



北京大学
PEKING UNIVERSITY

语言模型中是否隐藏了空间异形同义信息？

秦宇航

北京大学中文系

July 10, 2026

目录

① 引入

② 方法及提案

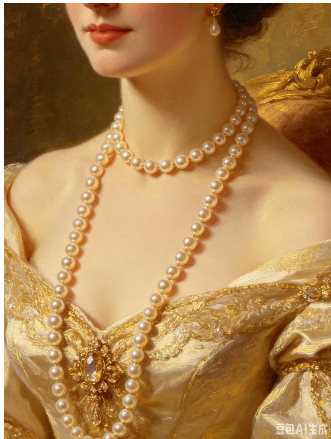
③ 实验及分析

④ 小结

空间异形同义：“一景”对“多句”




 北京大学
 PEKING UNIVERSITY



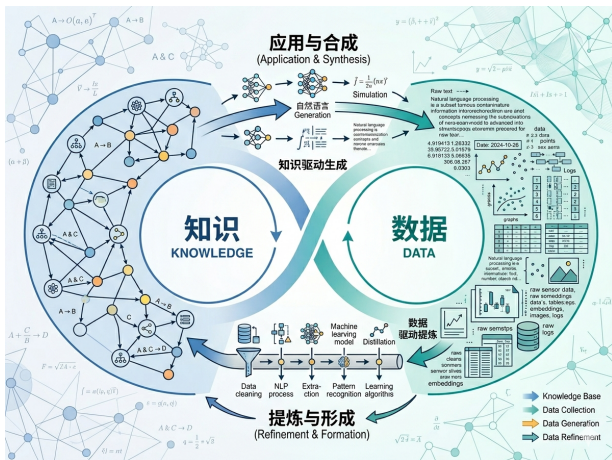
→ 贵妇人的脖子上挂着一串珍珠项链。

北京大学
PEKING UNIVERSITY



- 贵妇人的脖子上挂着一串珍珠项链。
- 贵妇人的脖子下挂着一串珍珠项链。

研究范式：知识与数据的联动



问题：循环不成立



■ 缺知识

¹Qin et al., 2025; 詹卫东 等, 2024, 2025.

²Qin et al., 2025; 詹卫东 等, 2025.

问题：循环不成立



■ 缺知识

- 当前的知识¹并未完全覆盖所有的异形同义类型

¹Qin et al., 2025; 詹卫东 等, 2024, 2025.

²Qin et al., 2025; 詹卫东 等, 2025.

问题：循环不成立



■ 缺知识

- 当前的知识¹并未完全覆盖所有的异形同义类型

■ 缺数据

¹Qin et al., 2025; 詹卫东 等, 2024, 2025.

²Qin et al., 2025; 詹卫东 等, 2025.

问题：循环不成立



■ 缺知识

- 当前的知识¹并未完全覆盖所有的异形同义类型

■ 缺数据

- SpaCE2025 异形同义数据集仅 1100 条²，每个词对对应的语料 ≤ 20

¹Qin et al., 2025; 詹卫东 等, 2024, 2025.

²Qin et al., 2025; 詹卫东 等, 2025.

自大语言模型 (LLM) 中发掘隐藏知识



大语言模型 (LLM) 在许多语言任务上都具有良好的表现。

LLM 的训练数据中包含了大量的文本信息，可能隐含了与空间异形同义相关的语言知识，可以用于筛选异形同义语言数据。

可能的提取方案：

- 基于最终输出：通过提示工程 (prompt engineering) 设计问题，从 LLM 的输出中获得数据；

自大语言模型 (LLM) 中发掘隐藏知识



大语言模型 (LLM) 在许多语言任务上都具有良好的表现。

LLM 的训练数据中包含了大量的文本信息，可能隐含了与空间异形同义相关的语言知识，可以用于筛选异形同义语言数据。

可能的提取方案：

- 基于最终输出：通过提示工程 (prompt engineering) 设计问题，从 LLM 的输出中获得数据；
- 基于中间表示：通过 LLM 的嵌入 (embedding)，提取潜在的知识。



基于最终输出的方法：模型表现存在不平衡

模型名称	参考答案	预测同义	预测异义	回复无效 ¹	F1 score
gpt-5.5	异形同义	244	296	10	0.6085
	异形异义	8	528	14	0.7686
deepseek-v4-flash	异形同义	346	204	0	0.7261
	异形异义	57	493	0	0.7907
deepseek-v4-pro	异形同义	277	271	2	0.6525
	异形异义	22	524	4	0.7792

Table: 3 种先进模型在 SpaCE2025 异形同义数据集¹上的表现。模型对参考答案为“异形异义”的试题判断准确率显著高于参考答案为“异形同义”的试题，模型对两类试题的判断存在不平衡。

¹Qin et al., 2025.

