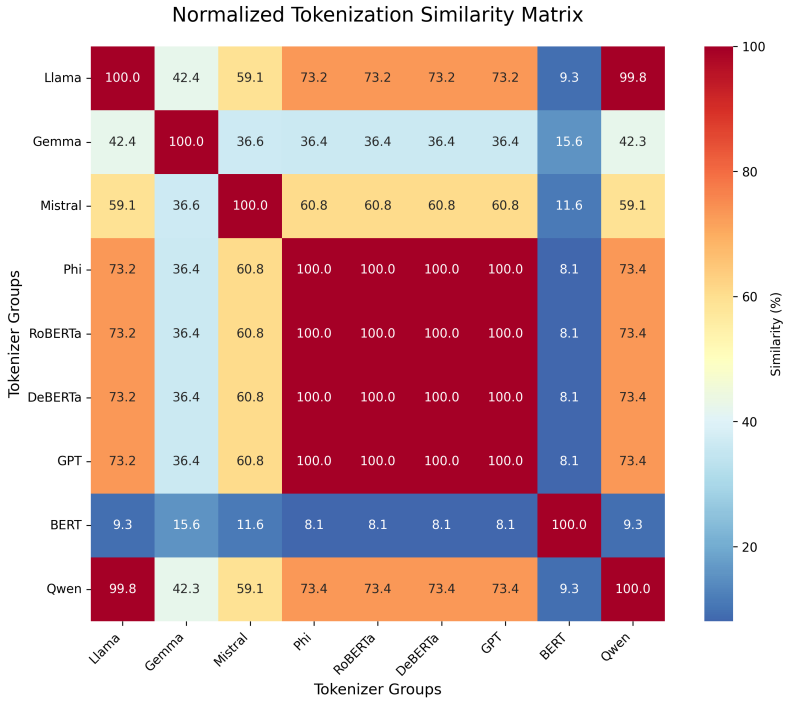


Figure 11: **Substantial lexical overlap suggests that tokenization differences alone cannot explain the observed performance variations in our experiments.** Vocabulary overlap between different tokenizers; most tokenizer families share substantial lexical overlap.

Rand Index (ARI) for the k-means clusters derived from LLM embeddings against human-defined categories. ARI measures the similarity between two data clusterings, correcting for chance. Like AMI, a score of 1 indicates perfect agreement and 0 indicates chance agreement.

Across all tested LLMs, the ARI and NMI scores largely mirrored the trends observed with AMI, showing significantly above-chance alignment with human categories and similar relative model performances. Silhouette scores, while more variable, generally indicated reasonable cluster cohesion for both LLM-derived and human categories. Detailed tables of these scores are provided below.

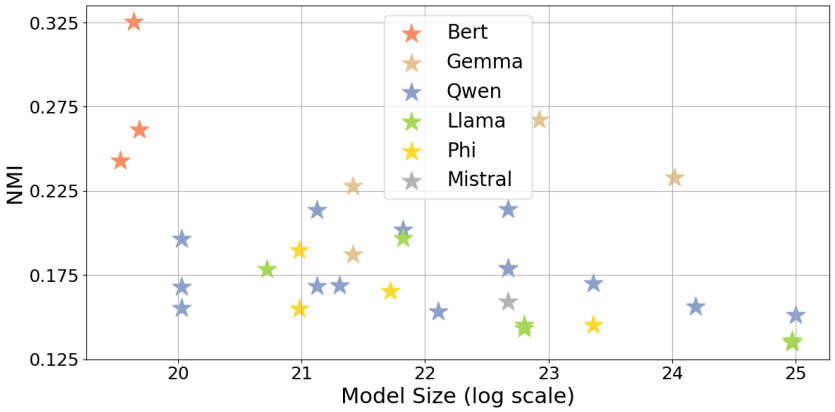


Figure 12: **LLM-derived Clusters Show Above-Chance Alignment with Human Conceptual Categories.** Normalized Mutual Information (NMI) between human-defined categories and clusters from static LLM embeddings. Results are averaged over three psychological datasets. All models perform significantly better than random clustering. BERT's performance is notably strong.

