从符号到思想：LLMs 和人类如何用压缩换取意义

陈珊妮
斯坦福大学
cshani@stanford.edu

丹·朱拉夫斯基
斯坦福大学
jurafsky@stanford.edu

Yann LeCun
纽约大学，Meta - FAIR

Ravid Shwartz-Ziv
纽约大学，Wand.AI

Abstract

人类通过 语义压缩 将知识组织成紧凑的类别，通过将多样化的实例映射到抽象表示来保留意义（例如，知更鸟 和 蓝鸟 都属于 鸟类；大多数鸟类 会飞）。这些概念反映了表达保真度和表示简单性之间的权衡。大型语言模型 (LLMs) 展现了卓越的语言能力，但它们内部表示在压缩和语义保真度之间是否达到类似人类的权衡尚不清楚。我们介绍了一种新的信息论框架，借鉴了率失真理论和信息瓶颈原理，以定量比较这些策略。通过分析来自一系列 LLMs 的标记嵌入，并与经典的人类分类基准进行对比，我们发现了关键的差异。虽然 LLMs 形成了与人类判断一致的广泛概念类别，但它们难以捕捉对人类理解至关重要的细粒度语义区分。更根本的是，LLMs 表现出强烈的统计压缩倾向，而人类的认知系统似乎优先考虑适应性细微差别和上下文丰富性，即使这导致我们的衡量标准下压缩效率较低。这些发现揭示了当前人工智能与人类认知架构之间的关键差异，为开发更符合人类概念表示的 LLMs 指明了方向。

1 引言：大型语言模型中意义的谜团

> “人类语言中构式定义的类别可能因语言而异，但它们被 映射到一个共同的概念空间，这代表了一种共同的认识遗产，确实是人类心智的地理。”– *Croft [2001]p. 139*

人类形成概念的能力是智能的基石，使我们能够通过从复杂信号中提取意义来管理信息过载。我们通过识别关键特征并将经验压缩成认知上易于处理的摘要 [Murphy, 2004] 来实现这一点。这种概念架构通常是分层的（例如，一个 知更鸟 是一个 鸟，一个 动物 [Rosch et al.,1976]），这是一种强大的语义压缩：不同的实例被映射到紧凑的表示。关键在于，这个过程在表示效率（压缩）和保留关键语义保真度（意义）之间取得了平衡，这是学习和理解的基本权衡。

预印本。审稿中。

大型语言模型（ LLMs ）在处理和生成人类语言方面展现出惊人的能力，执行的任务往往似乎需要深刻的语义理解 [Singh 等人，2024，Li 等人，2024]。尽管如此，一个基本之谜仍然存在：**LLMs 是否真正像人类一样掌握概念和意义，还是它们的成功主要根植于对海量数据集的复杂统计模式匹配?** 鉴于人类能够毫不费力地将大量输入提炼为紧凑、有意义的概念，这一过程受制于信息压缩和语义保真度之间的固有权衡 [Tversky，1977，Rosch，1973b]。

作为人类认知的心理支架，概念能够实现高效的解释、从稀疏数据中进行泛化以及丰富的交流。为了使 LLMs 超越表面层次的模仿并实现更类似人类的理解，调查它们内部表示如何在信息 压缩 和 语义意义的保留 之间进行关键权衡至关重要。LLMs 是否发展出与人类思维的效率和丰富性相媲美的概念结构，还是它们采用了根本不同的表示策略?

为解决这一问题，我们引入了一种基于信息论的新型定量方法。我们开发并应用了一个框架，该框架借鉴了 **率失真理论** [香农，1948] 以及**信息瓶颈**原则 [Tishby 等人，2000] 以系统性地比较 LLMs 和人类概念结构如何在表征复杂性（压缩）与语义保真度之间取得平衡。作为关键的人类基线，我们利用了认知心理学中的开创性数据集，详细描述了人类的分类 [Rosch, 1973a, 1975, McCloskey 和 Glucksberg，1978]。这项工作的贡献之一是这些经典数据集的数字化和公开发布，它们提供了具有高实证严谨性的基准，往往超过现代众包替代方案。我们的框架旨在剖析这些不同系统如何应对压缩 - 意义权衡。

我们的跨多种 LLMs 的比较分析揭示了不同的表征策略。虽然 LLMs 通常形成与人类判断一致的广泛概念类别，但它们往往无法捕捉对人类理解至关重要的细粒度语义区分。更关键的是，我们发现了一个明显的优先级差异：LLMs 表现出强烈的追求激进统计压缩的倾向，而人类概念系统似乎更倾向于适应性的微妙性和语境丰富性，即使这可能会以我们衡量 sheer 压缩效率的潜在成本为代价。这种分歧突出了根本差异，并为开发更具人类对齐概念理解的 AI 提供了途径。

2 研究问题与范围

推动人工智能超越模式匹配迈向更深层次语义理解，关键在于 LLMs 是否发展出类似于人类认知的概念结构。人类概念有效地在语义丰富性与认知可管理性之间取得平衡，这是一种意义与信息压缩之间的权衡。本文探讨了 LLMs 是否以及如何复制这种基本平衡。
先前工作已探索 LLMs 的概念景观，包括它们对关系知识 [Shani 等人，2023]，提取可解释概念的方法 [Hoang-Xuan 等人，2024 年，Maeda 等人，2024]，通过稀疏激活产生的表示 [Li 等人，2024]，嵌入几何层次结构 [Park 等人，2024], 以及自回归概念预测 [Barrault 等人，2024]。虽然富有见地，但这些研究通常缺乏从信息论视角基准对比丰富人类认知数据，对压缩 - 意义权衡 进行深入、定量的比较，或者它们可能未将概念定义建立在既定的认知理论基础上。因此，如何严格比较 LLMs 和人类在表示效率与语义保真度之间取得平衡，仍然是一个关键的开放性问题。此外，认知科学已将信息论应用于人类概念学习 [Imel 和 Zaslavsky，2024 年，Tucker 等人，2025 年，Wolff，2019 年，Sorscher 等人，2022], ，但通常未将其与现代 AI 模型联系起来。

这项工作旨在通过整合认知心理学、信息论和现代自然语言处理来弥合这一差距。我们提出了三个核心研究问题来指导我们的调查：

[RQ1]: LLMs 中涌现的概念在多大程度上与人类定义的概念类别一致? **[RQ2]:** LLMs 和人类在这些概念内部是否表现出相似的内部几何结构，特别是在项目典型性方面?

[研究问题 3]: 人类和 LLMs 在形成概念时，在平衡表征压缩与语义保真度方面的策略有何不同?

