计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(9): 2187−2201, 2026

大语言模型的齿状回:基于模式分离的场景认知能力诊断及增强框架

马博翔[1] 李 茹[1] 郭少茹[1] 李英豪[1] Víctor Gutiérrez-Basulto[2]
[1](山西大学计算机与信息技术学院 太原 030006)
[2](卡迪夫大学计算机与信息工程学院 威尔士卡迪夫 CF24 4AG)
(maboxiang@sxu.edu.cn)

Dentate Gyrus of Large Language Models: A Scenario Cognition Diagnosis and Enhancement Framework Based on Pattern Separation

Ma Boxiang[1], Li Ru[1], Guo Shaoru[1], Li Yinghao[1], and Víctor Gutiérrez-Basulto[2]
[1](*School of Computer Science and Technology*, *Shanxi University*, *Taiyuan* 030006)
[2](*School of Computer Science and Informatics*, *Cardiff University*, *Cardiff*, *Wales CF24 4AG*)

Abstract Large language models (LLMs) exhibit systematic generalization limitations in scenario cognition tasks, revealing potential functional deficiencies in their knowledge memorization processes. Inspired by the pattern separation mechanism of the dentate gyrus in the human hippocampus, we hypothesize that failures in scenario cognition are associated with limitations in the internal pattern separation capability of LLMs. Based on this hypothesis, we first design a cross-entity representation diagnosis (CERD) method, which quantitatively measures entity representation confusion across different layers without additional training. Our analysis reveals that standard autoregressive training fails to induce sufficiently effective pattern separation capabilities, leaving substantial cross-entity correlation confusion within LLM representations. Building upon this observation, we further propose latent-space pattern separation (LPS), which explicitly incorporates pattern separation constraints into the training process by introducing a triplet-based contrastive regularization at target layers with severe entity confusion, thereby reducing correlations among easily confused knowledge representations. Extensive experiments across LLMs ranging from 0.8B to 9B parameters demonstrate that LPS consistently improves scenario cognition capabilities. Compared with standard autoregressive training, our method achieves up to 101.5% relative improvement in exact match (EM), providing a new perspective for understanding and enhancing scenario cognition and knowledge memorization mechanisms in LLMs.

Key words large language models; scenario cognition; pattern separation; knowledge memorization; representation learning

摘 要 大语言模型在场景认知任务上表现出系统性的泛化局限,暴露了其知识记忆过程中可能存在的功能性缺陷。对此,受人脑海马体中齿状回结构的模式分离机制启发,提出了场景认知失败与模型内部模式分离能力局限之间的相关性假设。基于这一假设,首先设计了一种跨实体表示诊断(cross-entity representation diagnosis, CERD)方法,在无需额外训练的条件下量化模型各层的实体表示混淆程度,观察到标准自回归训练并未使模型形成充分可用的模式分离能力,其内部仍存在跨实体的相关性混淆。在此

收稿日期: 2026-06-06; **修回日期**: 2026-07-16
基金项目: 国家自然科学基金项目(62376144, 62506218); 山西省基础研究计划资助项目(202403021211092)
This work was supported by the National Natural Science Foundation of China (62376144, 62506218) and the Fundamental Research Program of Shanxi Province (202403021211092).
通信作者: 李茹(liru@sxu.edu.cn)

基础上,进一步提出隐空间模式分离(latent-space pattern separation, LPS)方法,在实体混淆较为严重的目标层引入三元组对比正则化,将模式分离约束显式融入大语言模型的训练过程,以降低模型内部易混淆知识间的相关性。覆盖 0.8B~9B 参数规模的大量实验结果表明,该方法能够相对稳定地提升大语言模型的场景认知能力,相较于标准自回归训练,该方法在精确匹配(EM)指标上获得了最高 101.5% 的相对提升,为理解与改进大语言模型的场景认知能力及知识记忆机制提供了新的视角。

关键词 大语言模型;场景认知;模式分离;知识记忆;表示学习

中图法分类号 TP391.1; TP183

DOI: 10.7544/issn1000-1239.202660437 **CSTR:** 32373.14.issn1000-1239.202660437

现如今的大语言模型(large language model, LLM)经过在大规模语料上的预训练后,展现出了令人瞩目的知识记忆能力,并能够在所记忆知识的基础上流畅地回答各种不同领域的问题[1-2]。这种能力源于自回归语言建模目标对海量文本中统计规律的捕捉:模型通过反复预测下一个词,将分散在训练语料中的知识及其间关联编码到高维参数空间中[3-4]。