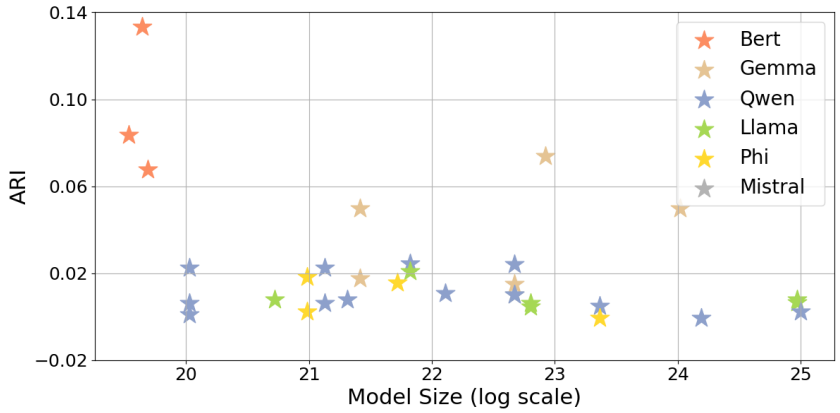


Figure 13: **LLM-derived Clusters Show Above-Chance Alignment with Human Conceptual Categories.** Adjusted Rand Index (ARI) between human-defined categories and clusters from LLM static embeddings. Results are averaged over three psychological datasets. All models perform significantly better than random clustering. BERT's performance is notably strong.

These supplementary metrics reinforce the conclusion that LLMs capture broad human-like conceptual groupings.

A.9 Detailed AMI Scores per Model and Dataset

Table 2 provides a more granular view of the static AMI scores for each LLM across the three individual psychological datasets.

Dataset	Model	NMI	AMI	ARI
[Rosch, 1973c]	bert-large-uncased	0.19453	0.2011	0.11336
[Rosch, 1975]	bert-large-uncased	0.16547	0.27324	0.2216
[McCloskey and Glucksberg, 1978]	bert-large-uncased	0.12003	0.15934	0.06306
[Rosch, 1973c]	FacebookAI/roberta-large	0.1021	0.10666	0.03393
[Rosch, 1975]	FacebookAI/roberta-large	0.12138	0.23938	0.14165
[McCloskey and Glucksberg, 1978]	FacebookAI/roberta-large	0.06271	0.08873	0.03173
[Rosch, 1973c]	google-t5/t5-large	0.16583	0.16855	0.03676
[Rosch, 1975]	google-t5/t5-large	-0.03799	0.04179	0.00758
[McCloskey and Glucksberg, 1978]	google-t5/t5-large	0.06146	0.08825	0.0082
[Rosch, 1973c]	google/gemma-2-27b	0.08523	0.09065	0.04158
[Rosch, 1975]	google/gemma-2-27b	0.04276	0.10062	0.06244
[McCloskey and Glucksberg, 1978]	google/gemma-2-27b	0.07814	0.10274	0.04364
[Rosch, 1973c]	google/gemma-2-2b	0.04029	0.04107	0.01212
[Rosch, 1975]	google/gemma-2-2b	0.04529	0.14844	0.07596
[McCloskey and Glucksberg, 1978]	google/gemma-2-2b	0.09953	0.13593	0.06326
[Rosch, 1973c]	google/gemma-2-9b	0.1222	0.12757	0.06053
[Rosch, 1975]	google/gemma-2-9b	0.07841	0.16126	0.09617
[McCloskey and Glucksberg, 1978]	google/gemma-2-9b	0.10879	0.13997	0.06439
[Rosch, 1973c]	google/gemma-2b	0.04336	0.04616	0.01593
[Rosch, 1975]	google/gemma-2b	-0.00353	0.04483	0.01577
[McCloskey and Glucksberg, 1978]	google/gemma-2b	0.03472	0.05484	0.02142
[Rosch, 1973c]	google/gemma-7b	0.04459	0.04547	0.01052
[Rosch, 1975]	google/gemma-7b	-0.03055	0.02644	0.01506
[McCloskey and Glucksberg, 1978]	google/gemma-7b	0.03338	0.05724	0.02176
[Rosch, 1973c]	meta-llama/Llama-3.1-70B	0.03008	0.03528	0.01936
[Rosch, 1975]	meta-llama/Llama-3.1-70B	-0.07026	0.02636	0.00392
[McCloskey and Glucksberg, 1978]	meta-llama/Llama-3.1-70B	-0.04773	0.00972	0.00236
[Rosch, 1973c]	meta-llama/Llama-3.1-8B	0.00473	0.00393	0.00023
[Rosch, 1975]	meta-llama/Llama-3.1-8B	-0.03928	0.05489	0.01884
[McCloskey and Glucksberg, 1978]	meta-llama/Llama-3.1-8B	-0.02671	0.02208	6.00E-05