这三个问题引导我们的研究，每个问题都通过第 4 节中详细描述的信息论框架来探讨。RQ1 首先通过考察广义概念类别的对齐来研究信息是如何被压缩的。 RQ2 接着深入探讨这些类别的内部精细结构，探究语义细微差别的保留，例如项目典型性。在这些分析的基础上， RQ3 利用整个框架全面比较 LLMs 和人类在压缩与意义之间的权衡优化上的差异。为了使这些比较具有实证基础，我们始终使用经典的人类分类数据集 [Rosch, 1973a, 1975, McCloskey and Glucksberg, 1978] 作为实证基准。我们的总体目标是利用这种比较性的信息论方法，不仅评估当前的 LLMs，而且推进我们对人工和自然智能中高效和有意义表征的理解。

3 与人类认知进行基准测试

要实证研究 LLM 表示与人类概念结构之间的关系，需要两个关键组成部分：人类分类的稳健基准和多样化的 LLM 选择。本节详细介绍了这些组成部分。

3.1 Human Conceptual Baselines: Empirical Data from Seminal Cognitive Science

我们的比较以认知心理学中开创性研究的资料为基础，这些研究绘制了人类分类过程。这些研究提供了丰富的实证证据，展示了人类如何形成概念、判断类别成员资格以及感知典型性。关键在于，与许多现代众包数据集可能存在的噪声不同，这些经典基准是由认知科学专家精心策划的，反映了深层的认知模式而非表面的关联，并且基于当时正在发展的概念结构理论。我们关注三个有影响力的作品：

Rosch (1973): 这项由 Rosch [1973a] 进行的基础性工作探索了语义类别，作为原型理论 [Rosch, 1973c] 研究计划的一部分。该理论认为，类别是围绕 “ 原型 ” 成员组织的，而不是严格的、平等共享的特征。数据集包括八个常见语义类别（例如，家具、鸟类）中的 48 个项目，以及典型性排名（例如， ‘ 知更鸟 ‘ 作为典型的鸟类， ‘ 蝙蝠 ‘ 作为非典型的）。

Rosch (1975): 在原型理论的基础上，Rosch [1975] 进一步详细阐述了语义类别是如何在认知中表征的。这项工作为十类中的 552 个物品提供了广泛的典型性评分（例如， ‘orange ‘ 作为典型的水果， ‘ squash ‘ 则不太典型）。

McCloskey & Glucksberg (1978): McCloskey 和 Glucksberg [1978] 研究了自然类别的 “ 模糊 “ 边界，表明成员资格通常是渐变的而非绝对的。他们的数据涵盖了 18 类中的 449 个物品，包括典型性评分和成员资格确定性评分（例如， ‘dress ‘ 是典型的服装， ‘bandaid ‘ 则不太典型）。

虽然起源于不同的研究小组且具有不同的理论侧重，但这些数据集共享严谨的实验设计，并提供了关于类别分配和物品典型性的数据。我们从这些研究中汇总了数据，创建了一个包含 34 类 1049 个物品的统一基准。这个汇总数据集，我们已经数字化并公开提供（见附录 A.1 ），为评估计算模型的人类相似性提供了至关重要的、高保真度的实证基础，我们鼓励将其用于未来的研究。

3.2 大型语言模型研究

我们包含了一系列多样化的 LLMs，以评估概念表示如何随计算架构和规模而变化。这个选择涵盖了流行的架构范例（仅编码器，仅解码器）和广泛的各种模型大小，从 3 亿到 720 亿参数。

我们的分析包括 BERT 家族的仅编码器模型（例如，BERT-Large [Devlin 等人，2019 年，He 等人，2020 年，Zhuang 等人，2021] ）。大多数是仅解码器的自回归模型，

包括：六个 Llama 家族模型（1B 到 70B，例如 Llama 3.1 70B [Touvron 等人，2023a,b，Grattafiori 等人，2024]）；五个 Gemma 家族模型（2B 到 27B [团队 等人，2024, 2025]）；十三个 Qwen 家族模型（0.5B 到 72B [Bai 等人，2023，Yang 等人，2024]）；四个 Phi 家族模型（例如 Phi-4 [Javaheripi 等人，2023，Abdin 等人，2024，Abouelenin 等人，2025]）；以及一个 Mistral 7B 模型 [Karamcheti 等人，2021]。附录 A.2 提供了所有模型变体、标识符和架构细节的完整列表。

对于每个 LLM，我们从其输入嵌入层（ ‘E ‘ 矩阵）中提取静态、词元级别的嵌入。这种选择使我们的分析与我们通常在人类分类实验中遇到的典型无上下文刺激的性质保持一致，确保了可比的表示基础。这些嵌入构成了我们后续分析中推导 LLM 生成的概念集群的基础。

4 比较压缩和意义的框架

为了理解 LLM 和人类认知如何应对表示意义的根本挑战，我们引入了一个信息论框架。该框架旨在 **分析压缩信息为高效表示和保留对真正理解至关重要的丰富语义保真度之间的关键权衡，或张力**。借鉴 率失真理论 ***(RDT)*** [Shannon, 1948] 以及 信息瓶颈 ***(IB)*** 原则 [Tishby 等人的核心原理，我们的方法为解决我们所有三个研究问题提供了一个连贯的视角。我们的研究通过首先探索与表示紧凑性和语义保留相关的这个权衡的不同方面，然后综合这些见解来评估概念表示的整体效率而进行。通过这个逐步的信息论视角来看，我们的研究问题被如下处理：

> **[RQ1] 通过分类对齐探测表示紧凑性：** 我们首先检查信息如何压缩成分类结构。人类分类和 LLM 推导的聚类都简化了各种项目 X 成结构化的组 C。对于 RQ1，我们通过量化共享信息（例如，通过调整后的互信息）来评估模型集群（ C_{LLM} ）和人类分类（ C_{Human} ）之间的对齐，从而提供了一个初步的视角来了解紧凑性是如何实现的。这里的有效输入表示原则与我们的框架的 “ 复杂性 ” 方面相关。
>
> **[RQ2] 通过内部结构探测语义保留：** 接下来，我们评估这些压缩表示在多大程度上保留了语义。一个有效的系统必须保留关键的语义细微差别。对于 RQ2，我们通过将 LLM 内部的项中心性度量与人类的典型性判断相关联来研究这一点，探究 LLM 是否能够忠实地表示细粒度的语义信息，即 LLM 能否捕捉到 C_{Human} 的内部结构？这与我们框架中的 “ 失真 ” （或保真度）方面相关。
>
> **[RQ3] 评估总表示效率的综合权衡：** 最后，在探索了紧凑性和保留之后，我们利用我们的完整框架。RQ3 采用一个统一的目标准则 $\mathcal{L}$ (详细说明如下)，定量评估 LLM 和人类系统在处理这一基本权衡时的总效率。

The following subsections detail the theoretical underpinnings of this framework.

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

为了严格形式化表示紧凑性和保留语义之间的平衡，我们借鉴了信息论。**率失真理论（ RDT ）** [Shannon，1948] 提供了基础语言。RDT 量化了在最大 “ 失真 ” D (保真度损失) 限制下，表示源 X 为 C 所需的最低 “ 率 ” R (表示复杂性)。目标是优化 $R + \lambda D$，从而对表示效率进行原则性评估。

The **信息瓶颈 (IB) 原则** [Tishby etal., 2000] 是一种相关方法。IB 寻求输入 X 的压缩表示 C，该表示在最大化相关变量 Y 的信息的同时，最小化 $I(X;C)$，即互信息 C 保留关于 X 的信息（瓶颈的 “ 成本 ” ）。这通常被表述为最小化 $I(X;C) - \beta I(C;Y)$。

我们的分析框架直接应用了 RDT 的核心思想，即平衡速率和失真。我们构建了一个目标函数，$\mathcal{L}$，旨在明确平衡一个**复杂度项**（类似于 RDT 的速率），它量化了通过概念簇 C 表示项目 X 的信息成本，以及一个 **失真项**（类似于 RDT 的 D），它测量了这些簇中丢失或模糊的语义信息。我们的复杂度项，结合了 $I(X;C)$，与 IB 原则相呼应。然而，我们的失真项直接测量了簇内语义保真度损失（具体来说，是项目嵌入相对于其簇质心的方差），这与经典的 IB 公式不同，在经典的 IB 公式中，失真通常隐含地与一个外部相关变量 Y 相关联。这种直接方法使我们能够评估任何给定的聚类 C，无论是来自人类认知数据还是 LLM 嵌入，如何本质上平衡其自身的结构紧凑性和其组成部分相对于原始数据 X 的意义。

