# 大型语言模型嵌入中的语义结构

**Austin C. Kozlowski**
University of Chicago

**Callin Dai**
University of Chicago

**Andrei Boutyline**
University of Michigan

## Abstract

心理学研究一致发现，人类对单词在不同语义尺度上的评分可以简化为低维形式，而且信息损失相对较小。我们发现，大型语言模型（LLM）的嵌入矩阵中编码的语义关联表现出类似的结构。我们展示了由反义词对（例如，kind - cruel）定义的语义方向上的单词投影与人工评分高度相关，并进一步发现这些投影实际上可以有效地简化为 LLM 嵌入中的三维子空间，与人类调查响应中得出的模式紧密相似。此外，我们发现沿一个语义方向移动标记会对几何对齐特征造成目标外的影响，这些影响与其余弦相似度成正比。这些发现表明，语义特征在 LLM 中具有如同在人类语言中互相关联的纠缠关系，尽管语义信息表面上看似复杂，但实际上是惊人的低维。此外，考虑到这种语义结构，可能是避免在引导特征时产生意外后果的关键。

## 1 介绍

大型语言模型（LLMs）展示出模仿人类语言行为的非凡能力，但这些模型的内部表征在多大程度上与人类认知模型相似仍不清楚 [1] 。一方面，LLMs 有可能会建立与人类产生这些训练数据时相似的内部表征，因为它们在接受了大量关于思维、行为和互动的记录的训练。然而，另一方面，LLMs 使用的架构与人脑不同，它们的训练数据在质上不同于人类在发展过程中接收到的刺激，而它们的“下一个词预测”任务可能与人类学习的目标根本不同。改善我们对 LLMs 内部意义表征的理解不仅具有科学价值，还有可能对与模型安全、审计和控制相关的实际应用有价值 [2, 3] 。

在本文中，我们从社会心理学中的一个长期发现入手，探讨其与 LLM 内部表征的相关性。具体而言，大量文献发现，人类在表面上看似不同的语义尺度上的评分往往遵循一个强烈且系统的相关结构；例如，被认为“柔软”的东西往往也被标记为“善良”，“强壮”的东西往往是“大”。因此，对于一组广泛的语义属性的评分可以有效地简化为一个三维解，信息损失相对较小 [4, 5] 。诸如语义差异研究等传统识别这些三种潜在维度为评价（好与坏），效力（强与弱）和活动性（运动与静止），而其他关于语义评分的研究也发现了类似的潜在因素，如温暖与能力 [6] 或作为价度、唤起和支配 [7] 。

利用单词嵌入模型开发的技术并成功扩展到 LLM [8, 9] ，我们从 LLM 嵌入矩阵中提取特征方向，这些方向对应于 28 个关键语义轴（例如，善良-残忍、愚蠢-聪明）。我们将单词（标记）的向量投影到这些特征向量上，并显示这些投影与人类对这些单词的评级在相应语义尺度上高度相关。在确认了标记投影与语义关联之间的对应关系后，我们对投影应用主成分分析（PCA），发现三维解决方案保留了 28 个原始特征之间 40 到 55 % 的方差，并且这些主成分的负载暗示了一个与先前研究中人类受试者所识别的评价、效能和活动维度类似的结构。

在识别出这个语义结构后，我们考虑特征对齐对于模型行为和引导的影响。具体来说，我们假设对一个特征进行干预可能会在其他特征上造成可预测的非目标效应，这种效应与它们的余弦相似度成比例。例如，如果软硬与善恶紧密对齐，我们预计对软硬的干预在这一方向上的非目标效应会比在与其正交的特征（如愚昧明智）上的非目标效应更强。为了验证这一假设，我们提示大型语言模型对一组单词进行语义联想报告，就像被试者在心理问卷中做的一样。收集关于这组单词的 LLM 语义联想的基线数据后，我们对模型的词元嵌入进行干预，引导相应的词向量朝某一语义特征的方向，然后测量对所有其他语义特征报告的联想的影响。我们的结果支持这一假设，即非目标效应的大小与目标和非目标特征向量之间的余弦相似度成比例。

我们的研究做出了几个贡献：

1. 我们发现，LLM 嵌入中的语义特征表现出强烈且有意义的相关结构。在 $n$ 维空间中可以表示的“几乎正交”向量的数量相对于 $n$ [10] 是指数级的。这意味着具有 $n > 1000$ 的 LLM 嵌入能够表示大量近似正交的特征。尽管存在这种理论可能性，我们发现对于表达文化关联和意义的许多关键特征之间存在实质性的相关性。
2. 这种特征结构紧密反映了人类语义评级中常见的模式，表明特征的相关性不是嵌入表示的产物，而是人类语言和理解的一个有意义的特征。
3. 最后，我们证明了这种结构对隔离和引导单个特性具有重要意义。先前的研究已经确定了这种“脱靶效应”，但通常将其视为随机的 [11, 9] 。我们表明，特性之间的相关结构可以预测这些脱靶效应，这一见解有助于预见和减轻特性引导的意外后果。

## 2 背景

### 2.1 大型语言模型嵌入

自回归解码器仅由 LLM 组成，其结构为在起始位置为嵌入矩阵，在末端为去嵌入矩阵而构成的变换器块的堆栈 [12, 13, 14, 15, 16] 。模型开始处的嵌入矩阵起到将离散的符号映射到分布式表示的向量的角色。这些向量随后被传递通过变换器堆栈，之后再由去嵌入矩阵映射回符号。

虽然大多数现有的大型语言模型可解释性研究集中在模型的变换器模块 [17, 18, 19, 20] 中的激活模式，但嵌入和反嵌入矩阵同样值得关注。首先，因为嵌入矩阵中的向量直接与单个标记关联，这些空间中的方向比模型的变换器模块中的向量更容易与可解释的概念相联系，因为变换器模块中的向量与特定标记的联系较远。其次，嵌入矩阵由权重组成，而不是激活。这意味着在这些矩阵中识别出的语义关系并不依赖于特定的输入文本，而是影响到所有的推理过程。

### 2.2 特征纠缠与叠加

深度神经网络编码了多种“特征”或概念表示，但这些特征很少对应于单个神经元。相反，特征往往以方向或区域的形式出现在模型的潜在空间中，这个空间是通过多个神经元的激活组合而成，这些神经元在其他组合中用来代表不同的特征 [17, 21] 。最近被称为“叠加”的这种现象被假设是因为当 $n$ 足够大时，$n$ 维空间可以代表 $\gg n$ 个几乎正交的向量 [22] 。因此，代表更多特征的好处通常超过了由于略微非正交的特征向量之间的最小干扰所产生的代价。因此，一个与“汽车”相关的神经元可能也与“猫”高度相关，但最终被激活的特征取决于这个神经元如何与其他神经元组合激活 [23] 。

叠加通常被描述为一种新兴的方法，用于将比维度更多的特征打包到模型中，其结果可能是随机特征集最终会处于叠加状态 [22] 。然而，关于分布式表示的长篇文献表明，特征可以在有意义的情况下是非正交的，而且基于它们的语义关系，特征之间可能会出现实质性的对齐 [24, 25, 26] 。事实上，在他们关于分布式表征的奠基性工作中，Hinton、McClelland 和 Rumelhart [27] 认为，使用正交表示来表示有意义相关的特征将消除“分布式表示最有趣的性质之一：它们自动引发概括”（第 82 页） [27] 。

从这个角度来看，分布式表示就像奇异值分解（SVD）等降维技术一样，通过将 $k$ 观测特征表示为较少数量的 $n$ 潜在特征的线性组合，尽可能减少信息的失真或丢失。对于浅层神经网络的分析，如 word2vec，这种解释尤其有成效，其中表示的基方向不对应于可解释的特征，但具有相似意义的单词聚集在一起，并且空间中的方向对应于连续特征，如好坏、男性女性或富裕贫穷 [24, 29, 8, 30, 31] 。虽然现在普遍认为，像 word2vec 这样的浅层神经网络的表示通过特征的几何对齐来编码人类的语义结构，但这种观点尚未在 LLM 的特征表示研究中占据主导地位。

然而，如果语义特征在大语言模型（LLM）中具有有意义的非正交性，这可能会使“特征引导”的努力变得复杂，即用户通过在模型内部隔离和修改特征来放大或抑制特征的过程。最近的努力通过特征引导来控制 LLM 输出并减少不必要行为，的确报告了当目标特征被修改时对其他特征的意外“偏离目标”影响 [11, 9] 。然而，这些文献将偏离目标影响主要视为特征随机叠加的结果，并且几乎没有努力去识别这些影响的系统模式。先前的研究试图通过减少特征之间的对齐来最小化偏离目标影响。比如，Park 等 [9] 通过对白解向量施加白化操作，使得对一个特征的干预对其他特征造成的偏离目标影响最小化。尽管最小化特征相关

性确实减少了偏离目标影响，我们假设如果特征在语义上有意义地非正交且其对齐反映了真实的语义关系，那么强制正交可能会扭曲底层概念的表示。也就是说，如果模型潜在空间中特征的对齐编码了语义信息，那么非正交性可能是分布式表示中的一个“特性”而非“缺陷”。