[Rosch, 1973c]	meta-llama/Llama-3.2-1B	0.01936	0.01567	0.00246
[Rosch, 1975]	meta-llama/Llama-3.2-1B	-0.01876	0.05663	0.00782
[McCloskey and Glucksberg, 1978]	meta-llama/Llama-3.2-1B	0.03625	0.06798	0.01352
[Rosch, 1973c]	meta-llama/Llama-3.2-3B	0.03757	0.03537	0.00876
[Rosch, 1975]	meta-llama/Llama-3.2-3B	0.01893	0.09619	0.03193
[McCloskey and Glucksberg, 1978]	meta-llama/Llama-3.2-3B	0.03914	0.07395	0.0202
[Rosch, 1973c]	meta-llama/Meta-Llama-3-70B	0.02289	0.03133	0.01514
[Rosch, 1975]	meta-llama/Meta-Llama-3-70B	-0.06428	0.0185	0.00554
[McCloskey and Glucksberg, 1978]	meta-llama/Meta-Llama-3-70B	-0.04595	0.01068	0.00272
[Rosch, 1973c]	meta-llama/Meta-Llama-3-8B	0.03512	0.02852	0.00225
[Rosch, 1975]	meta-llama/Meta-Llama-3-8B	-0.06011	0.03694	0.00676
[McCloskey and Glucksberg, 1978]	meta-llama/Meta-Llama-3-8B	-0.0355	0.0219	0.00676
[Rosch, 1973c]	microsoft/deberta-large	0.03748	0.03909	0.01467
[Rosch, 1975]	microsoft/deberta-large	0.16568	0.28993	0.20527
[McCloskey and Glucksberg, 1978]	microsoft/deberta-large	0.03217	0.06175	0.03019
[Rosch, 1973c]	microsoft/phi-1_5	0.02102	0.01786	0.0075
[Rosch, 1975]	microsoft/phi-1_5	0.03989	0.13887	0.04305
[McCloskey and Glucksberg, 1978]	microsoft/phi-1_5	0.00895	0.05215	0.00639
[Rosch, 1973c]	microsoft/phi-1	0.0249	0.01698	0.00133
[Rosch, 1975]	microsoft/phi-1	-0.03625	0.02811	0.00217
[McCloskey and Glucksberg, 1978]	microsoft/phi-1	-0.01148	0.03085	0.00371
[Rosch, 1973c]	microsoft/phi-2	0.03703	0.02968	0.00404
[Rosch, 1975]	microsoft/phi-2	-0.03654	0.04227	0.03942
[McCloskey and Glucksberg, 1978]	microsoft/phi-2	-0.00254	0.02531	0.00533
[Rosch, 1973c]	microsoft/phi-4	0.03075	0.03043	0.01076
[Rosch, 1975]	microsoft/phi-4	-0.06737	0.00092	-0.01361
[McCloskey and Glucksberg, 1978]	microsoft/phi-4	-0.01789	0.02705	0.00066
[Rosch, 1973c]	mistralai/Mistral-7B-v0.3	0.0425	0.03507	0.00357
[Rosch, 1975]	mistralai/Mistral-7B-v0.3	-0.05018	0.01217	0.0177
[McCloskey and Glucksberg, 1978]	mistralai/Mistral-7B-v0.3	-0.01264	0.03902	0.00931
[Rosch, 1973c]	Qwen/Qwen1.5-0.5B	0.00148	-0.00225	0.00399
[Rosch, 1975]	Qwen/Qwen1.5-0.5B	-0.01538	0.04833	0.0095
[McCloskey and Glucksberg, 1978]	Qwen/Qwen1.5-0.5B	0.02559	0.06023	0.00771
[Rosch, 1973c]	Qwen/Qwen1.5-1.8B	0.03397	0.03232	0.01034
[Rosch, 1975]	Qwen/Qwen1.5-1.8B	-0.01129	0.05803	0.00683
[McCloskey and Glucksberg, 1978]	Qwen/Qwen1.5-1.8B	-0.00541	0.03614	0.00538
[Rosch, 1973c]	Qwen/Qwen1.5-14B	0.0372	0.02738	0.0028
[Rosch, 1975]	Qwen/Qwen1.5-14B	-0.02604	0.05153	0.01211
[McCloskey and Glucksberg, 1978]	Qwen/Qwen1.5-14B	0.00124	0.04136	0.00338
[Rosch, 1973c]	Qwen/Qwen1.5-32B	0.02638	0.02436	0.00409
[Rosch, 1975]	Qwen/Qwen1.5-32B	-0.03413	0.02526	-0.00665
[McCloskey and Glucksberg, 1978]	Qwen/Qwen1.5-32B	-0.01991	0.02124	-0.00059
[Rosch, 1973c]	Qwen/Qwen1.5-4B	0.03803	0.04058	0.01742
[Rosch, 1975]	Qwen/Qwen1.5-4B	-0.03309	0.03988	0.01678
[McCloskey and Glucksberg, 1978]	Qwen/Qwen1.5-4B	-0.03997	0.00548	-0.00028
[Rosch, 1973c]	Qwen/Qwen1.5-72B	0.03697	0.02892	0.00144
[Rosch, 1975]	Qwen/Qwen1.5-72B	-0.06184	0.02213	0.0017
[McCloskey and Glucksberg, 1978]	Qwen/Qwen1.5-72B	-0.02022	0.02918	0.00297
[Rosch, 1973c]	Qwen/Qwen2-0.5B	0.02266	0.01923	0.00662
[Rosch, 1975]	Qwen/Qwen2-0.5B	0.0515	0.14571	0.04999
[McCloskey and Glucksberg, 1978]	Qwen/Qwen2-0.5B	0.01508	0.04357	0.00643
[Rosch, 1973c]	Qwen/Qwen2-1.5B	0.02956	0.02779	0.00544
[Rosch, 1975]	Qwen/Qwen2-1.5B	-0.03595	0.03443	-0.01099
[McCloskey and Glucksberg, 1978]	Qwen/Qwen2-1.5B	0.01768	0.05407	0.01604
[Rosch, 1973c]	Qwen/Qwen2-7B	0.06424	0.06439	0.02067
[Rosch, 1975]	Qwen/Qwen2-7B	0.0333	0.09155	0.02832
[McCloskey and Glucksberg, 1978]	Qwen/Qwen2-7B	0.05329	0.07599	0.01977
[Rosch, 1973c]	Qwen/Qwen2.5-0.5B	0.03165	0.03291	0.01029
[Rosch, 1975]	Qwen/Qwen2.5-0.5B	-0.06534	-0.0196	-0.01165
[McCloskey and Glucksberg, 1978]	Qwen/Qwen2.5-0.5B	0.0062	0.04191	0.0054
[Rosch, 1973c]	Qwen/Qwen2.5-1.5B	0.04838	0.0489	0.0129