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

基于这些信息论基础，本节正式定义了我们框架的两个关键组成部分 – 复杂性 和 失真。这些组成部分使我们能够定量地处理先前引入的表示紧凑性（核心到 [RQ1]）和语义保留（中心到 [RQ2]）的方面。然后我们将这些组合成一个统一的目标函数 ,$\mathcal{L}$，旨在评估压缩 - 意义权衡的整体效率，这是 [RQ3] 的主要关注点。该 $\mathcal{L}$ 函数评估从项目 C 中派生的概念簇 X (例如，token 嵌入):

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

Here, $\beta \geq 0$ 是一个平衡两个项相对重要性的超参数。

复杂性（速率）项： 第一个组成部分，**复杂性** (X, C)，衡量通过将其分配到簇 C 来表示原始项目 X 的信息成本或复杂性。它由项目及其簇标签之间的互信息 $I(X;C)$ 量化。较低的 $I(X;C)$ 表示更大的压缩，这意味着簇分配 C 使特定项目 X 更具可预测性（即，除了簇标签之外，指定它们需要更少的信息）。定义 $I(X;C) = H(X) - H(X|C)$，并假设 $|X|$ 对于初始熵计算是等概率的唯一项目 ($H(X) = \log_2 |X|$)，条件熵是 $H(X|C) = \frac{1}{|X|} \sum_{c \in C} |C_c| \log_2 |C_c|$。这假设对于这个复杂性计算，每个簇 C_c (大小为 $|C_c|$) 内的项目在其共享标签 c 之外是不可区分的。因此：

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

该术语形式化了代表紧凑性方面，这是 [RQ1] 的核心。

失真项： 第二个组件， **Distortion**(X, C)，量化了将项目分组到簇中导致的语义保真度损失。它被测量为项目嵌入的平均簇内方差，反映了项目与其簇中心趋势的紧密程度以及簇的语义一致性。这直接关系到细粒度语义信息的保留，该思想在 [RQ2] 中探讨。对于每个簇 $c \in C$，其质心是 $x_c = \frac{1}{|C_c|} \sum_{x \in c} x$(其项目的平均嵌入)。其内部方差是 $\sigma_c^2 = \frac{1}{|C_c|} \sum_{x \in c} \|x - x_c\|^2$。聚类 C 的总失真是这些方差的加权平均值：

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

较低的失真值意味着，平均而言，项目与其各自的簇质心较近，表明每个簇内更好地保留了共享的语义特征。

统一目标函数： 将复杂度（公式 2 ）和失真（公式 3 ）的正式定义代入我们的一般公式 $\mathcal{L}$（公式 1 ）中，得到支撑我们比较分析完整的目标函数：

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

This $\mathcal{L}$ 函数提供了一个单一、原则性的度量标准，用于评估给定的聚类方法在多大程度上有效地平衡了信息压缩的需求与保留语义意义的要求，作为直接量化工具来处理 [RQ3]。

随着 $\mathcal{L}$ 目标现在完全明确，我们的信息论框架提供了一个全面的工具包。复杂度项（公式 2）使我们能够量化与 [RQ1]，相关的表示紧凑性方面，而失真项（公式 3）则能够评估语义保留，这对于 [RQ2] 至关重要。整体 $\mathcal{L}$ 函数（公式 4）然后直接促进了对集成压缩 - 意义权衡的评估，这是 [RQ3] 的核心。因此，该框架使我们能够系统地和定量地研究 LLM 和人类认知如何在信息效率和语义丰富性之间进行平衡。我们在第 5 节中详细介绍了该框架的实证研究应用。

5 Unpacking Representational Strategies: An Empirical Investigation

基于我们的信息论框架（第 4 节）和已建立的基准（第 3 节），我们现在实证地研究我们的研究问题。本节详细介绍了用于比较 LLM 和人类概念策略的具体方法，这些策略涉及概念一致性、内部语义结构和整体表示效率等关键维度。

[RQ1] 评估概念一致性 为了研究 LLM 得出的概念类别如何与人类定义的类别一致（RQ1），我们通过 k-means（K 由每个数据集的人类类别计数设置）对 LLM token 嵌入进行聚类。使用调整后的互信息（AMI）、归一化互信息（NMI）和调整后的兰德指数（ARI）来量化与人类类别的对齐度，并与随机聚类基线进行比较。

[RQ2] 检查内部聚类几何形状和语义保留 为了评估 LLM 表示如何捕捉人类般的典型性（RQ2），检查内部类别几何形状，我们计算每个项目的 token 嵌入与其人类分配的类别名称的 token 嵌入之间的余弦相似度（例如，‘robin‘ 到 ‘bird‘）。然后，这些 LLM 得出的相似度与来自我们认知科学数据集的人类典型性评分相关联（Spearman 的 ρ）。

[RQ3] 评估压缩 - 意义权衡的效率 为了评估压缩和意义的整体平衡（RQ3），我们应用我们的框架，通过计算 $\mathcal{L}$ 目标（公式 4，$\beta = 1$）来评估人类和 LLM 得出的概念结构（后者通过在 K 范围内进行 k-means）。这比较了每个系统如何平衡复杂性 $I(X;C)$ 与失真。簇熵是紧凑性的辅助度量。
为了健壮性，所有 k-means 聚类都涉及一百次随机初始化，并取平均结果。附录 A.3 提供了关于补充指标（如轮廓分数）的详细信息。

5.1 [RQ1] 整体概览：概念类别的对齐

我们首先调查 LLM 是否形成与人类判断对齐的概念类别。

主要发现：与人类类别的大致对齐

LLM 生成的聚类显著与人类定义的概念类别对齐，表明它们捕捉了人类概念组织的关键方面。值得注意的是，某些编码器模型表现出令人惊讶的强对齐，有时甚至优于规模大得多的模型，这突显了除纯粹规模之外，还有其他因素影响类似人类的类别抽象。

实验回顾：我们基准数据集的 LLM 词嵌入 [Rosch, 1973a, 1975, McCloskey and Glucksberg, 1978] 被聚类（k-means；K 匹配人类类别计数）。使用 AMI、NMI 和 ARI 测量与人类类别的对齐（AMI 显示在图 1 中；有关完整详细信息，请参阅附录 A.3、A.4）。

结果与观察：在所有测试的 LLM 中，推导出的概念簇与人类类别显著高于随机概率（图 1，显示平均 AMI 分数）。这表明它们的语义空间在宏观层面上编码了支持类人分组的信息。

值得注意的是，BERT 系列（尤其是 BERT-large-uncased ）表现出强大的对齐性，其效果通常与甚至超过许多更大的仅解码器模型。这表明，除了规模之外，架构或预训练因素也会影响类别人工智能结构的形成。

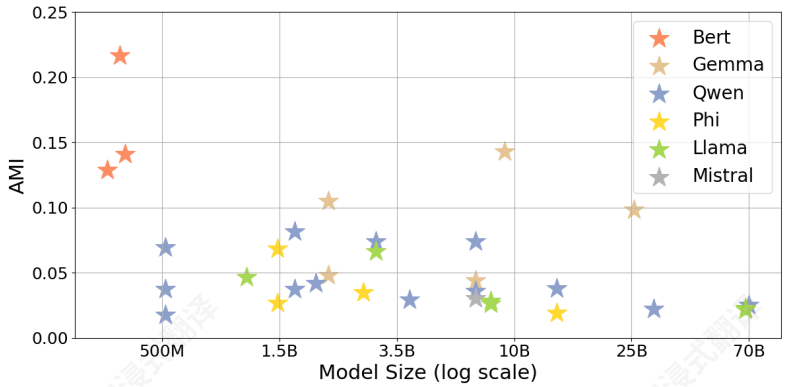


Figure 1: **LLM-derived Clusters Show Above-Chance Alignment with Human Conceptual Categories.** Adjusted Mutual Information (AMI) between human categories and LLM-embedding clusters versus model size. Results are averaged over three psychological datasets. All models perform significantly better than random clustering. BERT's performance is notably strong.