¹表示未能成功提取模型 prediction 的试题数量。



基于中间表示的方法：提取模型嵌入 (Embedding)

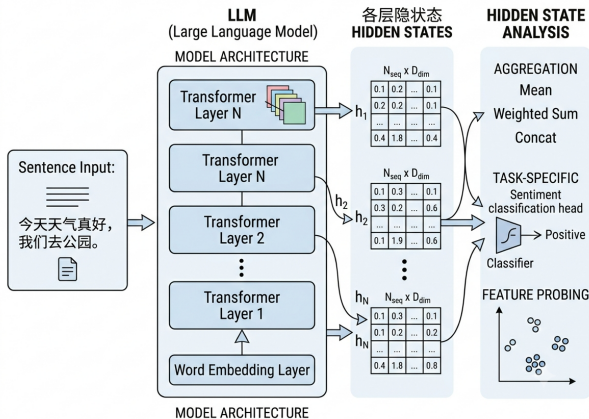


Figure: 基于模型的中间表示的方法示意图。将需要提取的文本输入模型中，提取模型各层的隐藏状态 (嵌入)，用于下一步的研究。

基于嵌入的下游任务



文本嵌入可以用于多种类型的下游任务中，例如¹：

- 文本语义相似度 (Semantic Textual Similarity, STS)
- 信息检索 (Information Retrieval, IR)
- 通用嵌入 (Universal Embedding, UE)
- 长上下文压缩 (Long Context Compression, LCC)
- 文本反转 (Text Inversion, TI)
- ...

¹Nie et al., 2025.

文本语义相似度任务

文本语义相似度 (Semantic Textual Similarity, STS) 任务¹是判断一个句子对的意思是否近似的任务，评估方式是对模型嵌入的相似度和人工标注的相似度进行比较。

sentence1	sentence2	similarity score
A plane is taking off.	An air plane is taking off.	5
A girl is flying a kite.	A girl running is flying a kite.	4
A woman is eating something.	A woman is eating meat.	3
A woman is peeling a potato.	A woman is peeling an apple.	2
A man is playing the piano.	A woman is playing the violin.	1
A woman is dancing.	A man is talking.	0

Table: MTEB 数据集²中的句子对示例。数据集中的句子对都标注了一个相似度分数 x , $x \in [0, 5]$ 。分数的高低与句子语义的相似度相对应。

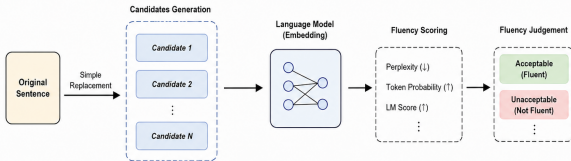
¹Cer et al., 2017; Enevoldsen et al., 2025; Marelli et al., 2014; Muennighoff et al., 2023.

²Muennighoff et al., 2023.



基于中间表示的异形同义数据合成 pipeline

1. Fluency Filtering



2. Similarity Computation

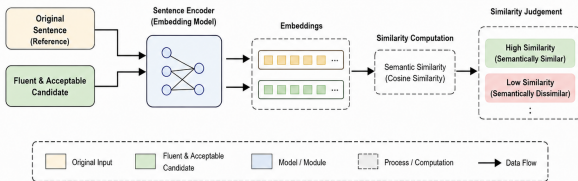


Figure: 基于中间表示的异形同义数据合成 pipeline。 **Step1:** 将原始句子做简单替换得到新的句子，然后通过模型嵌入计算困惑度等参数判断句子是不是可以说的句子。 **Step2:** 将 Step1 中得到的句子对输入到模型中，提取模型嵌入，然后计算句子对的相似度。