[Rosch, 1975]	Qwen/Qwen2.5-1.5B	0.03785	0.113	0.02761
[McCloskey and Glucksberg, 1978]	Qwen/Qwen2.5-1.5B	0.06166	0.08675	0.03162
[Rosch, 1973c]	Qwen/Qwen2.5-3B	0.03882	0.0348	0.00465
[Rosch, 1975]	Qwen/Qwen2.5-3B	0.03977	0.10821	0.04302
[McCloskey and Glucksberg, 1978]	Qwen/Qwen2.5-3B	0.03416	0.07307	0.02959
[Rosch, 1973c]	Qwen/Qwen2.5-7B	0.0529	0.05051	0.01605
[Rosch, 1975]	Qwen/Qwen2.5-7B	-0.00905	0.03227	0.01044
[McCloskey and Glucksberg, 1978]	Qwen/Qwen2.5-7B	0.00222	0.02759	0.00551

Table 2: Mutual information measures (normalized mutual information, adjusted mutual information, adjusted rand index) per model per dataset. Aggregated results are shown in the main paper and the Figures in the Appendix.

A.10 Correlation between Human Typicality Judgments and LLM Internal Cluster Geometry

The following tables present the Spearman correlation coefficients (ρ) between human typicality judgments and LLM internal representations across different analysis approaches:

Table 3: Static analysis correlations using embeddings from the E matrix. This approach captures the baseline semantic relationships between items and categories without contextual processing.

Table 4: Peak AMI layer analysis correlations using contextual embeddings from the layer that maximized AMI scores (as identified in RQ1). This approach leverages the optimal layer for semantic clustering to assess fine-grained semantic fidelity.

Both tables present correlations across three cognitive science datasets: Rosch (1973), Rosch (1975), and McCloskey (1978), with asterisks (*) indicating statistically significant correlations ($p < 0.05$). The modest correlation values across most models suggest limited alignment between LLM internal representations and human-perceived semantic nuances.

A.11 Typicality and Cosine Similarity [RQ2]

Figure 14 shows representative scatter plots illustrating the relationship between human typicality scores (or psychological distances) and the LLM-derived item-centroid cosine similarities for selected categories and models. These plots visually demonstrate the often modest correlations discussed in Section 5.2.

Figure 15 shows the aggregated Spearman correlation across model families and datasets. These correlations are very weak and mostly non-significant.

A.12 Theoretical Extreme Case Exploration for $\mathcal{L}$

In the case where $|C| = |X|$ (each data point is a cluster of size 1, so $|C_c| = 1 \; \forall c \in C$), then $H(X|C) = \frac{1}{|X|} \sum_{c \in C} 1 \cdot \log_2 1 = 0$. The distortion term $\sigma_c^2 = 0$ for each cluster as the item is its own centroid. Thus, $\mathcal{L} = I(X;C) + \beta \cdot 0 = H(X) - H(X|C) = H(X) = \log_2 |X|$. This represents the cost of encoding each item perfectly without any compression via clustering, and zero distortion.