解释： 这些发现证实了 LLM 可以从它们的嵌入中恢复广泛的、类人的类别，验证了更深入的对比分析。这种宏观层面的协议要求检查这些类别的更细粒度的内部几何形状，我们将在下文中解决。

5.2 [RQ2] 深入挖掘：对细粒度语义的保真度

在建立了 LLM 广泛与人类概念类别对齐（第 5.1 节）之后，我们接下来研究一个更细致的问题：LLM 是否也捕捉了这些类别的内部语义结构，特别是类人的项目典型性?

关键发现：对语义细微之处的有限捕捉

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.

实验回顾：对于这个问题，如本节引言中详细所述，我们比较了来自认知科学数据集的人类典型性判断 [Rosch, 1973a, 1975, McCloskey 和 Glucksberg, 1978] 与基于 LLM 的度量。具体来说，我们计算了每个项目的标记嵌入与其 * 人类分配的类别名称 * （例如，‘robin ‘ 与 ‘bird ‘ ）的标记嵌入之间的余弦相似度。然后，这些项目到类别标签的相似度与人类评定的典型性分数相关联（ Spearman 的 ρ [Wissler, 1905]）。

结果与观察 : LLM 生成的项目到类别标签相似性与人类典型性判断之间的 Spearman 相关性在大多数模型和数据集上通常较为一般（附录 A.5 中的表 2 ； 图 6 ）。尽管某些相关性达到统计显著性 ($p < 0.05$)，但它们的幅度通常表明对应关系有限。这种模式表明，人类认为高度典型于某个类别的项目，并不总是被 LLM 表示为与该类别标签的嵌入显著更相似。虽然 BERT-large-uncased 偶尔表现出稍强的相关性，但这些仍然保持中等水平（表 2 ）。因此，没有测试的模型能稳健地使用此指标复制人类典型性梯度。附录 A.6 提供了支持这些观察的进一步可视化。

解释 : 这些发现表明，虽然 LLM 可以识别用于广泛分类的特征，但它们围绕显式类别标签组织的语义空间并不完全反映人类对语义的细微

人类典型性判断中可见的原型结构。在 LLM 中，一个项目嵌入与其类别标签嵌入的相似性驱动因素可能与支撑人类典型性的丰富、多方面的标准（例如，感知属性、功能角色）不同。LLM 可能相反地捕获与类别标签的更统计上均匀的关联，从而低估了人类概念的分级、原型中心特性。这种在捕获细粒度语义上的分歧导致了我们后续对整体信息处理效率的探究。

5.3 [RQ3] 效率角度：压缩 - 意义权衡

在探索了类别对齐（ RQ1 ）和内部语义结构（ RQ2 ）之后，我们现在处理我们的核心问题：在平衡信息压缩与语义意义保留时，LLM 和人类的表征策略在整体效率上如何比较? 我们的信息论框架直接探究这种权衡。

关键发现：不同的效率策略

与人类概念结构相比，LLM 在其概念表征中表现出明显更优的信息论效率。 通过我们的 $\mathcal{L}$ 目标进行评估，LLM 衍生的簇始终如一地实现了更 “ 最优 “ 的平衡（按此衡量）在表征复杂性（压缩）和语义失真之间。人类概念化虽然更丰富，但似乎在统计上不太紧凑，表明优化是为了超越纯粹的统计可压缩性压力。

实验回顾 : 如本节引言中详细所述，我们分析了人类定义的类别和 LLM 生成的聚类（来自不同 K 的 k-means ），并使用了两种主要的信息论度量：平均聚类熵 (S_α) [Giraldo et al., 2014, Wei et al., 2025] 以及我们的 $\mathcal{L}$ 目标函数（方程 4，带有 $\beta = 1$ ）。

结果与观察 : 一个数据集（ Rosch, 1975 ）的说明性结果如图 2 所示；所有数据集的趋势都一致（完整结果在附录 A.8 中）。

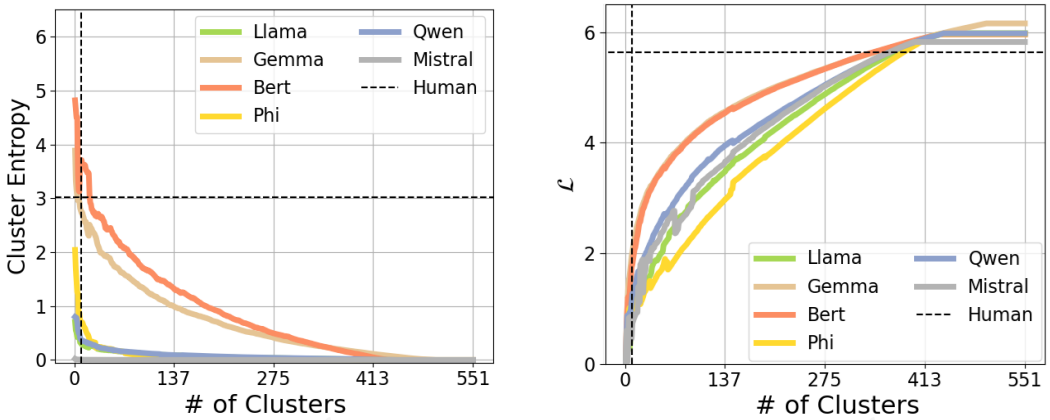
聚类熵洞察： 人类概念始终表现出比 LLM 生成的聚类更高的平均熵，即使在相似的 K 值下也是如此（图 2，左）。这表明，根据这个度量，人类类别在统计上 “ 不那么紧凑 ”，并且比 LLM 聚类包含更大的内部多样性。

信息论目标（ $\mathcal{L}$ ）洞察： 该 $\mathcal{L}$ 目标揭示了更明显的差异（图 2，右）。LLM 生成的聚类在大多数测试的 K 中始终比人类概念类别实现显著更低的 $\mathcal{L}$ 值。由于较低的 $\mathcal{L}$ 表示在我们的框架内最小化复杂性和失真之间的更优统计权衡，这意味着 LLM 根据这个特定的信息论基准更 “ 高效 ”。

解释 : 从熵和 $\mathcal{L}$ 目标中得出的综合结果表明，在表示策略上存在根本性差异。LLMs 表现出高度优化的统计紧凑性，通过最小化冗余和内部方差，实现了信息论上 “ 有效 ” 的表示。相比之下，人类的认知系统虽然在这些统计指标上 “ 次优 ”，但可能受到更广泛的功能需求的塑造。这些需求包括适应性强的一般化、丰富的因果和功能推理、神经嵌入的约束，以及细致沟通的要求 —— 这些压力可能更有利于统计上 “ 整洁 ” 度较低的表示，但最终更灵活和强大，能够应对复杂的世界。

6 讨论与结论

我们的信息论研究揭示了一个根本性分歧：LLMs 和人类在平衡信息压缩与语义意义时采用了截然不同的策略。虽然 LLMs 在与人类判断的广泛类别一致性方面取得了进展（ RQ1; 第 5.1 节），但在捕捉细粒度的语义细微差别方面存在不足（ RQ2; 第 5.2 节），并且关键在于，它们表现出截然不同的表示效率特征（ RQ3; 第 5.3 节）。这一模式强烈表明，LLMs 和人类正在优化不同的目标。



(a) **人类概念类别表现出更高的平均熵。** LLM 和人类类别的均值簇熵 (S_α) 与簇数 (K) （固定 K ）。更高的熵表示压缩程度更低。

(b)**LLM 实现了更优的 $\mathcal{L}$ 权衡。** 我们的信息论目标 ($\mathcal{L}$) 与 K。更低的 $\mathcal{L}$ 表示更优的统计压缩 - 意义平衡。

Figure 2: **LLMs Show More Statistically "Optimal" Compression Than Humans in Cluster Entropy and the $\mathcal{L}$ Measure.** (a) Mean cluster entropy as a function of K used for k-means clustering. (b) IB-RDT objective ($\mathcal{L}$) as a function of K used for k-means clustering. Human categories consistently show higher entropy and $\mathcal{L}$ values. Results shown for Rosch (1975) dataset; full results in Appendix A.8.