基于中间表示的异形同义数据合成 pipeline

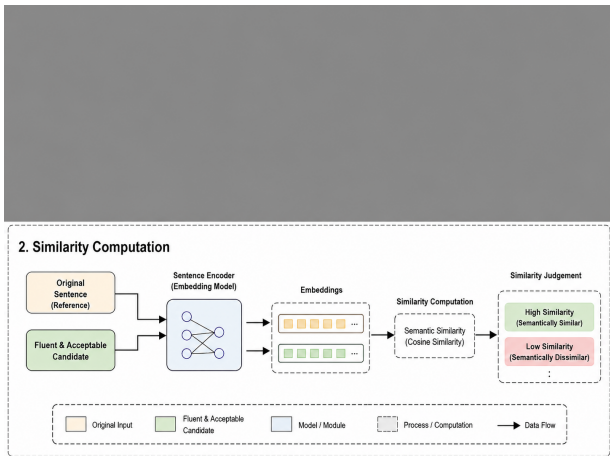


Figure: 基于中间表示的异形同义数据合成 pipeline。 **Step1:** 将原始句子做简单替换得到新的句子，然后通过模型嵌入计算困惑度等参数判断句子是不是可以说的句子。 **Step2:** 将 Step1 中得到的句子对输入到模型中，提取模型嵌入，然后计算句子对的相似度。

提案



- 数据形式：将 SpaCE2025 异形同义数据集¹中的句子对转换为类 STS 任务的形式，构建一个文本语义相似度测试集。
- 评估方式：使用文本嵌入 (embedding) 计算两个句子的相似度。

¹Qin et al., 2025.

实验设置



■ 模型选择：7B 级模型/30B 级模型

模型尺寸	模型名称
7B 级	Llama2-7B、Llama3.1-8B、Qwen2.5-7B Qwen3-8B、Qwen3.5-9B、GLM-4-9B-0414
30B 级	Qwen3-32B、Qwen3.5-27B

■ 池化方法：取最后一个 token 的 embedding

■ 相似指标：余弦相似度。对于句子 s_1 、 s_2 和模型层数 L ，余弦相似度为

$$\cos(s_1, s_2) = \frac{x_{s_1, L} \cdot x_{s_2, L}}{\|x_{s_1, L}\| \|x_{s_2, L}\|}$$

■ 同义预测：计算下述 prediction 和原始 label 的准确率

$$\text{pred}_{\text{syn}} = \begin{cases} 1, & \cos(s_1, s_2) > 0 \\ 0, & \text{otherwise} \end{cases}$$



实验 1：直接探测句嵌入

- 直接将句子输入模型，提取每层的句嵌入并逐对计算相似度
- 准确率：均在 0.5 附近，这是由于无论何种类型的句子，模型计算的相似度都极高，缺乏区分度
- label 和句对相似度的 Spearman 系数：相关性差，较大的模型相关性增强

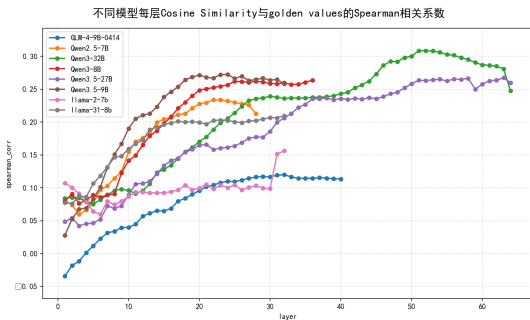


Figure: 直接提取句嵌入，label 和句对相似度的 Spearman 系数。



实验 2：增加 prompt 的嵌入探测

一系列研究表明，单纯将句子输入 **decoder-only** 模型的方法捕捉语义信息的能力不足。提高 **decoder-only** 模型在 STS 任务上的表现，目前的主流方法有 2 种：

- ① 有训练方法：调整模型架构，引入双向注意力，微调模型；
- ② 无训练方法：通过增加适当的 **prompt**，使得模型聚焦于特定的 **tokens** 上，以提高模型的表现。

实验 2：增加 prompt 的嵌入探测



我们基于已有的对 STS 任务的提升方法，尝试了一系列 prompt 设计：

- ① 单词释义 prompt¹: ‘句子”sentence” 的意思用一个词表示是’:
- ② 重复 prompt²: ‘被重复句:”sentence”; 重复句:”sentence”’
- ③ 假装思维链 (CoT) prompt³: ‘在一步步思考后，句子”sentence” 的意思是:’
- ④ 知识引入 prompt⁴: ‘方位义词的意义并不固定，和与其关联的目标物和参照物有关。在此基础上，句子”sentence” 中方位义词”location” 的意思是’

¹ Jiang et al., 2024.