然而,这一广泛应用的训练目标却导致了大语言模型场景认知能力不足的问题[5]:当多个实体共享同一个语义场景并在其中扮演不同的语义角色时,模型难以准确地根据给定描述从其参数化记忆中召回正确的实体以实现泛化。例如,对于“电影导演 Paxton 向制片人 Helen 和演员 Blake 提出了一个新的电影构想”这一事实,人类能够轻松的回答诸如“向制片人 Helen 和演员 Blake 提出了一个新的电影构想的人是谁”等问题;但对于大语言模型而言,尽管其在训练过程中准确记忆了大量语义相同、仅描述方式不同的事实表述,但在推理阶段,模型仍难以将参数化知识中准确召回目标实体作为回答,最终表现出系统性的泛化失败。

对此,本文深入分析了人类与模型在知识记忆过程中的天然差异,以定位模型场景认知失败的潜在原因。现有研究表明,人脑海马体结构中的齿状回承担着一项核心功能——模式分离[6]:它将人类接收到的高度重叠的感知输入进行正交化处理,将其投射为稀疏且去相关的神经编码,然后再传递至海马体进行存储[7]。正是因为齿状回的存在,人们能够在其所记忆的相同或高度相似的场景下清晰地区分不同个体的行为和特征。然而对于大语言模型而言,其知识记忆阶段所采用的标准自回归训练缺乏一种系统性的机制来主动分离不同实体的隐空间表示[8]。它通过下一词预测目标隐式地塑造其表示空间,但这一目标本身并不包含任何将不同实体表示相互推离的显式引导。

在此基础上,本文提出了一个核心研究假设:大语言模型的标准自回归训练并未使模型形成充分可用的模式分离能力,使得相似但不相同的知识表示间存在混淆,进而导致了其场景认知失败。

为验证这一核心假设,本文提出了一种“先诊断、后干预”的两阶段方法。具体地,本文首先设计了一种跨实体表示诊断(cross-entity representation diagnosis, CERD)方法。该方法在大语言模型隐空间中逐层提取实体名称和描述文本的表示,通过计算跨实体余弦相似度与同类实体自相似度来量化每层实体表示的模式分离程度,以判断其是否通过自回归训练过程隐式实现了模式分离功能。

基于其诊断结果,本文进一步设计了一种隐空间模式分离(latent-space pattern separation, LPS)方法,以显式模拟类似人脑中齿状回所具备的模式分离过程。其核心是在标准自回归训练的基础上,向大语言模型内部实体表示分离相对较差的 Transformer 层引入一个局部的三元组对比正则化信号。该信号强制要求每个实体的自身描述表示与其名称表示锚点之间的余弦相似度必须高于与其他实体描述表示之间的相似度。结合最难负样本挖掘策略,LPS 将形成一种“拉近同类、推远异类”的局部压力,在大语言模型内部模拟了齿状回的模式分离功能。特别地,LPS 仅在训练阶段作为辅助损失存在,不改变模型推理时的行为,也不需要任何外部数据或引入额外参数。

本文对 0.8B~9B 参数规模的大语言模型进行了系统的实验评估。结果表明,在 CERD 的诊断结果引导下,LPS 方法在不同规模的模型中均能有效改善其场景认知能力,进而有效佐证了本文所提出的核心假设的合理性。

总体而言,本文的贡献可归纳为 4 个方面:

1)以大语言模型的场景认知能力为出发点,对比人类与其知识记忆过程中存在的机制性差异,提出了其场景认知能力与模式分离能力间的相关性假设。

2)提出了 CERD 方法，为量化大语言模型隐空间中的模式分离能力提供了一个无需额外训练且可逐层分析的诊断框架。

3)提出了 LPS 方法，一种基于齿状回模式分离的理论启发和 CERD 的诊断结果，为大语言模型提供了一种无需引入外部知识或额外参数的类人知识记忆方法。

4)在跨 0.8B~9B 规模的多个大语言模型上的实验表明，本文的 CERD 方法能够有效测得模型内部的模式分离能力，且在其诊断结果的基础上引入 LPS 干预后，大语言模型的场景认知能力能够被有效提升至远超标准自回归基线的水平。