LLM 似乎在统计紧凑性方面进行了激进的优化。它们形成了信息论高效的表达，这从它们的较低簇熵和更 " 优 " 的 $\mathcal{L}$ 分数中得以证明。这表明它们最小化冗余并最大化统计规律性，这可能是它们在大量文本语料库上训练的结果。然而，这种对压缩的强烈关注限制了它们充分编码丰富、基于原型的语义细节的能力，而这些细节对于深度、类人理解至关重要。

人类认知优先考虑适应性丰富性、上下文灵活性和广泛的功能效用，即使这以我们的框架测量的统计紧凑性为代价。人类概念观察到的更高熵和 $\mathcal{L}$ 分数可能反映了针对更广泛复杂认知需求的优化。这些包括对稳健泛化的精细表示，支持强大的推理能力（因果、功能、目标导向），通过可学习和共享的结构实现有效沟通，以及将概念扎根于丰富的多模态体验中。大脑的神经架构本身可能天生倾向于分布式、上下文敏感和适应性表示，而不是静态最优压缩。因此，人类认知似乎在 " 投资 " 更好的适应性和多功能性，而我们的统计测量将其注册为低效。

像 BERT 这样的小型编码器模型在特定对齐任务中的出色表现（第 5.1 节）也表明，架构设计和预训练目标会显著影响模型抽象类人概念信息的能力。这一观察结果突出了未来人工智能发展的重要方向，即专注于增强人机对齐。

这些不同的表征策略具有重大意义。 **对于人工智能发展**，实现更类人的理解需要超越当前以规模化和统计模式匹配为中心的范式。未来的工作应探索明确促进更丰富、更细致概念结构的原理；我们的信息论框架和 $\mathcal{L}$ 目标（第 4 节）为引导和评估模型朝着这种更类人的平衡提供了潜在的工具类。 **对于认知科学**，具有独特优化偏好的 LLM 们是宝贵的计算挡板。将它们的操作策略与人类表现进行比较，可以揭示塑造人类概念形成的独特约束和多面目标，为认知理论提供了一个强大的试验场。

本质上，LLM 擅长统计可压缩性，其表征路径与人类认知根本不同，后者推崇适应性丰富性和功能性效用，往往高于纯粹的统计效率。这一核心差异至关重要：它突出了人工智能追求类人理解的当前局限性，并为未来研究指明了重要方向。推进人工智能

“从 token 到思想 “, 朝着真正能够理解和推理的系统发展，将需要拥抱培养这种更丰富、具有上下文感知能力的概念结构的原理。我们的框架在这一方向上提供了一定量的发展，鼓励进一步探索明显的 “ 低效 ” 实际上可能是强大、类人智能的标志的途径。

参考文献

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, 等 . Phi-3 技术报告：在您的手机上本地运行的强大语言模型。 *arXiv* 预印本 *arXiv:2404.14219*, 2024。

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, 等 . Phi-4-mini 技术报告：通过混合 LoRA 实现紧凑而强大的多模态语言模型。 *arXiv* 预印本 *arXiv:2503.01743*, 2025。

金则贝 , 白帅 , 崔云飞 , 崔泽宇 , 邓凯 , 邓晓东 , 范杨 , 葛文斌 , 韩宇 , 黄飞 , 等 . Qwen 技术报告 . *arXivpreprint arXiv:2309.16609*, 2023.

Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, ArtyomKozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R Costa-jussà, David Dale, 等 . 大概念模型：句法表示空间中的语言建模 . *arXiv preprintarXiv:2412.08821*, 2024.

威廉 · 克罗夫特 . 激进构式语法：类型学视角下的句法理论 . 牛津大学出版社 , 美国 , 2001.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, 和 Kristina Toutanova. BERT：用于语言理解的深度双向 Transformer 的预训练 . 在 Jill Burstein, Christy Doran, 和 Thamar Solorio 编的 *2019* 年北美计算语言学协会分会会议论文集：人机语言技术 , 第 *1* 卷（长篇和短篇论文） , 第 4171–4186 页 , 明尼苏达州明尼阿波利斯 , 2019 年 6 月 . 计算语言学协会 . doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Luis Gonzalo Sanchez Giraldo, Murali Rao, 和 Jose C Principe. 使用无限可分核从数据中测量熵 . *IEEE Transactions on InformationTheory*, 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, 等 . Llama 3 模型群 . *arXiv preprint arXiv:2407.21783*, 2024.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, 和 Weizhu Chen. Deberta: 解码增强型 BERT 与解耦注意力 . *arXivpreprint arXiv:2006.03654*, 2020.

Nhat Hoang-Xuan, Minh Vu, 和 My T Thai. LLM 辅助概念发现：自动识别和解释神经元功能 . *arXivpreprint arXiv:2406.08572*, 2024.

Nathaniel Imel 和 Noga Zaslavsky. 人类概念学习的最优压缩 . 在 认知科学学会年度会议论文集 , 卷 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, 等 . Phi-2: 小型语言模型的惊人力量 . *Microsoft Research* 博客 ,1(3):3,2023.

Siddharth Karamcheti, Laurel Orr, Jason Bolton, Tianyi Zhang, Karan Goel, Avanika Narayan, Rishi Bommasani, Deepak Narayanan, Tatsunori Hashimoto, Dan Jurafsky, 等 . Mistral—— 通往可复现语言模型训练的旅程 , 2021.

Yuxiao Li, Eric J Michaud, David D Baek, Joshua Engels, Xiaoqing Sun, 和 Max Tegmark. 概念的几何学：稀疏自动编码器特征结构 . *arXiv* 预印本 *arXiv:2410.19750*, 2024.

Akihiro Maeda、 Takuma Torii 和 Shohei Hidaka。将共现矩阵分解为可解释的组件作为形式概念。在 计算语言学协会发现 *ACL 2024*，第 4683–4700 页，2024 年。

Michael E McCloskey 和 Sam Glucksberg。自然类别：定义良好还是模糊集？ 记忆与认知，第 6 卷第 4 期：462–472，1978 年。

Gregory Murphy。 概念大书。麻省理工学院出版社，2004 年。

Kiho Park、 Yo Joong Choe、 Yibo Jiang 和 Victor Veitch。大型语言模型中类别和层次概念几何。 *arXiv* 预印本 *arXiv:2406.01506*，2024 年。

E Rosch。感知和语义类别的内部结构。认知发展与语言习得 / 纽约：学术出版社，1973a。

Eleanor Rosch。原型理论。 认知发展与语言习得，第 111–144 页，1973b 年。

Eleanor Rosch. 语义类别的认知表征 . 实验心理学杂志：综合 , 104(3):192, 1975.