### 2.3 文化情感子空间

概念的纠缠一直是心理学中一个长期受关注的话题。20 世纪 50 年代发展了语义差异法来测量认知联想，其中受试者在一组语义尺度上对词列表进行评分，这些尺度包括温暖-寒冷、柔软-坚硬和丑陋-美丽等 [4] 。使用这种方法的学者得出了两个关键发现。首先，他们发现对数百个项目的人的评分在数十个尺度上可以一致地减少到一个三维子空间，而信息的损失惊人地少。其次，他们发现相同的潜在因素大致对应于评估（好与坏）、效能（强与弱）和活动性（主动与被动），这一结构在广泛的国家和文化群体的分析中出现，这表明三因素结构可能是人类语义的共性 [32, 5] 。

这些发现表明，我们用于理解、分类和区分世界上物体的许多属性可以在一个相对低维的子空间中被大致捕获。此外，这些发现推动了对分布式语言表示的文化和心理学解释。如果我们在日常生活中用于做出区分和决策的许多关键属性可以很好地描述为少数潜在属性的线性组合，那么这些特征在语言模型的表示空间中可能会纠缠在一起，这并不是因为偶然的叠加，而是语言和思维中这些属性自然相关的结果。

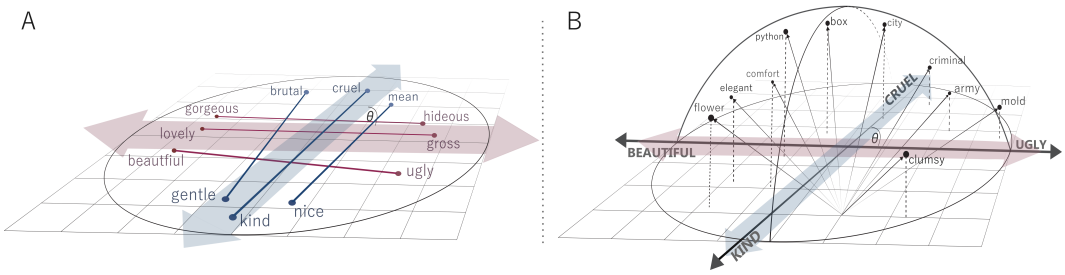


Figure 1: 图板（A）：通过计算反义词对之间的差异并取其平均值来提取语义特征向量。(B)：通过计算单个词元在嵌入空间中对语义特征向量的归一化线性投影来估算其与语义特征的关联。

## 3 数据与方法

我们的分析依赖于三种形式的数据：（i）通过调查收集的基于语义尺度的人类对单词的评分，（ii）相同单词和语义尺度集合在 LLM 嵌入空间中的投影，以及（iii）类似于调查任务的任务中 LLM 的下一个标记概率。我们对经过指令调优的 Gemma-3 系列模型（1B、12B 和 27B）进行了测试。附录中的结果显示，Llama-3 和 Qwen-2.5 模型在参数大小高达 72B 时也有类似的模式（尽管关系稍弱）。实验是在单个 H100 GPU 上运行的。

### 3.1 调查数据

我们使用了 Boutyline 和 Johnston 收集的关于语义关联的最近调查数据 [33] 。该调查通过 Prolific 平台对一份由 1750 名受访者组成的在线配额样本进行调查。该样本在主要社会人口统计类别上大致代表了美国人口。调查要求受访者从一组语义尺度上对 360 个词进行评分。为了减少受访者的疲劳，每个受访者仅被展示了全部词/尺度对中的一个子集；最终每个词平均被 24 名受访者在每个尺度上进行了评分。在我们的分析中，我们剔除了 59 个在我们的 LLM 嵌入空间中不对应于单个标记的词。我们还排除了原始 40 个尺度中的 12 个尺度，因为它们与其他尺度过于相似（例如，人机与自然-人工）、或因为方差过高导致难以进行稳健估计（例如，自由派-保守派）、或因为同义词-反义词对在嵌入空间中的数量不足以进行特征识别（例如，值得的-不值得的）。最终得到的分析样本是对 28 个不同语义尺度进行评分的 301 个词。

### 3.2 测量特征方向

为了从大型语言模型（LLM）的嵌入中提取特征方向，我们应用了一种最初为词嵌入模型开发的方法 [30, 8]，此方法已成功扩展到 LLM [19, 20]。该方法形式化地表示了一种直觉，即特征可以被识别为嵌入空间中反义词之间的方向，并且语义上相似的反义词对共享相似的方向向量（例如，good-bad 与 great-awful 具有很高的相似性）。具体来说，给定一个感兴趣特征的反义词对集合，我们计算每个反义词对的嵌入向量之间的归一化差异，然后取这些归一化差异的平均值，如方程 (1) 所示：

$$\boldsymbol{d}_f = \frac{1}{N}\sum_{j=1}^{N}\frac{\boldsymbol{ant}_j^+ - \boldsymbol{ant}_j^-}{\|\boldsymbol{ant}_j^+ - \boldsymbol{ant}_j^-\|} \tag{1}$$

，其中 $\boldsymbol{d}_f$ 是语义特征 $f$ 的向量，$N$ 是反义词对的总数（我们为每个特征使用 10 对），$\boldsymbol{ant}_j^+$ 和 $\boldsymbol{ant}_j^-$ 是与 $j$ -反义词对中的正、负反义词对应的嵌入词向量。我们将分析限制在能够完全用单个标记表示的词。

我们在嵌入空间中量化语义关联为目标标记与相应特征向量之间的余弦相似度，这等同于两个归一化向量的线性投影。我们构建了一个以这种方式计算的投影数据集，反映上述调查数据。具体而言，我们构建了 28 个对应于调查中测量的语义尺度的特征向量（例如，善良-残酷，愚蠢-明智等）。然后，对于调查中评价的 301 个词，我们从嵌入矩阵中取出相关的标记向量，并计算它与 28 个特征向量中的每一个的余弦相似度。结果数据集提供了 301 个标记在 28 个特征上的归一化线性投影，并且可以直接与调查数据集进行比较。

作为一个初步步骤，我们计算词向量投影到提取的语义特征向量上与人类对这些词在相应语义尺度上的评分之间的相关性。这些相关性用于验证我们从大语言模型嵌入中推导语义关联的方法。我们将我们的方法与 Park 等人提出的替代方法进行比较。[9]。他们在识别特征方向和进行干预以隔离“因果可分概念”并最小化非目标效应之前，对模型的去嵌入空间应用白化操作。具体而言，他们通过去嵌入空间协方差矩阵倒数的平方根（$\boldsymbol{\Sigma}^{-1/2}$）左乘其去嵌入向量，正交化跨标记的最大方差轴。尽管在减少非目标效应方面效果显著，我们假设这种变换可能会人为地分隔语义相关特征，从而扭曲特征表示。

在验证了我们的预测作为类似于调查评分的语义关联指标后，我们对调查和预测数据集进行了系列分析，以比较它们的相关结构。我们首先计算了调查中测量的所有 28 个语义尺度在我们分析样本中的所有词之间的完整相关矩阵。对于语义特征向量的词向量预测，我们也做了同样的处理，产生了一组可比较的相关矩阵。我们还计算了每对大型语言模型嵌入中的语义特征向量之间的余弦相似度。尽管预测之间的相关性揭示了数据点在这些轴上的相关性，余弦相似度则揭示了这些轴本身的非正交程度。

最后，对于调查和投影数据集，我们应用 PCA 并识别每个成分所解释的方差以及前三个成分中负载最高的变量。这些平行分析有助于揭示和描述由调查测量的 28 个语义标度与嵌入投影之间的相关结构，以研究它们的共同行为模式。

### 3.3 干预和非靶向效应

在评估特征之间的相关性之后，我们通过测量对嵌入空间进行干预对下一个标记概率的影响来评估因果关系。对于每个标记 $i$，每个反义词对 $j$，以及每个语义特征 $f$，我们使用以下模板构建提示：

> USER: 你认为 { token $_i$ } 更像是 { first_antonym $_{f,j}$ } 还是 { second_antonym $_{f,j}$ }？请从这两个词中选择一个，不要有格式。ASSISTANT: 在 { first_antonym $_{f,j}$ } 和 { second_antonym $_{f,j}$ } 之间，我认为 { token $_i$ } 更像是

然后我们计算两个反义词（例如 “kind” 和 “cruel”）作为下一个标记的概率，然后归一化概率使其总和为 1。我们对每个语义特征 $f$ 的所有十对反义词以及两个反义词的顺序进行重复此过程，以提高鲁棒性，并对所得的归一化概率取平均值，如方程 (2) 所述。

$$p_{\text{norm}}(ant_f^+|w_i) = \frac{1}{N}\sum_{j=1}^{N}\frac{p(ant_{f,j}^+|w_i)}{p(ant_{f,j}^+|w_i) + p(ant_{f,j}^-|w_i)} \tag{2}$$