In the case where $|C| = 1$ (one cluster C_X contains all $|X|$ data points, so $|C_{C_X}| = |X|$), then $H(X|C) = \frac{1}{|X|} |X| \log_2 |X| = \log_2 |X|$. Thus, $I(X;C) = H(X) - H(X|C) = \log_2 |X| - \log_2 |X| = 0$. This represents maximum compression (all items are treated as one). The distortion term becomes $\beta \cdot \frac{1}{|X|} |X| \cdot \sigma_X^2 = \beta \cdot \sigma_X^2$, where σ_X^2 is the variance of all items X with respect to the global centroid of X. So, $\mathcal{L} = 0 + \beta \cdot \sigma_X^2 = \beta \cdot \sigma_X^2$. This represents the scenario of maximum compression where the cost is purely the distortion incurred by representing all items by a single prototype.

A.13 Compression Figures

Figure 18 depicts the IB-RDT objective ($\mathcal{L}$) vs. K. Lower $\mathcal{L}$ indicates a more optimal balance between compression ($I(X;C)$) and semantic fidelity (distortion). Human categories (fixed K) show higher $\mathcal{L}$ values.

A.14 $\mathcal{L}$ objective vs. downstream task performance

Analysis of 13 instruction-tuned models across 5 families (Qwen, Llama, Gemma, Phi, Mistral; see Table 5 for results) indicates no statistical significance ($r = -0.202$, $p = 0.508$). This finding suggests that while the $\mathcal{L}$

Model	Dataset Correlation (Spearman ρ) - Static Layer		
	Rosch (1973)	**Rosch (1975)**	**McCloskey (1978)**
Deberta large (304M)	0.144	0.107*	0.075
Bert large (340M)	0.378*	0.275*	0.250*
Roberta large (355M)	0.005	0.038	-0.029
Gemma (2B)	0.069	0.078	0.007
Gemma 2 (2B)	0.236	0.119*	0.147*
Gemma (7B)	0.131	0.100*	0.007
Gemma 2 (9B)	0.280	0.135*	0.199*
Gemma 2 (27B)	0.112	0.122*	0.161*
Qwen1.5 (0.5B)	0.175	0.076	0.096*
Qwen2 (0.5B)	0.238	0.041	0.040
Qwen2.5 (0.5B)	0.212	0.027	0.037
Qwen2.5 (1.5B)	0.141	0.086*	0.078
Qwen1.5 (1.8B)	0.172	0.134*	0.154*
Qwen2 (7B)	0.036	0.087*	0.040
Qwen1.5 (14B)	0.154	0.086*	0.108*
Qwen1.5 (32B)	-0.032	0.100*	0.081
Qwen2.5 (32B)	0.035	0.105*	0.084
Mistral v0.3 (7B)	0.076	0.152*	-0.009
Llama 3.2 (1B)	0.301*	0.056	0.039
Llama 3 (8B)	-0.002	0.099*	0.080
Llama 3.1 (8B)	0.004	0.108*	0.081
Llama 3 (70B)	0.148	0.161*	0.155*
Llama 3.1 (70B)	0.122	0.161*	0.155*
Phi 1 (1.42B)	0.071	0.052	0.054
Phi 1.5 (1.42B)	-0.088	0.079	0.018
Phi 2 (2.78B)	-0.056	0.044	0.024
Phi 4 (14.7B)	0.079	0.086*	0.097*
T5 Large (770M)	0.235	0.259*	0.178*
GPT-2 Medium (355M)	-0.032	0.063	-0.017
ViT-B/32 Text (63.1M)	0.527*	0.315*	0.286*
ViT-B/16 Text (63.1M)	0.528*	0.289*	0.278*
Word2Vec (300D)	0.442*	0.349*	0.437*
Glove (300D)	0.315*	0.333*	0.350*

Table 3: **Correlation between Human Typicality Judgments and LLM Internal Cluster Geometry.** Spearman static-layer rank correlations between human-rated psychological typicality/distance (higher human scores = less typical/more distant) and item-to-centroid cosine similarity (higher similarity = more central to LLM cluster). $^*p < 0.05$.

objective successfully identifies models that compress semantic categories more effectively, this compression ability does not directly translate to improved performance on standard NLP benchmarks. The lack of correlation implies that concept compression and benchmark accuracy represent distinct aspects of model capability, with the former capturing semantic organization efficiency and the latter measuring general knowledge and reasoning abilities. We specifically chose instruction-tuned models to ensure fair comparison on MMLU, as base models would likely perform poorly on this instruction-following benchmark. While our analysis covers a diverse range of model families and sizes, this represents a subset of available models due to the limited availability of instruction-tuned variants.