² Springer et al., 2025.

³ Zhang et al., 2024.

⁴ Zhang et al., 2024.



实验 2 结果

- 准确率仍在 0.5 左右，句子相似度都极高；
- 增加 prompt 后，模型的表现并未显著提升；
- 不同 prompt 对于不同模型的效应存在不同，模型受到输入形式的影响；

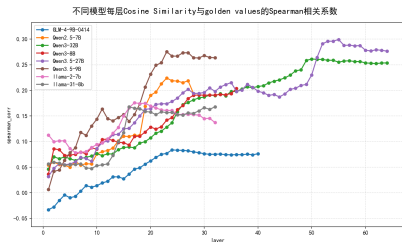


Figure: 以 prompt1 提取句嵌入, label 和句对相似度的 Spearman 系数。

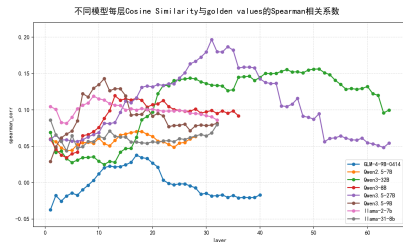


Figure: 以 prompt2 提取句嵌入, label 和句对相似度的 Spearman 系数。

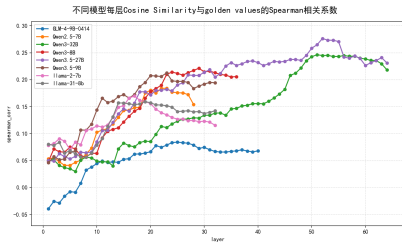


Figure: 以 prompt3 提取句嵌入, label 和句对相似度的 Spearman 系数。

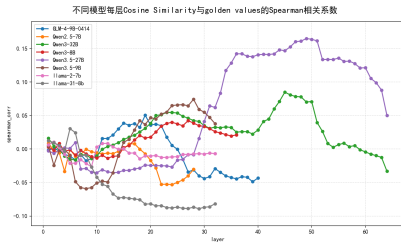


Figure: 以 prompt4 提取句嵌入, label 和句对相似度的 Spearman 系数。



实验 3：基于空间信息的 prompt 方法

基于提供任务相关的知识信息可以提升 STS 任务上的表现¹，我们尝试将空间语义角色标注信息引入输入的 prompt 中，包括如下两种 prompt:

- ① 仅引入目标物：' 在句子"sentence" 中，"target" 的位置或方向是： '
- ② 引入目标物和参照物：' 在句子"sentence" 中，"target" 相对于"reference" 的位置或方向是： '

¹Zhang et al., 2024.

空间信息的自动标注



由于 SpaCE2025 异形同义数据集并不完全包含这些信息，因此本文使用 Few-shot learning 方法，使用 deepseek-v4-pro 模型自动标注目标物和参照物。

- 输入 1100 条数据，成功返回目标物和参照物的数量为 1097 条¹，成功率为 99.73%；
- 通过人工检查，发现大部分标注可用，但标注结果中仍存在少量错误。

¹后续由人工补全

"id": 153,

"sentence1": " 母亲在家里做了很多墙布。厨房里的桌子上方挂着一块苹果和梨图案的墙布，上面还有一瓶葡萄酒和一只没有脑袋的烤鸡。图案下面有一行字：吃得好，忘烦恼。”，

"sentence2": " 母亲在家里做了很多墙布。厨房里的桌子上方挂着一块苹果和梨图案的墙布，下面还有一瓶葡萄酒和一只没有脑袋的烤鸡。图案下面有一行字：吃得好，忘烦恼。”，

"annotation":

"sentence1":

"target": " 葡萄酒和烤鸡”，

"reference": " 墙布”

"sentence2":

"target": " 葡萄酒和烤鸡”，

"reference": " 墙布”