在公式 2 中，$p_{\text{norm}}(ant_f^+|w_i)$ 表示在给定目标标记 $w_i$ 的情况下，选择与特征 $f$ 相关的正反义词的平均标准化概率。术语 $ant_{f,j}^+$ 和 $ant_{f,j}^-$ 分别指与特征 $f$ 相关的第 $j$ 对反义词中的正反义词，而 $N$ 表示为该特征考虑的反义词对的总数量。

在收集基线概率后，我们对每个提示执行逐标记的干预，推动目标标记朝向 28 个特征向量中的每一个调整。具体来说，我们根据上述过程（参见公式 (1) ）计算 28 个特征向量，然后对于每个标记 $i$ ，通过将其在嵌入矩阵中的表示朝着语义特征 $\boldsymbol{d}_f$ 的方向移动来修改该标记的表示。通过将缩放后的语义特征向量 $\boldsymbol{cd}_f$ 添加到目标标记 $\boldsymbol{w}_i$ 来实现这一点。干预向量被缩放为 $\pm 0.35 \cdot \|\boldsymbol{w}_i\|$ 的大小；初步测试表明，这种干预水平可以保持输出的一致性和标记的定义性意义。在应用干预后，我们对修改后的标记向量进行归一化，以恢复其原始大小 $\|\boldsymbol{w}_i\|$ 。


Figure 2: 在这个例子中，目标词“冬天”在美丑语义方向上被移动到美的方向。然后，LLM 计算在不同语义方向上表达的关联的概率，在这种情况下是善良-残酷。在此情况下，在潜在空间中使“冬天”更加美丽会对模型的表达关联产生离靶效应，使其在某种程度上显得更加善良。

## 4 调查中的语义结构和嵌入

### 4.1 投影数据

我们首先评估人类语义评分与 LLM 嵌入中词在相应特征向量上投影之间的相关性（图 3 ）。我们发现，在各个特征之间，嵌入投影与人类评分之间的相关性很高，范围从 0.3 到 0.7。鉴于这些方法在估计语义关联方面的显著差异，以及测量诸如语义评分等主观现象固有的高度测量误差，这些数值表明心理关联与 LLM 嵌入空间中标记的位置之间存在很强的一致性。

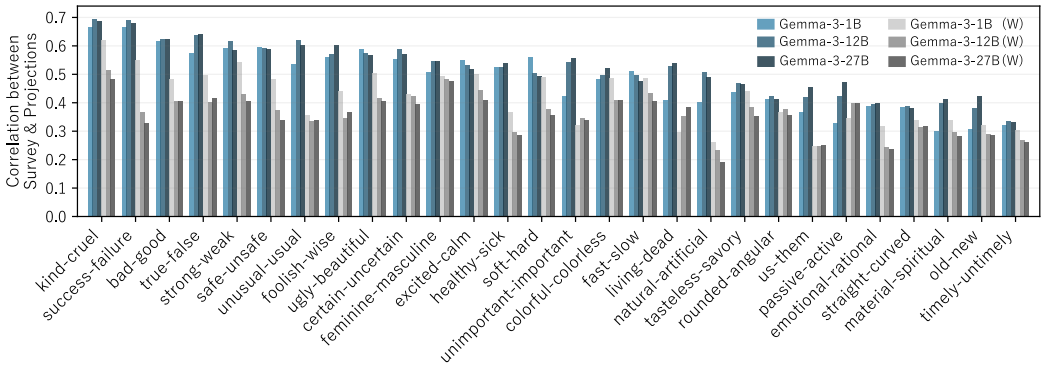


Figure 3: 人类评分与 28 个语义特征上 301 个词的词元嵌入投影之间的 Pearson 相关性。与“白化”嵌入空间的相关性标记为 (W)。

应用 Park 等人所描述的白化变换 [9] 到嵌入降低了我们 28 个投影分数与相应调查评分之间的平均标记层面相关性约 20 %。这样的下降与我们的假设一致：白化使标记云变得各向同性，并消除了之前共享方差的方向的人为关联。然而，人类评估判断确实依赖于这种共享方差，因此去除这些重叠会降低表示对心理和文化关联的忠实度。这个发现表明，嵌入中的轻微非正交性是语义结构的忠实反映，而不是噪声。

接下来，我们将调查数据中获得的语义结构与从大型语言模型嵌入中提取的语义结构进行比较。正如预期的那样，图 4 的 A 部分展示了调查数据中语义尺度之间的高度相关性。B 到

D 部分显示了 Gemma-3-12B 嵌入中特征之间的相关性（其他大型语言模型的结果在附录中包含）。大型语言模型特征之间的相关性虽然不如调查数据中观察到的那么高，但仍然显著，无论是将词汇限制为调查项目（B 部分）还是对整个嵌入词汇（D 部分）进行分析时，都是如此。这些相关性尤其显著，当与模型试图保持特征正交的替代假设相比时。此外，投影数据中的相关性结构与调查数据的非常相似。我们取调查相关矩阵的条目及相应的投影相关矩阵条目，发现当限制为调查词汇时，皮尔森相关系数为 0.82，对于完整词汇则为 0.73。然而，即使是对完整词汇，也有可能我们观察到的仍然是观察数据点的工件，而不是潜在特征本身的对齐。为了评估语义特征本身的对齐，我们在图 4 的 C 部分展示了跨特征余弦相似性的分布。虽然许多特征几乎正交，余弦相似性小于 0.05，但有很大比例显示出显著的相似性，许多在 0.1 到 0.3 之间。语义向量余弦相似性与调查数据集的相关系数之间的相关性为 0.76，这强烈地表明人类有意义的语义关联被编码在嵌入空间中特征向量的几何关系中。

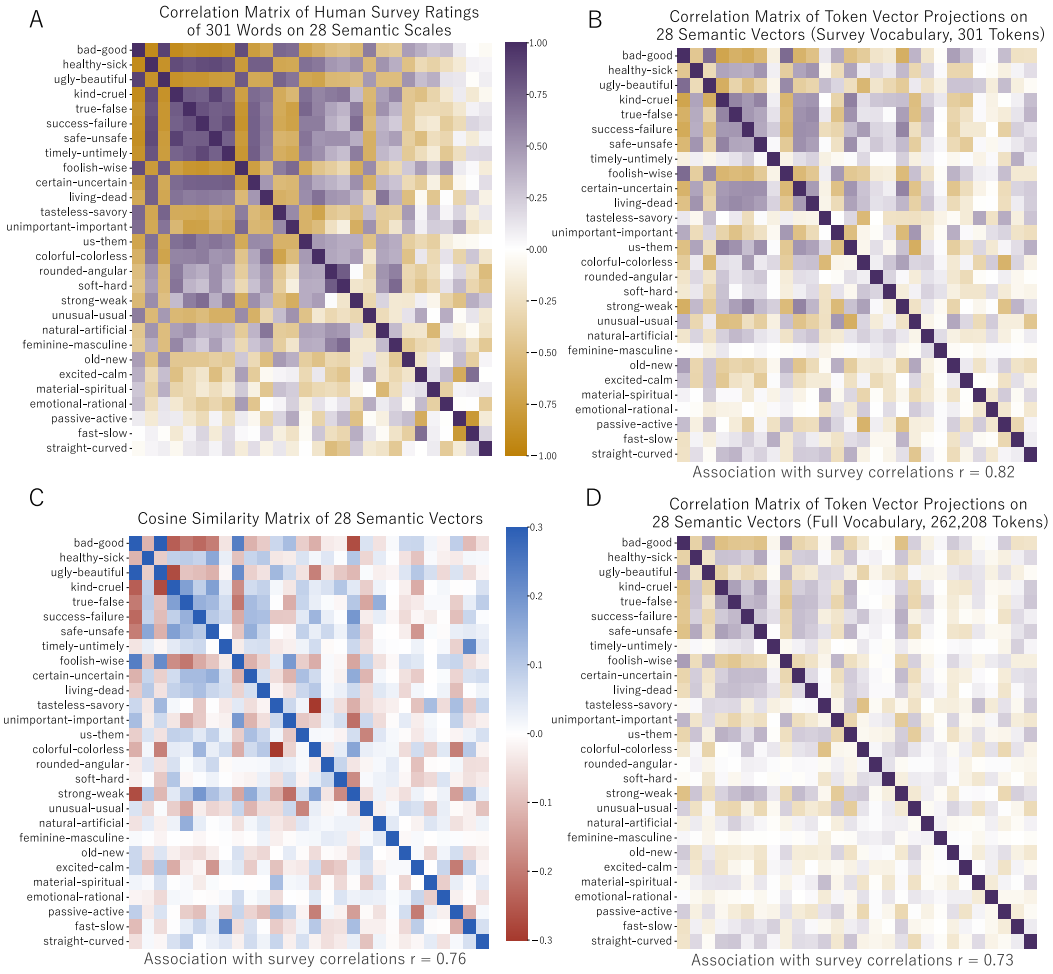


Figure 4: 面板 (A)：301 个单词在 28 个语义尺度上的人工评级的皮尔逊相关性。面板 (B)：Gemma-3-12B 嵌入中 301 个单词向量（调查词汇）在 28 个语义特征向量上的投影的皮尔逊相关性。面板 (C)：Gemma-3-12B 嵌入中语义特征向量对之间的余弦相似性。面板 (D)：Gemma-3-12B 嵌入中所有标记向量在 28 个语义特征向量上的投影的皮尔逊相关性。