Eleanor Rosch, Carol Simpson, 和 R Scott Miller. 典型性效应的结构基础 . 实验心理学杂志：人类感知与表现 , 2(4):491, 1976.

Eleanor H Rosch. 自然类别 . 认知心理学 , 4(3):328–350, 1973c.

Chen Shani, Jilles Vreeken, 和 Dafna Shahaf. 面向概念感知的大型语言模型 . 在计算语言学协会发现：*EMNLP 2023*, 页面 13158–13170, 2023.

Claude Elwood Shannon. 通信的数学理论 . 贝尔系统技术杂志 ,27(3):379–423, 1948.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, 和 Jianfeng Gao. 大型语言模型时代重新思考可解释性 . *arXiv* 预印本 *arXiv:2402.01761*, 2024.

Ben Sorscher、 Surya Ganguli 和 Haim Sompolinsky。神经表征几何学是少样本概念学习的基础。美国国家科学院院报，119(43)：e2200800119，2022 年。

Gemma 团队、 Thomas Mesnard、 Cassidy Hardin、 Robert Dadashi、 Surya Bhupatiraju、 Shreya Pathak、 Laurent Sifre、 Morgane Rivière、 Mihir Sanjay Kale、 Juliette Love 等。Gemma：基于 Gemini 研究和技术的开源模型。*arXiv* 预印本 *arXiv:2403.08295*，2024 年。

Gemma 团队、 Aishwarya Kamath、 Johan Ferret、 Shreya Pathak、 Nino Vieillard、 Ramona Merhej、 Sarah Perrin、 Tatiana Matejovicova、 Alexandre Ramé、 Morgane Rivière 等。Gemma 3 技术报告。*arXiv* 预印本 *arXiv:2503.19786*，2025 年。

Naftali Tishby、 Fernando C Pereira 和 William Bialek。信息瓶颈方法。*arXivpreprintphysics/0004057*，2000 年。

Hugo Touvron、 Thibaut Lavril、 Gautier Izacard、 Xavier Martinet、 Marie-Anne Lachaux、 Timothée Lacroix、 Baptiste Rozière、 Naman Goyal、 Eric Hambro、 Faisal Azhar 等。Llama：开放且高效的基础语言模型。*arXivpreprint arXiv:2302.13971*，2023a。

Hugo Touvron、 Louis Martin、 Kevin Stone、 Peter Albert、 Amjad Almahairi、 Yasmine Babaei、 Nikolay Bashlykov、 Soumya Batra、 Prajwal Bhargava、 Shruti Bhosale 等。Llama 2：开放基础和微调聊天模型。*arXivpreprint arXiv:2307.09288*，2023b。

Mycal Tucker，Julie Shah，Roger Levy 和 Noga Zaslavsky。通过效用、信息性和复杂性实现类人涌现通信。*Open Mind*，9:418–451，2025。

Amos Tversky。相似性的特征。*Psychological review*，84(4):327，1977。

兰伟 , 东王 , 和 王宇 . 广义相对熵 : 对 Rényi 熵的新视角及其从复杂度度量到稀疏度度量的探索 , 应用于机器状态监测 . *Mechanical Systems and Signal Processing*, 223:111917, 2025.

克拉克 · 威斯勒 . 斯皮尔曼相关系数公式 . *Science*, 22(558):309–311, 1905.

J· 杰拉德 · 沃尔夫 . 信息压缩作为人类学习、感知和认知的统一原则 . *Complexity*, 2019(1):1879746, 2019.

安阳 , 杨宝松 , 张贝辰 , 胡彬源 , 郑波 , 余 Bowen, 李成元 , 刘代恒 , 黄飞 , 魏浩然 , 等 . Qwen2.5 技术报告 . *arXiv preprintarXiv:2412.15115*, 2024.

刘庄 , 林伟 , 石亚 , 和 赵军 . 一种具有后训练的鲁棒优化 BERT 预训练方法 . 在沈立 , 孙茂松 , 刘杨 , 吴华 , 刘康 , 车万祥 , 何世珠 , 和 饶高奇 , 编者 , 第 *20* 届中国计算语言学会议论文集 , 第 1218–1227 页 , 呼和浩特 , 中国 , 2021 年 8 月 . 中国信息处理学会 . URL https://aclanthology.org/2021.ccl-1.108/.

A Limitations

尽管这项研究提供了有价值的见解，但仍应考虑一些局限性。

- 我们的分析主要关注英语；跨不同结构的语言的可推广性仍是一个开放性问题。
- 人类分类数据作为基准可能无法完全捕捉认知复杂性，并可能引入偏差。
- 我们的 IB-RDT 目标应用于特定的 LLM ； 其他模型或表示可能表现不同。
- 我们关注静态、无上下文表示。LLM 可能无法捕捉上下文敏感性，因为人类概念受超出原始压缩效率的因素影响（经验、社会互动、文化背景）。
- 我们的分析仅限于文本输入，并未探索基于图像的表示形式。

未来的工作可以通过扩展到其他语言、探索替代的认知模型、动态表示，以及在不同的架构或实际应用中测试这些原则来解决这些问题。

A.1 数据集访问详情

Rosch [1973a 、 1975],McCloskey 和 Glucksberg [1978] 的聚合和数字化的人类分类数据集以 CSV 格式提供，链接已缩短以保护匿名性：[Link reduced for anonymity]。

A.2 LLM 详情

- **BERT 家族：** deberta-large 、 bert-large-uncased 、 roberta-large [Devlin 等人，2019 年，He 等人，2020 年，Zhuang 等人，2021]。
- **QWEN 家族：** qwen2-0.5b, qwen2.5-0.5b, qwen1.5-0.5b, qwen2.5-1.5b, qwen5-1.5b, qwen1.5-1.5b, qwen1.5-4b, qwen2.5-4b, qwen2-7b,qwen1.5-14b, qwen1.5-32b, qwen1.5-72b [Bai 等人，2023，Yang 等人 , 2024].
- **Llama 家族：** llama-3.2-1b, llama-3.1-8b, llama-3-8b, llama-3-70b, llama-3.1-70b [Tou- vron 等人，2023a,b，Grattafiori 等人， 2024].
- **Phi 家族：** phi-1.5, phi-1, phi-2, phi-4 [Javaheripi 等人，2023 年，Abdin 等人，2024 年，Aboue- lenin 等人，2025].
- **Gemma 家族：** gemma-2b, gemma-2-2b, gemma-7b, gemma-2-9b, gemma-2-27b [团队等人，2024, 2025].
- **Mistral 家族：** mistral-7b-v0.3 [Karamcheti 等人， 2021].