Model	Dataset Correlation (Spearman ρ) - Peak AMI Layer		
	Rosch (1973)	**Rosch (1975)**	**McCloskey (1978)**
Deberta large (304M)	0.277	0.107*	0.126*
Bert large (340M)	-0.120	0.148*	0.026
Roberta large (355M)	-0.011	0.038	-0.022
Gemma (2B)	0.127	0.092*	0.039
Gemma 2 (2B)	0.034	0.139*	0.103*
Gemma (7B)	0.004	-0.050	0.088
Gemma 2 (9B)	0.047	0.110*	0.098*
Gemma 2 (27B)	-0.135	0.090*	-0.101*
Qwen1.5 (0.5B)	-0.025	-0.064	0.122*
Qwen2 (0.5B)	0.090	0.064	-0.077
Qwen2.5 (0.5B)	-0.012	0.121*	0.072
Qwen2.5 (1.5B)	-0.114	0.084*	-0.028
Qwen1.5 (1.8B)	0.092	-0.044	-0.004
Qwen2 (7B)	0.004	0.049	-0.087
Qwen1.5 (14B)	0.032	0.091*	0.043
Qwen1.5 (32B)	0.026	0.082	0.111*
Qwen2.5 (32B)	0.045	0.079	-0.060
Mistral v0.3 (7B)	0.013	0.039	0.107*
Llama 3.2 (1B)	0.063	0.094*	0.070
Llama 3 (8B)	-0.045	0.138*	0.084
Llama 3.1 (8B)	-0.096	0.130*	0.087
Llama 3 (70B)	-0.108	0.023	0.050
Phi 1 (1.42B)	0.216	0.185*	-0.029
Phi 1.5 (1.42B)	0.162	0.098*	0.015
Phi 2 (2.78B)	0.325*	0.142*	-0.033
Phi 4 (14.7B)	0.079	0.129*	0.007
T5 Large (770M)	0.219	0.226*	0.282*
GPT-2 Small (117M)	-0.046	0.118*	0.010
GPT-2 Medium (355M)	0.077	0.109*	-0.029
ViT-B/32 Text (63.1M)	0.128	0.055	0.152*
ViT-B/16 Text (63.1M)	0.089	0.086*	0.127*
Word2Vec (300D)	0.442*	0.349*	0.437*
Glove (300D)	0.315*	0.333*	0.350*

Table 4: **Correlation between Human Typicality Judgments and LLM Internal Cluster Geometry.** Spearman peak-AMI layer rank correlations between human-rated psychological typicality/distance (higher human scores = less typical/more distant) and item-to-centroid cosine similarity (higher similarity = more central to LLM cluster). $^*p < 0.05$.

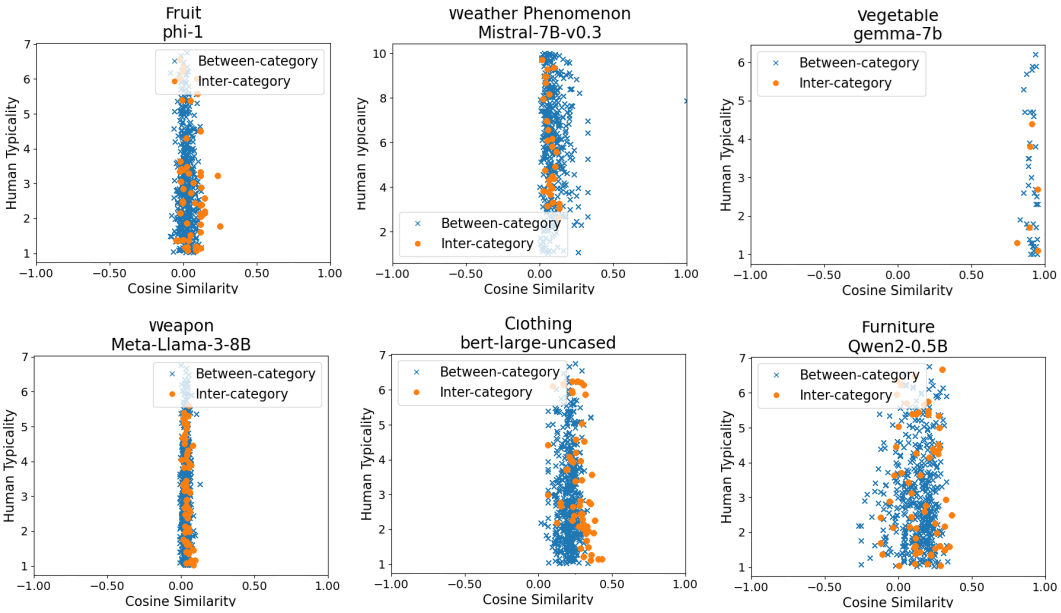


Figure 14: **Weak-to-No Correlation Between LLM Embedding Distance and Human Typicality Judgments.** Scatter plot examples of the cosine similarity versus the human typicality of items belonging to the category compared to items from other categories.