接下来，我们检查主成分分析（PCA）的结果，这些结果展示在图 5 中，以评估调查的相关性和来自 LLM 预测的相关性是否可以简化为类似的潜在因子结构。面板 A 显示了调查数据中的每个主要成分及三个 LLM 嵌入的解释方差的比例。虽然调查数据中的前三个成分解释了较大比例的方差，但 LLM 嵌入的前三个成分也捕捉到了大量的方差，特别是相对于特征的完全正交基线，在这种情况下，28 个成分中的每一个只解释 3.6 % 的总方差。

### 4.2 测量语义结构

在检视前三个成分的因子负荷时，我们首先发现调查清晰地再现了评估、力量和活动这三个维度。第一个成分最接近于坏-好和病-健康，极端值出现在像“骗子”和“小偷”这样的词的一端和“优雅”和“和平”这样的词的另一端。乍一看，第二个成分似乎与活动相关，消极-积极和慢-快是最高的负荷，但“懒散”和“柔软”与“剑”和“军队”的极端值显示了弱点与力量的对比。相比之下，第三个成分最极端的词语，“龙卷风”、“孩子”和“小狗”在一端，而“平原”、“椅子”和“盒子”在另一端，更清楚地表明了活动与运动对比于无生命的静止。

所有三种大小的 Gemma 模型显示出相似的结果。与调查中的情况一样，第一个成分显然捕捉到了评估，坏-好成为所有模型中第一个成分的最高载荷特征。虽然第二个成分与强度并不完全对应，但更像是“活力”。在各个模型尺寸中，最高载荷的特征是空白-多彩和乏味-可口，最高载荷的词语在一端是“普通”、“缓慢”和“基础”，在另一端是“辛辣”、“火热”和“迷人”。第三个成分大致对应于活跃性，激动-平静和快速-缓慢始终出现在高载荷特征中，而“迅速”和“危险”与“平静”和“缓慢”等词语展现了最极端的数值。总体而言，这些结果表明，在 28 个语义特征向量上的投影简化为一个低维子空间，虽然不完全相同，但与从调查结果中产生的子空间非常相似。

## 5 预测非靶向效应

先前的分析建立了大型语言模型嵌入中的语义特征与人类调查数据中结构相似部分之间的有意义相关性。然而，这种特征“纠缠”的含义尚不明确。在这最后的分析阶段，我们探讨一个特征向量上的位置是否与模型对词语在非目标但相关语义特征上的关联评估存在因果关系。结果如图 6 所示。每个点代表一对特征向量，x 轴上是特征对的余弦相似度，y 轴上是非目标效应的大小。我们注意到特征向量之间“纠缠”与其非目标效应大小之间存在很强的线性关联。尽管效应在最小的模型中最强，其中最对齐的特征对表现出 10 个百分点的平均非目标效应，12B 和 27B 模型也显示出明显的关联性和显著的非目标效应。这些结果表明，沿着特征向量的词语位置可预测地影响模型对它们的语义判断。此外，多个特征向量之间的对齐强度足以使得显著的非目标效应可预测地发生。

## 6 结论

大型语言模型以其高维表征而闻名，其性能随参数数量的幂律函数在许多数量级上呈一致缩放此现象似乎表明，构建一个有效的世界模型需要大量的维度。然而，先前的心理学研究表明，我们用来交流和理解世界的许多看似多样的意义轴大致可以表示为仅三个潜在因素的线性组合。我们的研究结果表明，与人类一样，LLM 嵌入中的语义关联表示为相对低维。该发现对技术 AI 安全工作具有鼓舞意义，因为它表明许多相关的语义特征（例如善良-残酷，安全-不安全，我们-他们）可能共享一个共同的低维子空间，从而简化了识别、审计和控制的任务。

此外，本研究通过明确测量和建模多个特征之间的关系，超越了现有大多数关于特征识别的研究。虽然识别单个特征向量对于模型审计和引导仍然重要，但对大型语言模型内部操作的更完整理解必须考虑特征如何彼此关联，以及这种定位如何影响模型行为。然而，本研究仅为绘制特征空间几何图形提供了初步步骤。未来的研究需要考虑这些几何形状如何在模型的层次和组件中变化，以及激活空间中特征之间的关系如何可能被上下文序列中的前续标记动态重新配置。人类理解世界并不是一组孤立的特征，而是通过关系、共性和差异。提升我们对大型语言模型如何理解世界的认知，可能不仅需要将特征组装成电路，还要组成概念图。

## References

[1] Alexander Ku, Declan Campbell, Xuechunzi Bai, Jiayi Geng, Ryan Liu, Raja Marjieh, R Thomas McCoy, Andrew Nam, Ilia Sucholutsky, Veniamin Veselovsky, et al. Using the tools of cognitive science to understand large language models at different levels of analysis. *arXiv preprint arXiv:2503.13401*, 2025.

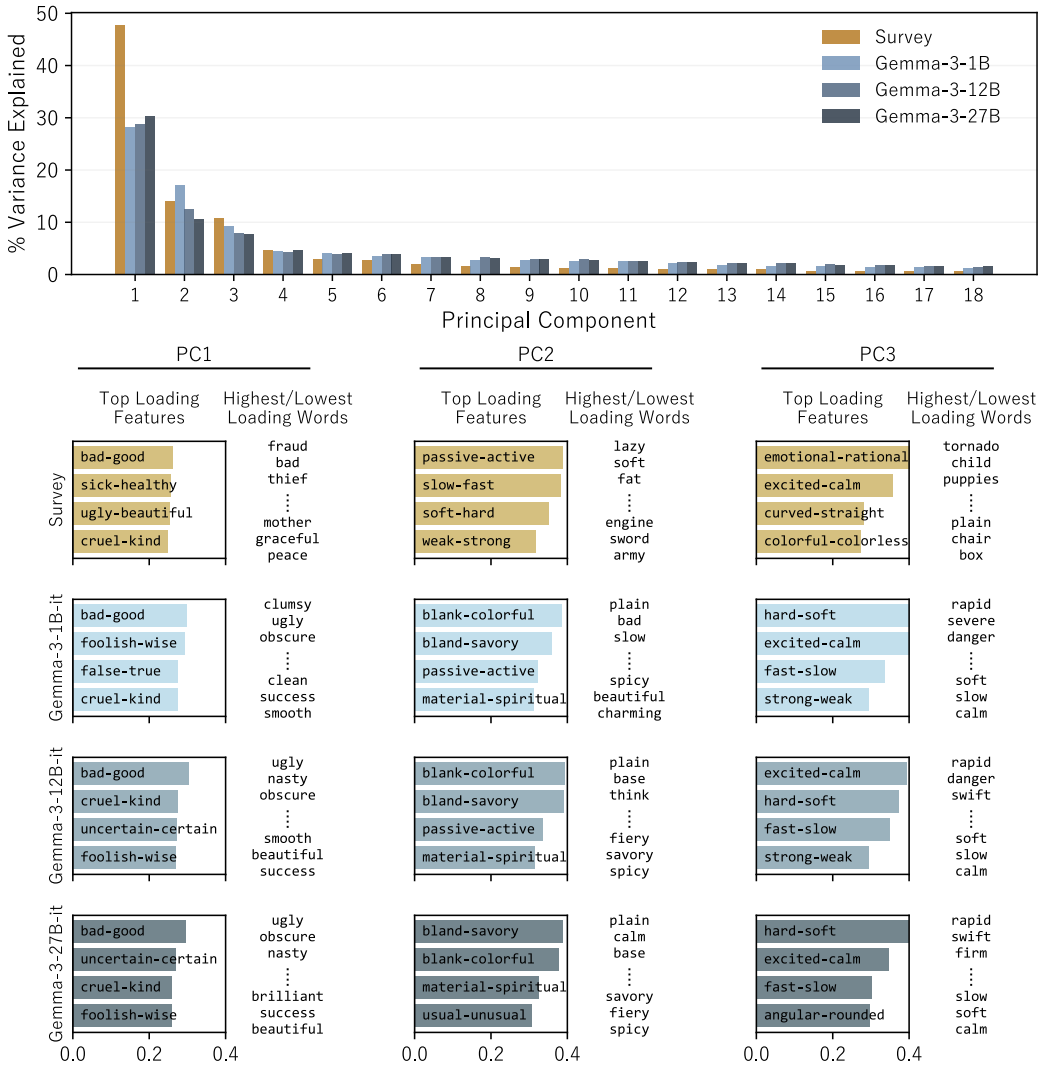


Figure 5: 对 301 个单词在 28 个语义尺度上的调查评分进行主成分分析的结果，以及这些 301 个词向量在 Gemma 3-1B、12B 和 27B 的嵌入中的 28 个对应语义特征向量上的投影。面板 (a)：前 18 个成分各自解释的方差的碎石图。面板 (b)：第一、第二和第三主成分上的最高载荷特征，以及每个成分上得分最高和最低的单词。


[2] Rohin Shah, Alex Irpan, Alexander Matt Turner, Anna Wang, Arthur Conmy, David Lindner, Jonah Brown-Cohen, Lewis Ho, Neel Nanda, Raluca Ada Popa, et al. An approach to technical agi safety and security. *arXiv preprint arXiv:2504.01849*, 2025.