1 预备知识

1.1 场景认知

场景认知任务借鉴了 Fillmore 所提出的框架语义学理论[9]，是大语言模型将文本语义场景要素与语境中对应论元进行精准关联匹配的核心认知能力。其中语义场景是经由现实经验形成的心智结构，场景要素为构成场景的角色、身份、行为、地点等参与要素，论元则是对应场景要素所指代的具体实体[10]。场景认知本质上要求模型跳出表层文本序列记忆，建立场景要素与具体论元实体之间稳固的语义绑定关系，实现对文本深层角色关系与情境结构的理解。

特别地，如图 1 所示，Ma 等人[5]面向大语言模型场景认知能力评估构建了一套面向虚构事实、标注场景要素的专用场景认知数据集 SBD(scenario based dataset)，整体由原子知识集、记忆集、理解集 3 部分构成。原子知识集采用多模型生成、语义相似度过滤、多模型投票及人工校验协同的构建方案，共计形成 500 条原子知识，且每条原子知识至少包含 3 个不同语义角色。记忆集在原子知识集基础上通过句式变换、同义复述进行语义扩充，使每条原子知识包含 10 种语义一致但表达不同的描述，形成共计 5 000 条样本供大语言模型记忆相关知识。理解集通过人工与大模型协同标注每条文本中的场景要素及对应论元，并基于标注信息构造填空式场景推理问题，最终构建包含 1 581 条问答样本的理解集，用于检验模型是否能够基于所记忆的原子知识泛化至场景要素-论元关联推理中。SBD 数据集全部围绕多角色复杂情境构建，为后续场景认知相关评测与模型能力优化提供了标准数据基准。

现有工作基于 SBD 数据集对大语言模型进行了评估[5]，结果表明，当前大语言模型虽具备极强的训练数据表层复刻记忆能力，但难以形成可靠的场景认知，无法稳定编码场景要素与论元的内在关联，仅依赖表面句式模式拟合而非真正的语义情境理解。这一“场景认知失败”的缺陷正是本文所关注的核心问题。

1.2 海马体中齿状回结构的模式分离功能

哺乳动物大脑的记忆系统由 2 个功能互补的学习系统构成，这一理论框架被称为互补学习系统[11]。新皮层作为慢速学习系统，通过渐进式的突触可塑性变化提取环境中的统计规律，形成稳定的语义知识和认知图式。海马体作为快速学习系统，能够在少量经验甚至单次经历后迅速绑定任意关联。新皮层的慢速学习天然地保护了既有知识的稳定性，而

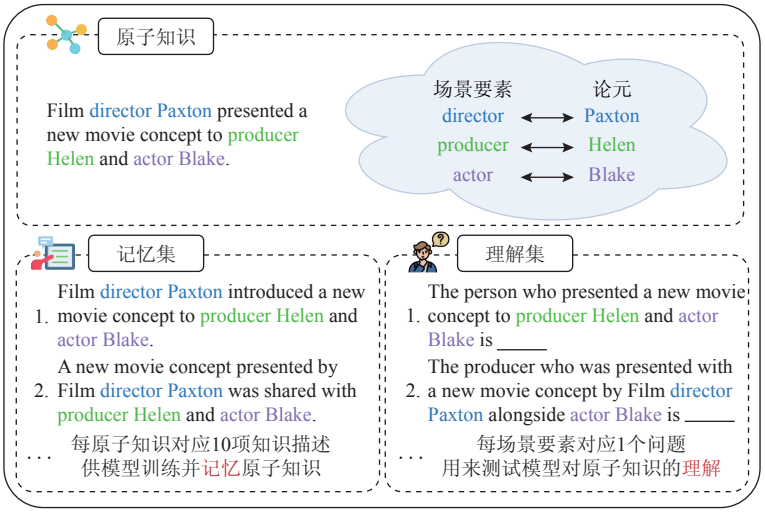


Fig. 1 Examples of the scenario cognition dataset

图 1 场景认知数据集示例

海马体的快速学习确保了对新异信息的敏感性。两者各司其职，共同支撑了人类的记忆与推理能力。

特别地，在海马体内部，齿状回占据了一个独特而重要的位置。来自内嗅皮层的兴奋性投射通过穿通路径首先到达齿状回的颗粒细胞，经过处理后再经由苔藓纤维传递至 CA3 区域的锥体神经元[12]。这一过程赋予了齿状回一项核心功能：模式分离。