A.3 额外聚类指标

为了进一步验证我们的聚类对齐结果（第 5.1 节），除了调整后的互信息（ AMI ）和归一化互信息（ NMI ）之外，我们还计算了 k-means 聚类从 LLM 嵌入到人类定义的类别中的调整兰德指数（ ARI ）。ARI 测量两个数据聚类之间的相似性，并校正偶然性。像 AMI 一样，得分为 1 表示完全一致，得分为 0 表示偶然一致。

在所有测试的 LLM 中，ARI 和 NMI 分数在很大程度上反映了 AMI 所显示的趋势，显示出显著高于偶然性的与人类类别的一致性，以及相似的相对模型性能。轮廓分数虽然更具变异性，但通常表明 LLM 衍生和人类类别的合理聚类凝聚力。这些分数的详细表格提供如下。

这些补充指标加强了 LLM 捕获广泛的人类概念性分组的结论。

A.4 按模型和数据集的详细 AMI 分数

表 1 提供了每个 LLM 在三个单独心理学数据集上 AMI 分数的更细粒度视图。

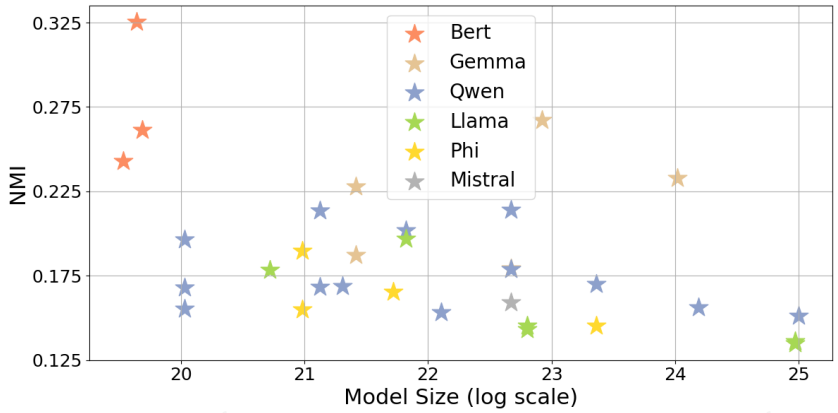


Figure 3: **LLM-derived Clusters Show Above-Chance Alignment with Human Conceptual Categories.** Normalized Mutual Information (NMI) between human-defined categories and clusters from LLM embeddings. Results are averaged over three psychological datasets. All models perform significantly better than random clustering. BERT's performance is notably strong.

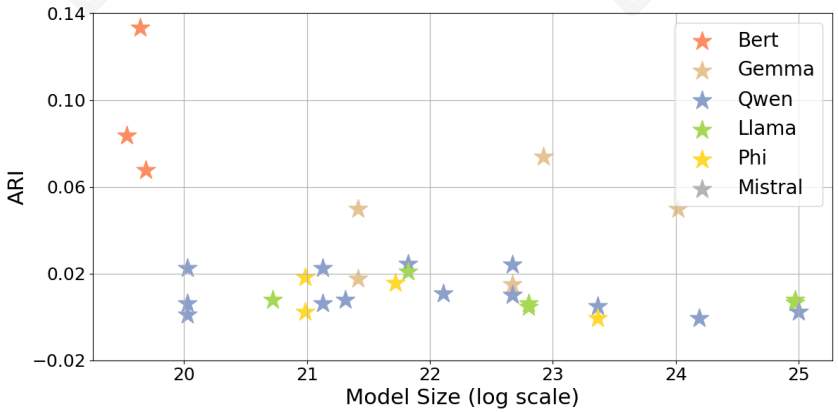


Figure 4: **LLM-derived Clusters Show Above-Chance Alignment with Human Conceptual Categories.** Adjusted Rand Index (ARI) between human-defined categories and clusters from LLM embeddings. Results are averaged over three psychological datasets. All models perform significantly better than random clustering. BERT's performance is notably strong.

数据集	模型	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/robert二个大的	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 和 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 1: 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.5 人类典型性判断与 LLM 内部聚类几何形状的相关性

A.6 典型性与余弦相似度 [RQ2]

图 5 显示了代表性散点图，说明了人类典型性分数（或心理距离）与选定类别和模型的 LLM 导出的项目质心余弦相似度之间的关系。这些图直观地展示了第 5.2 节中讨论的通常微弱的关联。

图 6 显示了跨模型系列和数据集的 Spearman 相关性的汇总。这些相关性非常弱且大多不显著。

A.7 理论极端情况探索 $\mathcal{L}$

(内容来自您原始附录部分 A: “ 理论极端情况探索 ”，确保它引用了方程 4 中定义的 $\mathcal{L}$)。

在 $|C| = |X|$ 的情况下（每个数据点是一个大小为 1 的簇，所以 $|C_c| = 1 \; \forall c \in C$ ），则 $H(X|C) = \frac{1}{|X|} \sum_{c \in C} 1 \cdot \log_2 1 = 0$。每个簇的失真项 $\sigma_c^2 = 0$，因为项目是其自己的质心。因此 ,$\mathcal{L} = I(X;C) + \beta \cdot 0 = H(X) - H(X|C) = H(X) = \log_2 |X|$。这代表了通过聚类完美编码每个项目而不进行任何压缩的成本，以及零失真。

在 $|C| = 1$ (一个簇 C_X 包含所有 $|X|$ 数据点，所以 $|C_{C_X}| = |X|$) 的情况下，则 $H(X|C) = \frac{1}{|X|} |X| \log_2 |X| = \log_2 |X|$。因此，$I(X;C) = H(X) - H(X|C) = \log_2 |X| - \log_2 |X| = 0$。这表示最大压缩（所有项目被视为一个）。失真项变为 $\beta \cdot \frac{1}{|X|} |X| \cdot \sigma_X^2 = \beta \cdot \sigma_X^2$，其中 σ_X^2 是所有项目 X 相对于 X 的全局质心的方差。所以，$\mathcal{L} = 0 + \beta \cdot \sigma_X^2 = \beta \cdot \sigma_X^2$。这表示最大压缩的场景，其中成本仅仅是通过单个原型表示所有项目所造成的失真。

A.8 压缩图表

图 7 显示了不同 LLM 系列中，平均簇熵 (S_α) 与簇数量 (K) 的关系，并与人类定义的类别（表示为数据集固定 K 值处的不同点或线）进行比较。更高的熵值表示压缩程度较低或聚类更多样化。

图 8 描绘了 IB-RDT 目标 ($\mathcal{L}$) 与 K 的对比。较低的 $\mathcal{L}$ 表示在压缩 ($I(X;C)$) 和语义保真度（失真）之间达到了更优的平衡。人类类别（固定的 K ）显示出更高的 $\mathcal{L}$ 值。

模型	数据集相关性（Spearman ρ）		
	Rosch (1973)	Rosch (1975)	McCloskey (1978)
Qwen1.5-72B	-0.237	-0.049	-0.016
Llama-3-70B	-0.124**	-0.085	0.016
Llama-3.1-70B	-0.125**	-0.084	0.015
Qwen1.5-32B	-0.051	-0.064**	0.007
gemma-2-27b	-0.166	-0.116	0.038
Qwen1.5-14B	-0.197	-0.052	-0.029
phi-4	0.061	-0.044	0.025
gemma-2-9b	-0.282	-0.074	0.117
Llama-3.1-8B	-0.184	-0.075	-0.058
Llama-3-8B	-0.162	-0.073	-0.053
Mistral-7B-v0.3	0.015	-0.112	0.040
Qwen2-7B	-0.021	-0.105	-0.008
Qwen2.5-7B	0.033	-0.066	-0.030
gemma-7b	-0.135	-0.047**	0.010
Llama-3.2-3B	-0.007	0.000	0.001
phi-2	0.049	-0.108**	-0.001
gemma-2b	-0.176	-0.055	0.052
gemma-2-2b	-0.283	-0.107	0.117
Qwen1.5-1.8B	-0.106	-0.085	0.021
Qwen2.5-1.5B	-0.003	-0.035	0.015
phi-1.5	0.134	-0.134	0.007
phi-1	0.219	-0.138	0.013
Llama-3.2-1B	-0.062	-0.004**	-0.003
Qwen1.5-0.5B	-0.122	-0.004	-0.001
Qwen2-0.5B	-0.044	0.009	-0.009
Qwen2.5-0.5B	-0.018	-0.009	-0.007
roberta-large	0.088	-0.047	-0.074
bert-large-uncased	-0.427	-0.198**	0.206**
deberta-large	0.016	-0.042	-0.023

Table 2: **Correlation between Human Typicality Judgments and LLM Internal Cluster Geometry.** Spearman 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). Negative correlations suggest alignment. $^{**}p < 0.05$.