[3] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

[4] Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. *The measurement of meaning*. University of Illinois press, 1957.

[5] David R Heise. *Surveying cultures: Discovering shared conceptions and sentiments*. John Wiley & Sons, 2010.

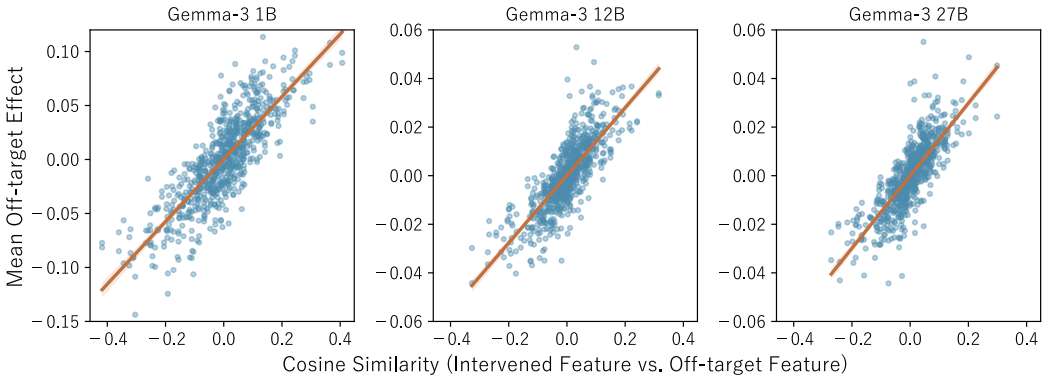


Figure 6: Gemma 3 模型中，通过目标和非目标语义特征向量的余弦相似性来衡量非目标效应的平均大小。效应被测量为下一个标记作为反义词 1 相对于反义词 2 的标准化概率的变化。所有关联在 $p < 0.001$ 处具有显著性。


[6] Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. In *Social cognition*, pages 162–214. Routledge, 2018.

[7] Lisa Feldman Barrett and James A Russell. The structure of current affect: Controversies and emerging consensus. *Current directions in psychological science*, 8(1):10–14, 1999.

[8] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

[9] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

[10] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[11] Esin Durmus, Alex Tamkin, Jack Clark, Jerry Wei, Jonathan Marcus, Joshua Batson, Kunal Handa, Liane Lovitt, Meg Tong, Miles McCain, Oliver Rausch, Saffron Huang, Sam Bowman, Stuart Ritchie, Tom Henighan, and Deep Ganguli. Evaluating feature steering: A case study in mitigating social biases. *Anthropic Blog*, 2024.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu,

Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

[18] Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025.

[19] Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*, 2023.

[20] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

[21] 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.

[22] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

[23] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. https://distill.pub/2020/circuits/zoom-in.

[24] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[26] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[27] Geoffrey Hinton, James L. McClelland, and David E. Rumelhart. Distributed representations. In James L. McClelland and David E. Rumelhart, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, pages 77–109. MIT Press, Cambridge, MA, 1986.

[28] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27, 2014.

[29] Austin C Kozlowski, Matt Taddy, and James A Evans. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949, 2019.

[30] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[31] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[32] Charles E Osgood. Semantic differential technique in the comparative study of cultures. *American anthropologist*, 66(3):171–200, 1964.

[33] Andrei Boutyline and Ethan Johnston. Forging better axes: Evaluating and improving the reliability of semantic dimensions in word embeddings. *SocArXiv*, 2025.

[34] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

# A 补充材料

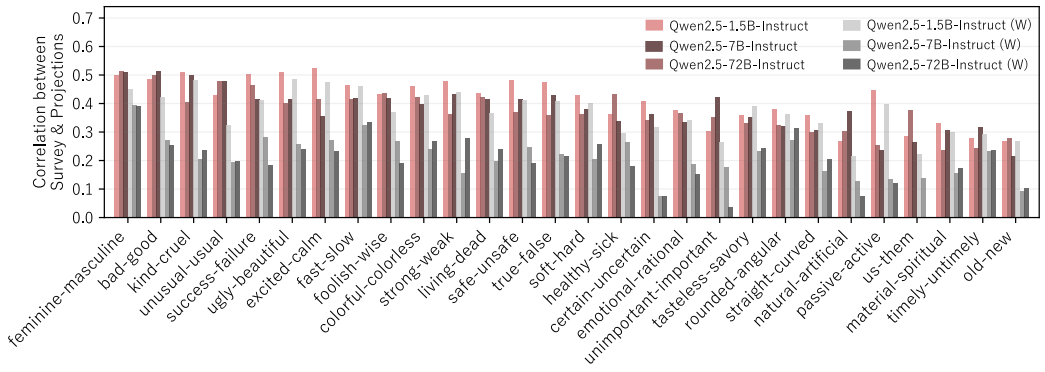


Figure 7: 在人类评分与 Qwen2.5 系列模型中 28 个语义特征的 301 个词之间进行的词嵌入投影之间的 Pearson 相关性。使用“白化”嵌入空间的相关性被标记为 (W)。

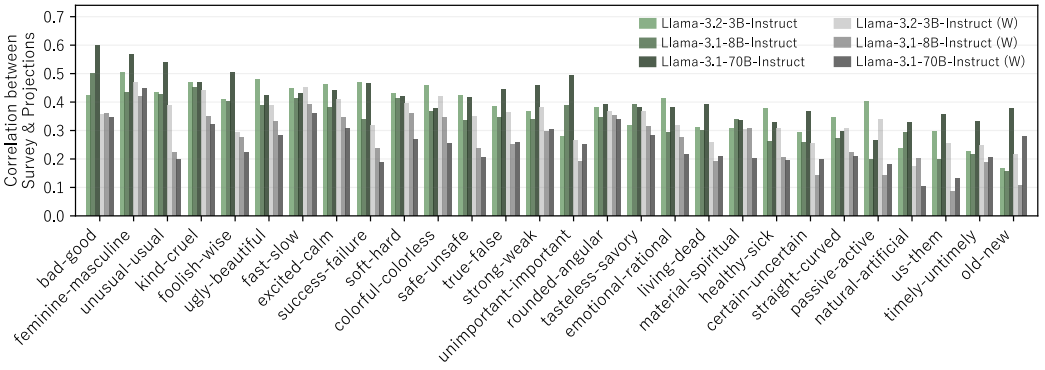


Figure 8: 在 Llama 3 系列模型中，301 个词在 28 个语义特征上的人类评分与词元嵌入投影之间的 Pearson 相关性。使用“白化”嵌入空间的相关性以 (W) 标记。