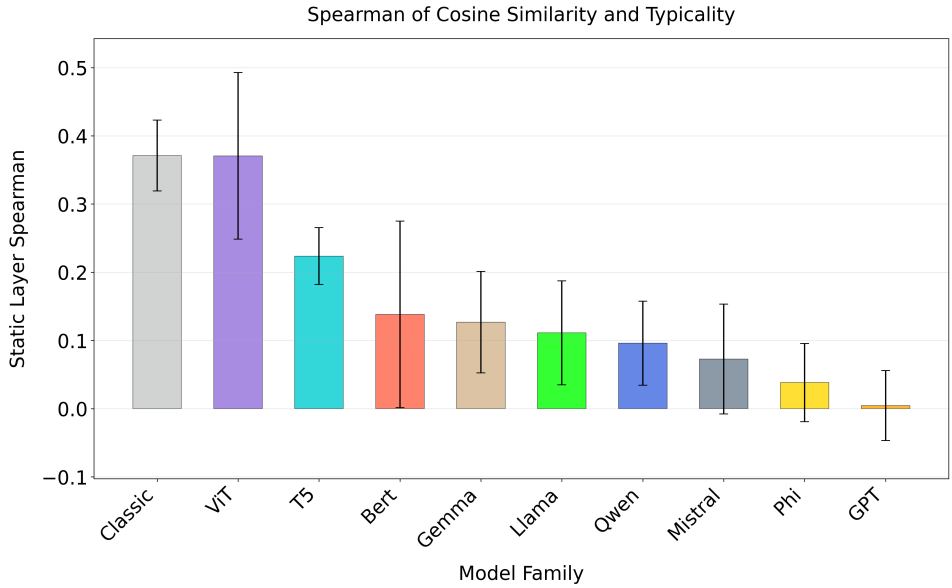


Figure 15: **Weak and Mostly Non-Significant Spearman Correlation Values Between Human Typicality Judgments and LLM Cosine Similarity Indicating Different Structure Representing Concepts.** Mean Static Layer Spearman correlation values across the models belonging to the same family and across the three datasets.

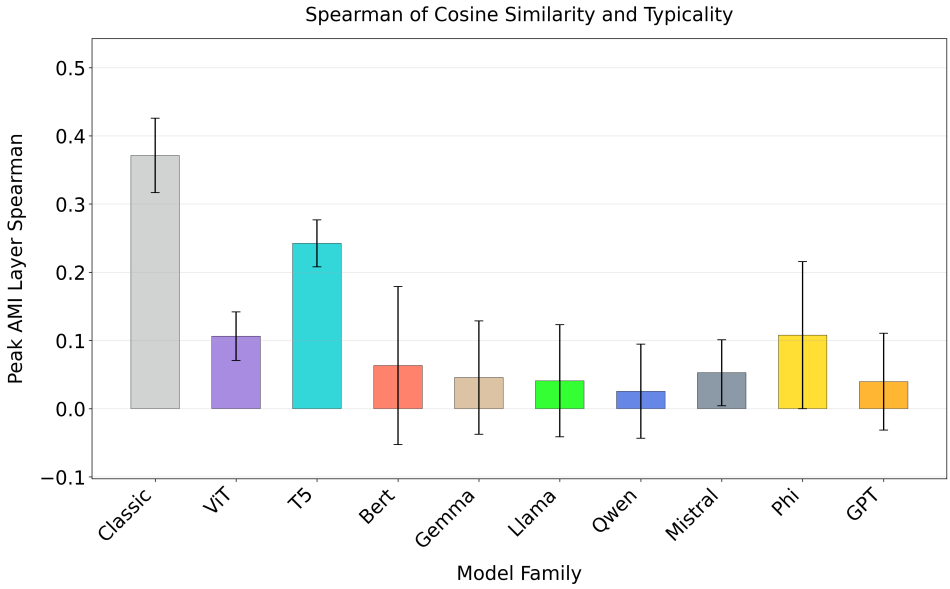


Figure 16: **Weak and Mostly Non-Significant Spearman Correlation Values Between Human Typicality Judgments and LLM Cosine Similarity Indicating Different Structure Representing Concepts.** Mean Peak AMI Layer Spearman correlation values across the models belonging to the same family and across the three datasets.