实验 3 结果

增加相关 prompt 并未提升模型的表现，准确率仍在 0.5 左右，相似度并未有区分。

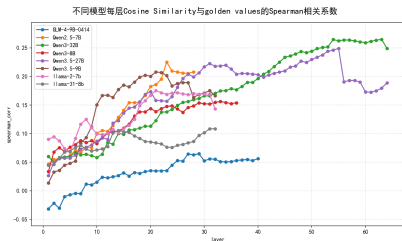


Figure: 以 prompt1 提取句嵌入，label 和句对相似度的 Spearman 系数。

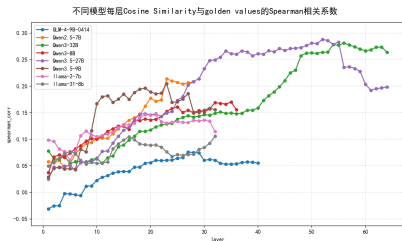


Figure: 以 prompt2 提取句嵌入，label 和句对相似度的 Spearman 系数。

实验 4：方位义词词嵌入探测



- 空间异形同义关系不同句子中存在方位义词语的差别。
 - 贵妇人的脖子上/下挂着一串珍珠项链。
- 这些方位词的区别是否体现在方位义词语的词嵌入中？

直接提取词嵌入的问题：语义不精确

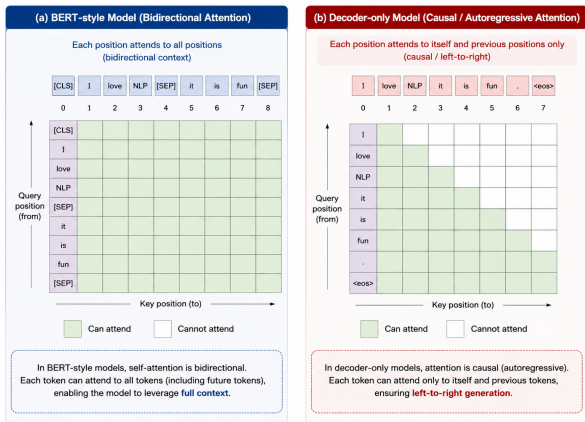


Figure: Bert 和 Decoder-only 模型的注意力机制区别。Decoder-only 模型的注意力只能注意到上文而无法注意到下文，因此直接提取的结果并不准确。

词嵌入提取方法



采用重复 prompt¹: '被重复句:"sentence"; 重复句:"sentence"', 将句子输入模型, 提取最后一层上、第二次重复中两句差异部分的词嵌入 (取算术平均)。

提取后, 对这些词嵌入做 PCA 降维, 观察不同方位义词的词嵌入是否存在区分。

¹ Springer et al., 2025.



实验 4 结果

PCA 降维结果表明，大量的方位义词语的词向量聚集在空间中一个较小的区域，区分度较低，这可能是空间同义和异义难以区分的原因。

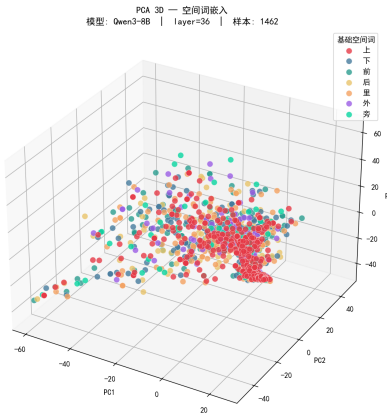


Figure: Qwen3-8B 模型的 PCA 结果

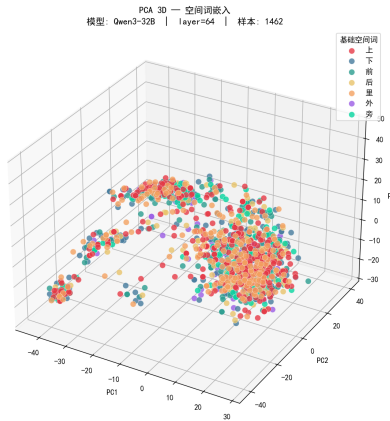


Figure: Qwen3-32B 模型的 PCA 结果

结论



- 对于模型，空间异形同义/异形异义对在语义上均接近，模型计算的相似度缺乏区分度；
- 适用于常规 STS 任务的 prompt 方法并不能在当前任务上取得良好效果；
- 在方位义词的词嵌入中，模型并未学习到区分不同方位义词的能力。
- 通过当前方法将模型嵌入用作空间异形同义判别器的可行性不高，当前测试模型的内部并不具备判定空间异形同义现象的能力。

问题与展望



- 目前计算出的语义相似度接近——由于句子过长，导致模型缺乏区分度，模型的表现不佳；需要探索更有区分度的 **embedding** 分析方法，或者模型不适用这一方法；
- 观察到参数更大的模型对两者区分能力更高——需要进一步在更大参数模型上进行实验。

谢谢!