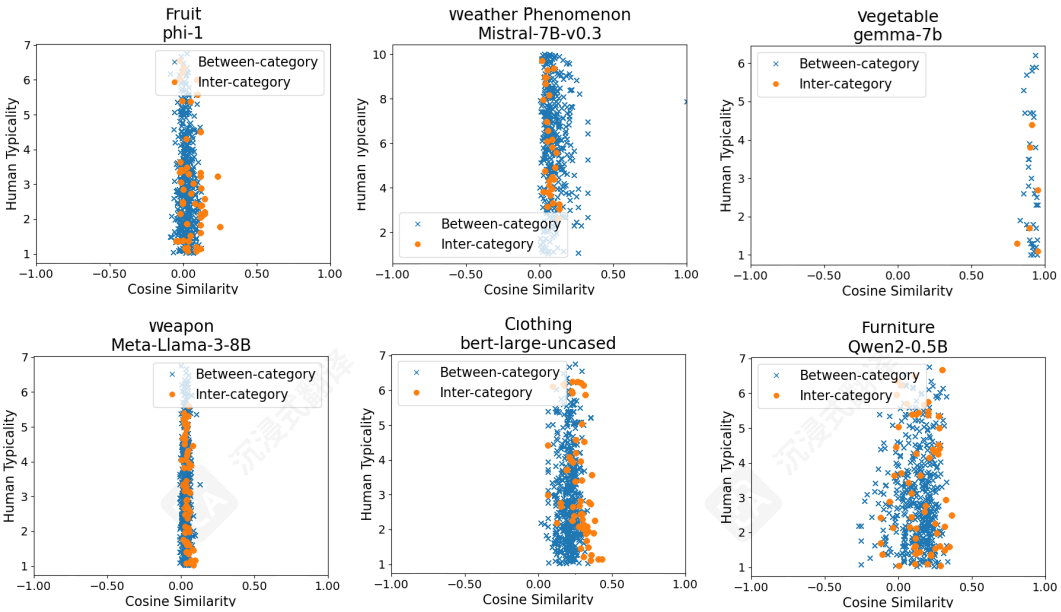


图 5：**LLM 嵌入距离与人类典型性判断之间弱相关或无相关性。** 散点图示例，展示了属于比较类别的物品与属于其他类别的物品在余弦相似度与人类典型性方面的关系。

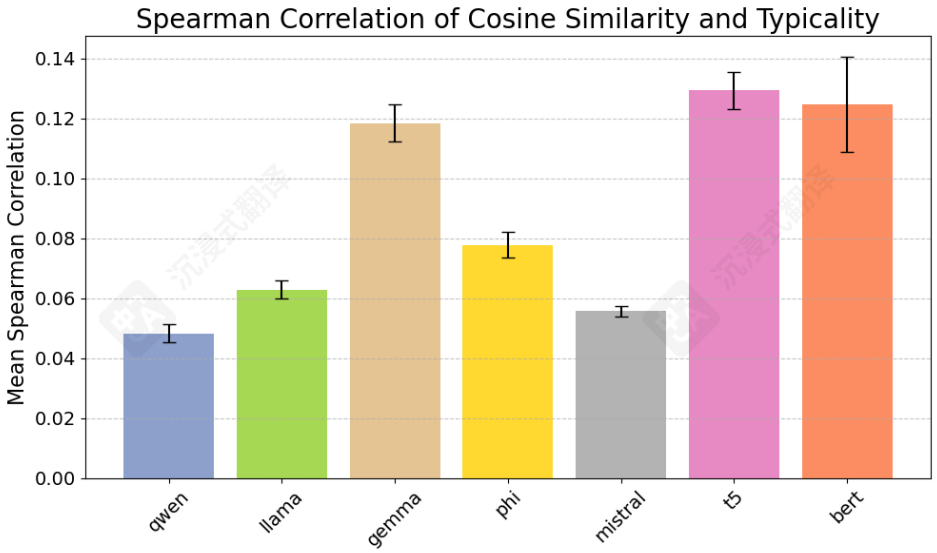


Figure 6: **Weak and Mostly Non-Significant Spearman Correlation Values Between Human Typicality Judgments and LLM Cosine Similarity Indicating Different Structure Representing Concepts.** Mean Spearman correlation values across the models belonging to the same family and across the three datasets.

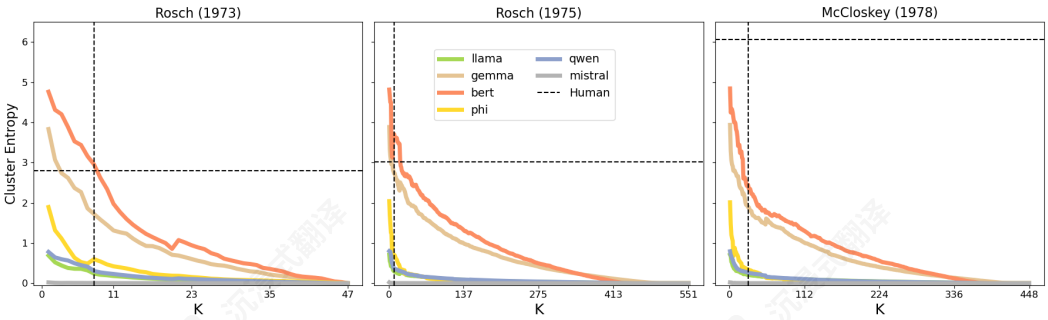


Figure 7: **Human Conceptual Categories Exhibit Higher Mean Entropy than LLM-Derived Clusters.** Mean cluster entropy (S_α) versus the number of clusters (K) for various LLMs, compared against human-defined categories (represented as distinct points or lines at their fixed K values from the datasets). Higher entropy values indicate less compressed or more diverse clusterings.

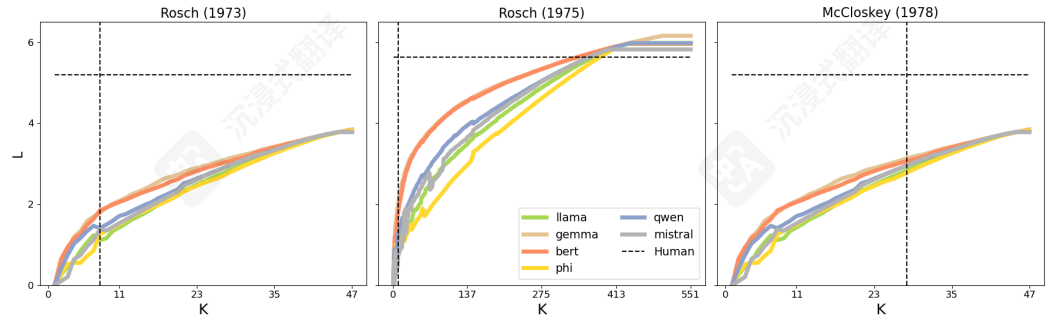


Figure 8: **LLMs Achieve a More "Optimal" Compression-Meaning Trade-off by the $\mathcal{L}$ Measure.** IB-RDT objective ($\mathcal{L}$) vs. K. Lower $\mathcal{L}$ indicates a more optimal balance between compression ($I(X;C)$) and semantic fidelity (distortion). Human categories (fixed K) show higher $\mathcal{L}$ values.