模式分离指的是将重叠的、相似的输入模式转化为更加正交、更少重叠的输出模式的过程[6]。内嗅皮层传递到海马体的输入具有高度重叠的特征，例如不同经历可能发生在同一场景中，涉及相似的人或物。如果这些重叠的输入直接进入 CA3 进行存储，它们将形成重叠的吸引子，在回忆时相互干扰，导致相似记忆之间的混淆。齿状回通过其独特的生理特性解决了这一问题：它拥有数量远超内嗅皮层输入神经元数量的颗粒细胞[12]，这一数量优势使其能够进行稀疏编码，用少数高度活跃的神经元来表征每个输入模式，而不同输入模式激活的神经元集合之间重叠极小[13]。此外，齿状回颗粒细胞之间存在强烈的侧抑制，进一步增强了输出编码的去相关性。

大量实验证据支持齿状回的模式分离功能。在啮齿类动物的行为实验中，当需要区分 2 个高度相似的空间位置或物体时，齿状回的失活会导致区分能力的显著下降，而对简单的新异检测则影响较小[14]。在人类的功能性磁共振成像研究中，齿状回区域的激活模式能够预测个体区分相似记忆的成功率[15]。这些发现共同确立了一个共识：齿状回是大脑中实现模式分离的关键结构，它将容易混淆的输入正交化，使海马体能够将不同经历存储为可区分的记忆痕迹。

1.3 自回归训练下的大语言模型模式分离能力局限假设

现阶段主流的大语言模型通常采用标准自回归训练通过下一词预测目标塑造其内部隐空间表示[16]。这一目标函数本质上是统计性的，它最大化训练语料中出现的词序列的似然，隐式地拉近语义相似的上下文在表示空间中的距离。对于语言建模本身而言，这种表示组织方式是合理的，语义相似的词或短语理应在表示空间中相对靠近，这有助于模型泛化到未见过的表达方式。

然而，对于场景认知任务，这种表示组织方式产生了一个根本性的问题。当多个实体共享相同的语义场景、出现在几乎相同的训练上下文时，由于其统计特征高度相似，模型会天然地将这些相似性忠实地编码为表示空间中的邻近性。但当推理阶段要求模型根据细致的角色描述区分这 2 个实体时，这种邻近性就变成了出现混淆并给出错误回答的潜在原因。

对此，本文的核心假设是：大语言模型的标准自回归训练缺乏一个等价于齿状回模式分离功能的显式约束，进而导致模型模式分离能力受限并表现出了其在场景认知上的系统性失败。齿状回的分离是一个显式的、局部化的过程，它不依赖于大脑其他区域的学习信号，而是在输入进入存储系统之前的一个独立的预处理步骤。相比之下，大语言模型的实体表示分离完全依赖于自回归目标通过多层非线性变换和残差连接所产生的隐式效应，但缺乏显式去除不必要相关性的驱动力。倘若该假设成立，则解决方案的设计方向将是明确且清晰的：应当在大语言模型训练过程中引入一个类齿状回的、显式的模式分离约束。

2 跨实体表示诊断

为验证本文所提出的核心假设，本文需要首先回答一个最为基础的问题：在场景认知这一任务背景下，标准自回归训练后的 Transformer 模型中跨实体表示混淆究竟存在于哪些层？混淆状况有多严重？不同参数规模的模型是否呈现出相同的混淆模式？由此可见，形成一个能够逐层量化跨实体表示混淆程度的诊断工具是构建有效干预方法的必要前提。对此，本文提出一种如图 2 所示的 CERD 方法。

2.1 跨实体表示诊断方法

CERD 方法的核心思想是直接测量实体名称表示与实体描述表示之间的跨实体相似度。给定一个包含实体名称与实体描述的、大小为 B 的批次样本 $\{x_i|i\in\{1,2,\cdots,B\}\}$。将这些样本以标准的对话模板格式化后输入模型并获取所有 Transformer 层的隐藏状态。