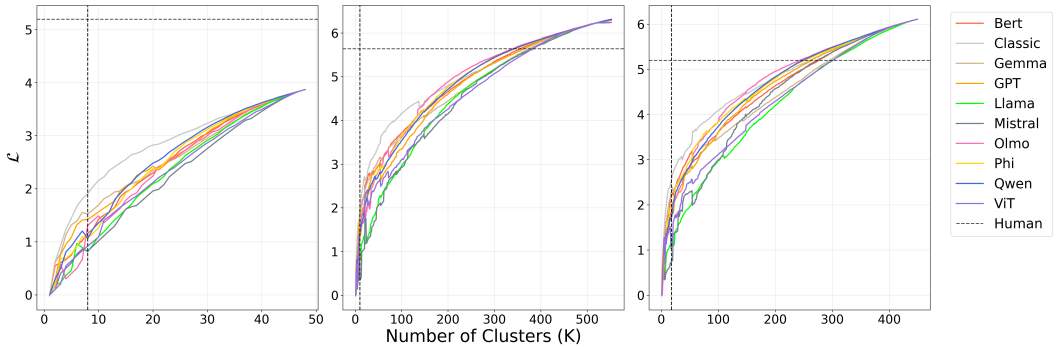


Figure 17: **Static Embeddings Achieve a more "Optimal" Compression-Meaning Trade-off by the $\mathcal{L}$ Measure.** IB-RDT objective ($\mathcal{L}$) vs. K across all datasets. Lower $\mathcal{L}$ indicates a more optimal balance between compression ($I(X;C)$) and semantic fidelity (distortion). Static embeddings consistently achieve lower $\mathcal{L}$ values than both human categories and contextual embeddings. The plots correspond to the three datasets in the following order: Rosch [1973a], Rosch [1975], McCloskey and Glucksberg [1978].

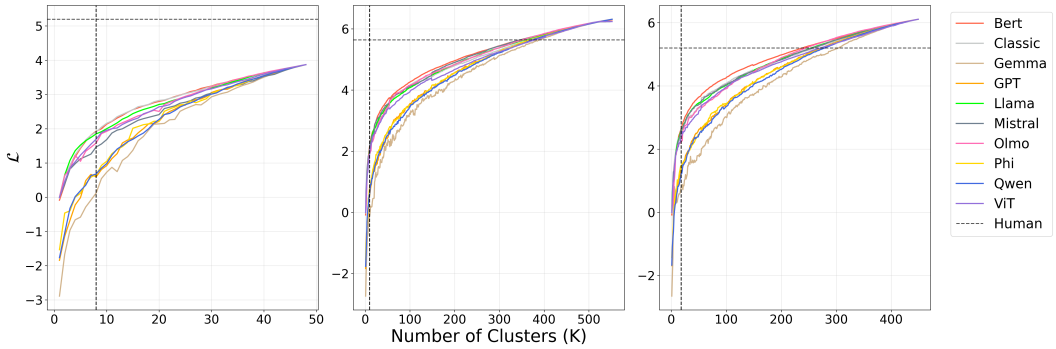


Figure 18: **Contextual Embeddings Achieve Better-than-Human Compression-Meaning Trade-off by the $\mathcal{L}$ Measure.** IB-RDT objective ($\mathcal{L}$) vs. K across all datasets. Lower $\mathcal{L}$ indicates a more optimal balance between compression ($I(X;C)$) and semantic fidelity (distortion). Contextual embeddings outperform human categories but achieve higher $\mathcal{L}$ values than static embeddings. The plots correspond to the three datasets in the following order: Rosch [1973a], Rosch [1975], McCloskey and Glucksberg [1978].

Model	**Size**	**$\mathcal{L}$ Score**	**MMLU Score**
Qwen2-0.5B-Instruct	494M	1.930	0.433
Qwen2.5-0.5B-Instruct	494M	1.982	0.469
Llama-3.2-1B-Instruct	1.2B	1.876	0.454
Qwen2.5-1.5B-Instruct	1.5B	2.071	0.597
Gemma-2B-IT	2.5B	2.382	0.366
Gemma-2-2B-IT	2.6B	2.263	0.565
Phi-4-mini-instruct	3.8B	1.905	0.678
Mistral-7B-Instruct-v0.3	7.2B	1.714	0.603
Qwen2-7B-Instruct	7.6B	2.319	0.700
Meta-Llama-3-8B-Instruct	8.0B	1.348	0.647
Llama-3.1-8B-Instruct	8.0B	1.320	0.679
Gemma-7B-IT	8.5B	2.372	0.512
Gemma-2-9B-IT	9.2B	2.467	0.723

Table 5: **No correlation between $\mathcal{L}$ objective scores and MMLU scores across different sizes and families.** Table displays $\mathcal{L}$ objective values vs. MMLU scores for 13 instruction-tuned models. Correlation measured ($r = -0.202$, $p = 0.508$).