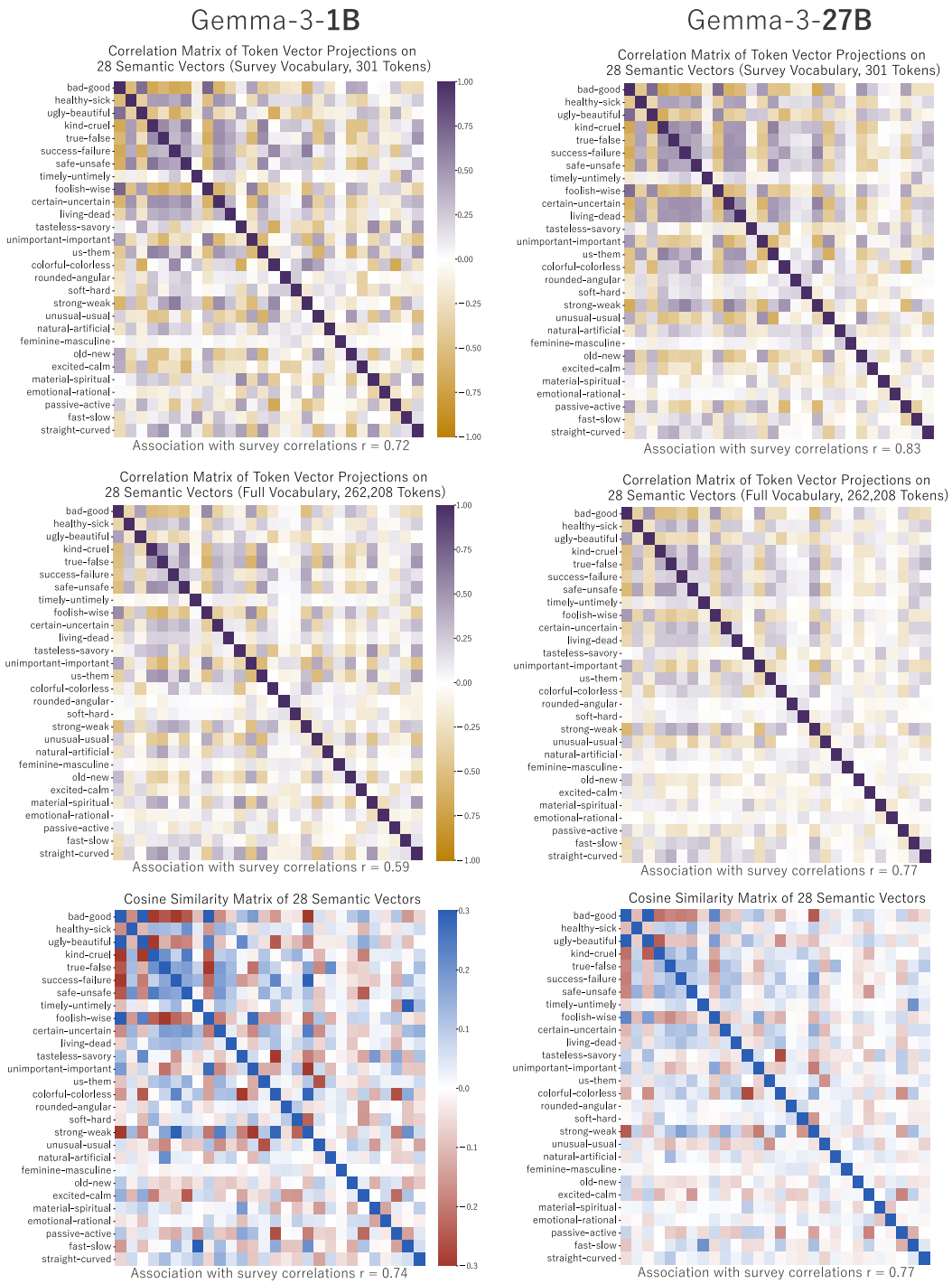


Figure 9: 来自 Gemma-3-1B 和 Gemma-3-27B 的结果：301 个词向量（调查词汇）在 28 个语义特征向量上的投影的皮尔逊相关系数，所有标记向量在 28 个语义特征向量上的投影的皮尔逊相关系数，以及成对语义特征向量之间的余弦相似度。

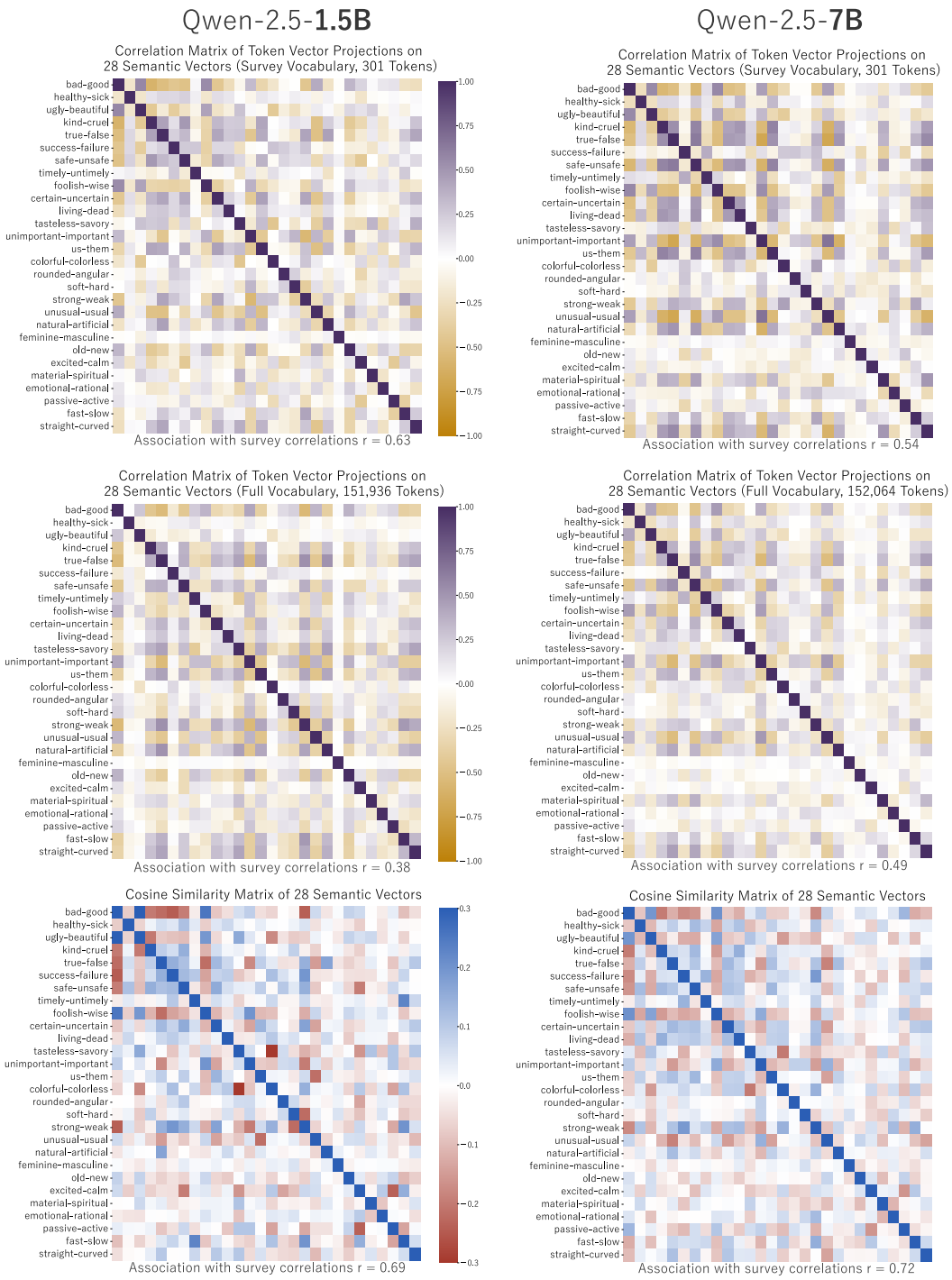


Figure 10: Qwen-2.5-1.5B 和 Qwen-2.5-7B 的结果：301 个词向量（调查词汇）在 28 个语义特征向量上的投影之间的 Pearson 相关性，所有标记向量在 28 个语义特征向量上的投影之间的 Pearson 相关性，以及语义特征向量对之间的余弦相似性。

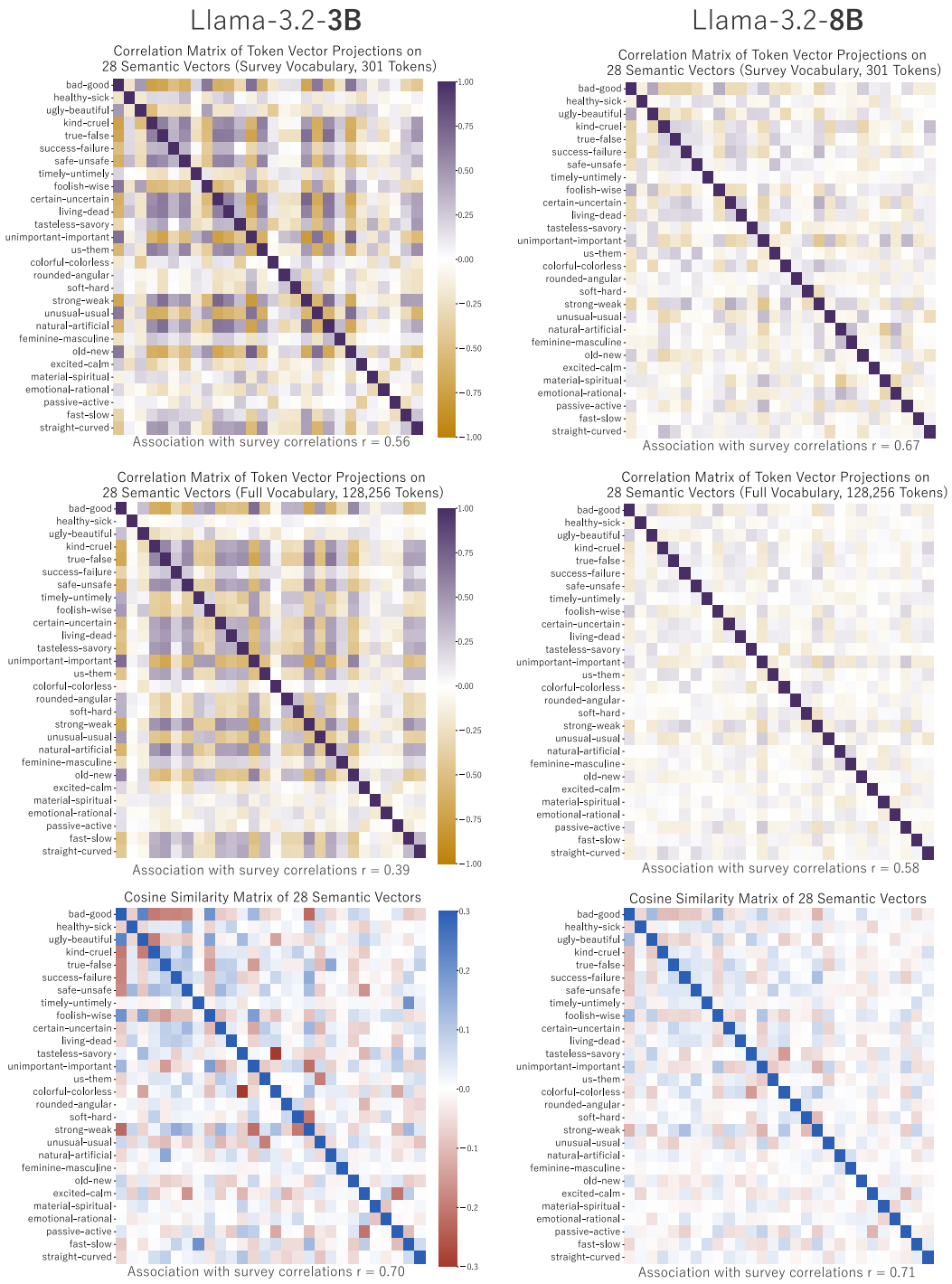


Figure 11: 来自 Llama-3.2-3B 和 Llama-3.1-8B 的结果：301 个词向量（调查词汇）在 28 个语义特征向量上的投影之间的皮尔逊相关，所有标记向量在 28 个语义特征向量上的投影之间的皮尔逊相关，以及语义特征向量对之间的余弦相似度。

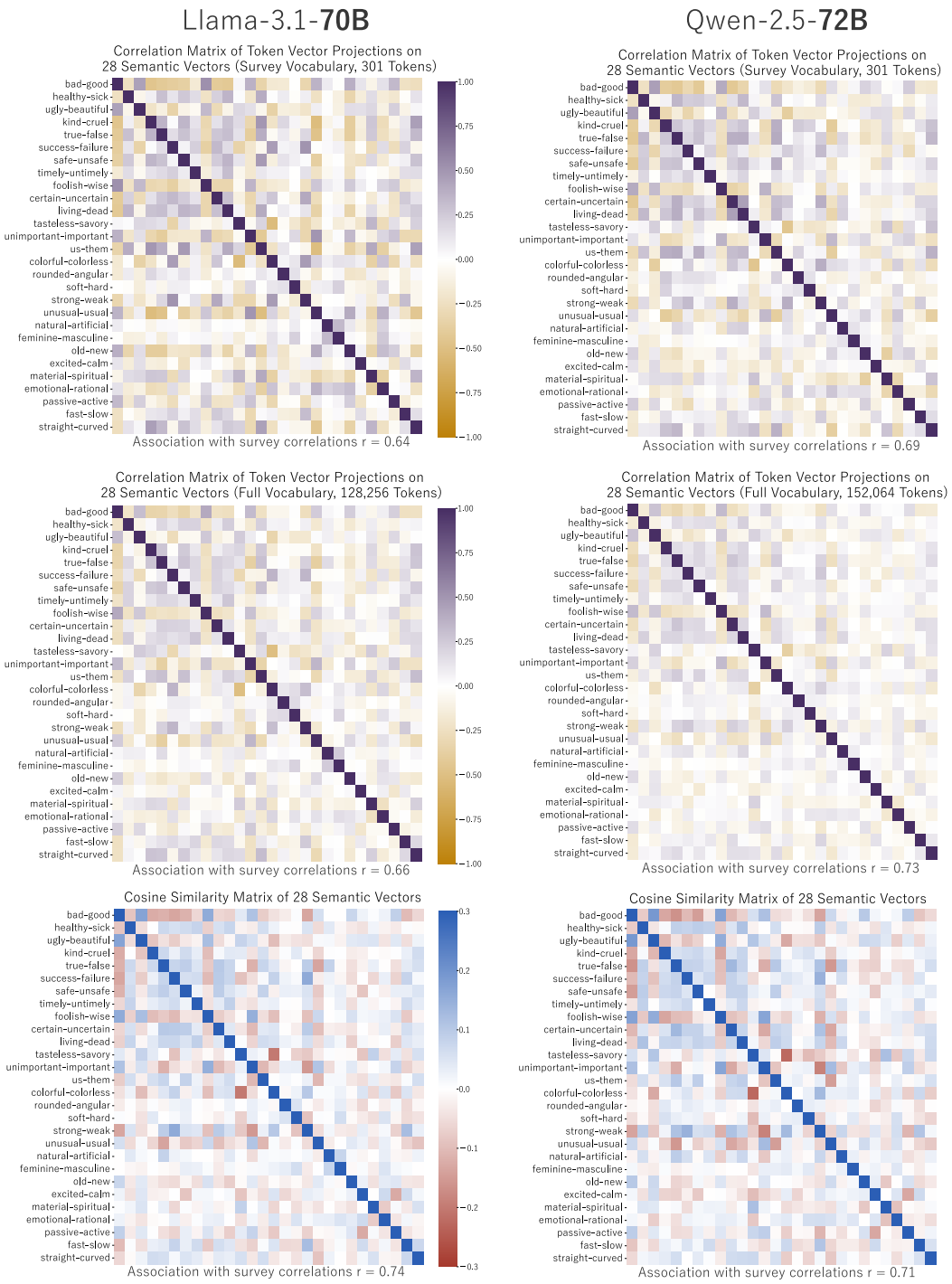


Figure 12: 来自 Llama-3.1-70B 和 Qwen-2.5-72B 的结果：301 个词向量（调查词汇）在 28 个语义特征向量上的投影的 Pearson 相关性，所有标记向量在 28 个语义特征向量上的投影的 Pearson 相关性，以及语义特征向量对之间的余弦相似度。

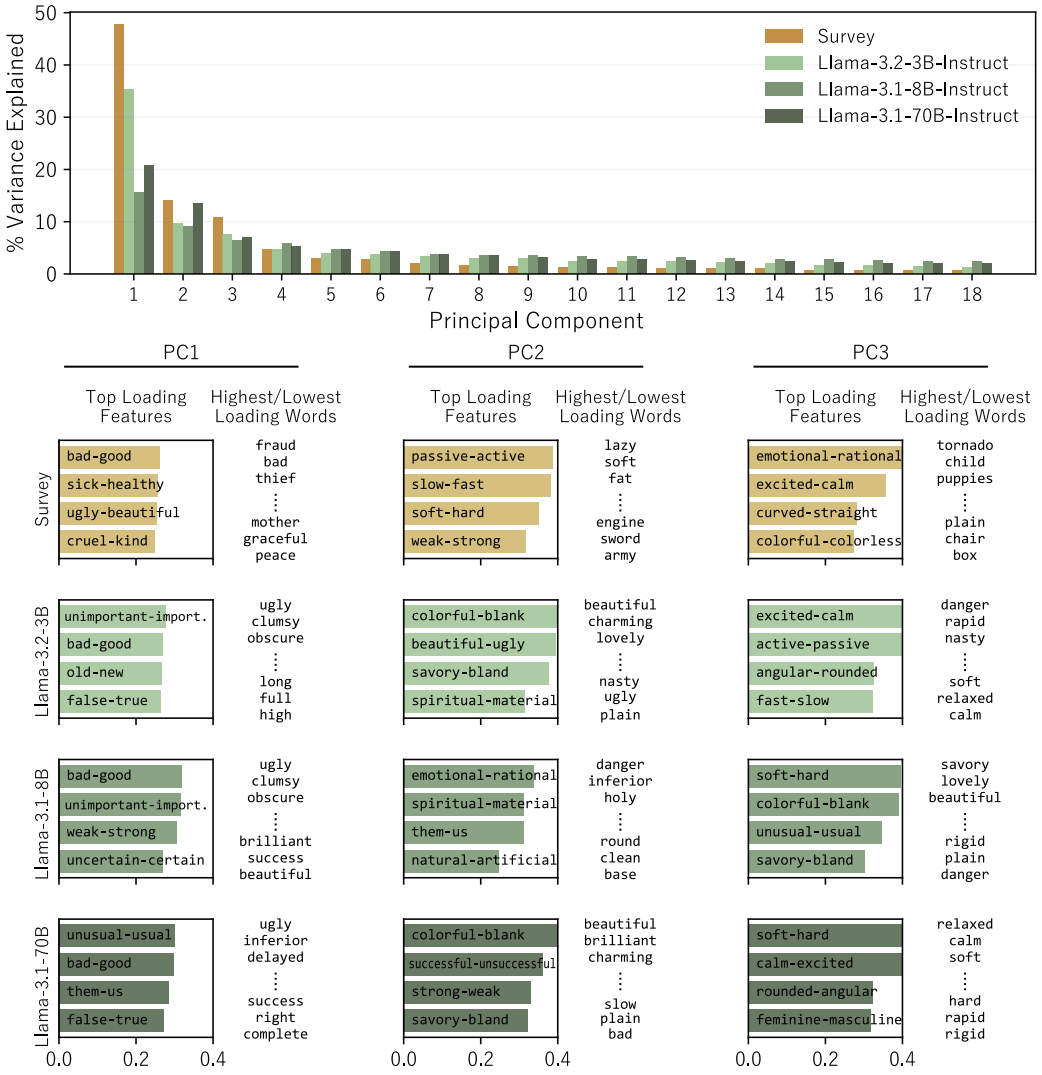


Figure 13: 主成分分析应用于对 301 个单词的 28 个语义尺度的调查评分，及在 Llama-3 3B、8B 和 70B 模型的嵌入中，这 301 个单词向量在 28 个对应语义特征向量上的投影。图 (a)：解释每个前 18 个成分的方差的碎石图。图 (b)：在第一、第二和第三主成分上加载最高的特征，以及每个成分上得分最高和最低的单词。

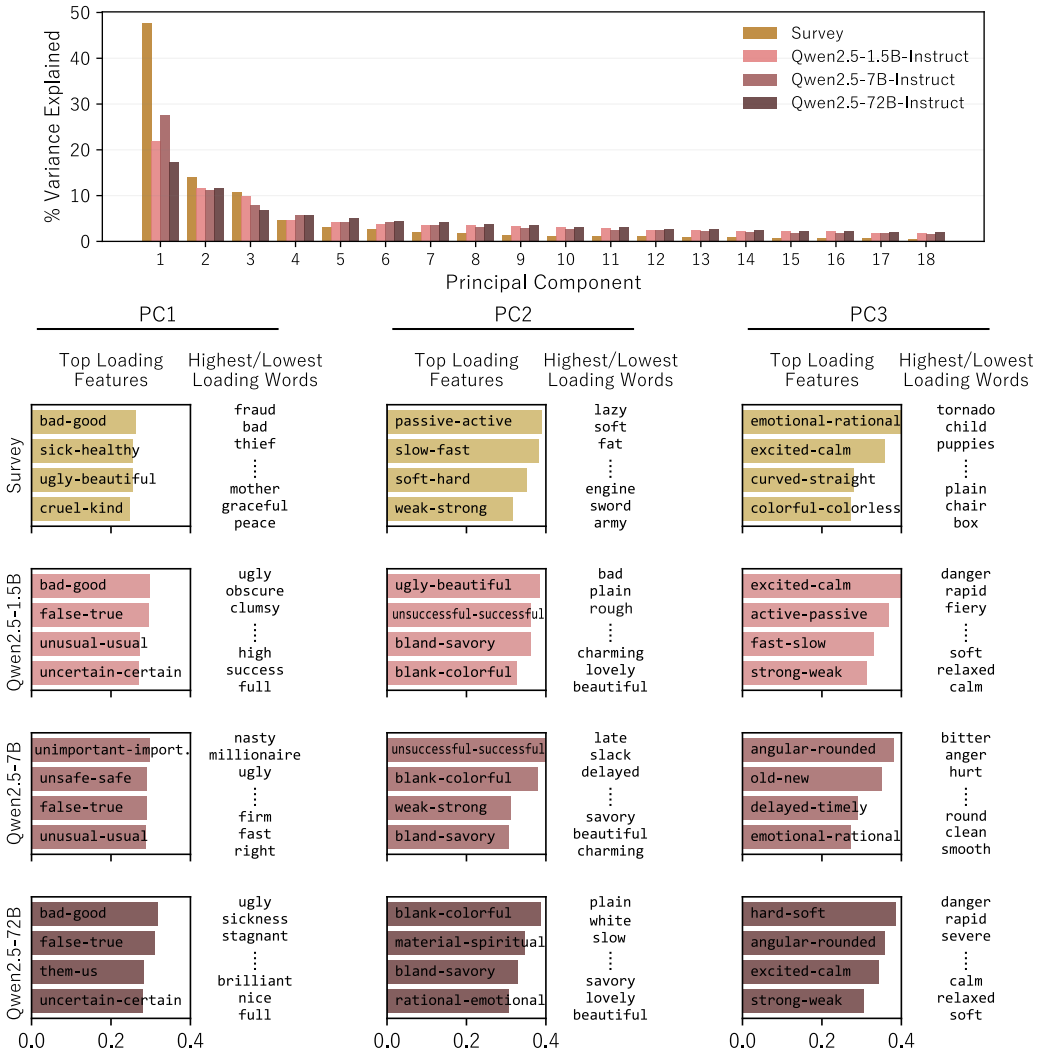


Figure 14: 主成分分析的结果应用于从 28 个语义尺度对 301 个词的调查评分，以及这 301 个词向量在 Qwen2.5 1.5B、7B 和 72B 模型中 28 个相应语义特征向量的嵌入中的投影。图（a）：方差的碎石图，由前 18 个成分解释。图（b）：在第一、第二和第三主成分上的最高负载特征，以及在每个成分上的最高和最低评分的词。

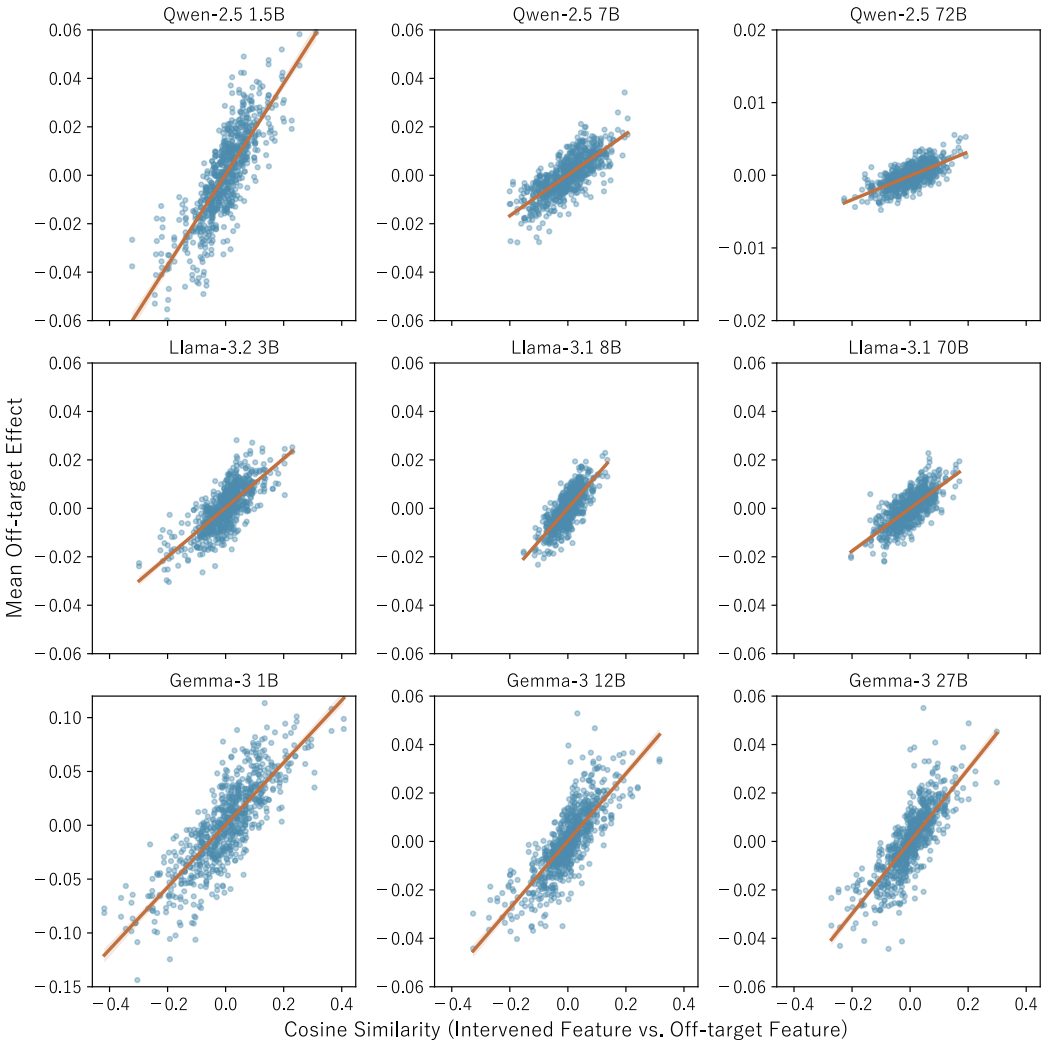


Figure 15: 在 Qwen、Llama 和 Gemma 模型中，通过目标和非目标语义特征向量的余弦相似性来计算非目标效应的平均大小。这些效应是通过下一个标记是反义词 1 相对于反义词 2 的归一化概率变化来衡量的。