对于每个 Transformer 层 l，CERD 方法从隐藏状态序列中提取实体段和描述段的表示向量。设该层输出的隐藏状态序列为 $\boldsymbol{H}^l\in\mathbb{R}^{S\times d}$，其中 S 为序列长度，d 为隐藏维度。值得注意的是，由于 CERD 的目标是在不依赖人工标注信息或额外处理的情况下分析模型内部实体表示的分离特性，因此本文不显式利用数据集中的实体边界或角色标注信息，而是采用基于序列位置的无监督表示抽取策略。具体而言，在本文构造的数据样本中，实体名称作为输入提示中的核心查询对象，固定位于描述文本之前。考虑到对话模板前缀以及生成格式可能引入的额外干扰，

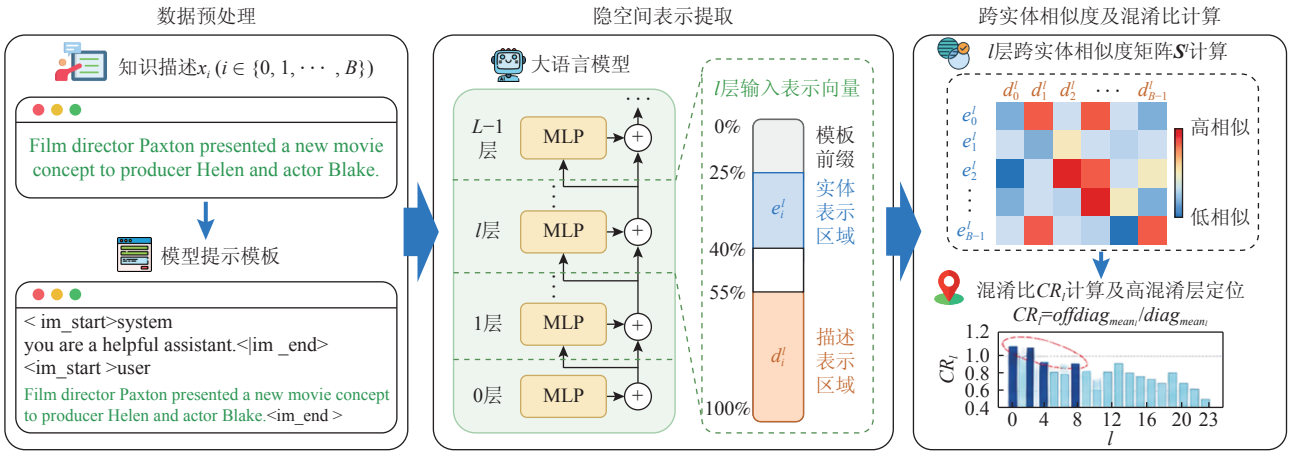


Fig. 2 Illustration of the cross-entity representation diagnosis method

图 2 跨实体表示诊断方法示意图

本文首先排除与系统提示模板相关的无意义头部 token，随后根据实体名称与描述文本在输入序列中的相对分布特征进行区域采样。

具体地，对于隐藏状态序列中靠近实体名称区域的 token，本文选择 25%~40% 的位置区间进行平均池化，作为实体表示 e_i^l；对于描述区域，由于其通常占据输入序列的主要长度范围，因此选 55%~100% 的位置区间进行平均池化作为描述表示 d_i^l。2 个区域之间保留间隔区域，以避免实体结束位置和描述开始位置附近局部语义过渡造成表示污染。

设第 l 层的实体表示矩阵为 $\boldsymbol{E}^l=(e_1^l;e_2^l;\cdots;e_B^l)\in\mathbb{R}^{B\times d}$，描述表示矩阵为 $\boldsymbol{D}^l=(d_1^l;d_2^l;\cdots;d_B^l)\in\mathbb{R}^{B\times d}$。对 2 个矩阵按行进行 L2 归一化后，计算其二者间的相似度：

$$\boldsymbol{S}^l=\hat{\boldsymbol{E}}^l(\hat{\boldsymbol{D}}^l)^{\mathrm{T}}\in\mathbb{R}^{B\times B},$$

其中 $\hat{\boldsymbol{E}}^l$ 和 $\hat{\boldsymbol{D}}^l$ 分别为行归一化后的实体矩阵和描述矩阵。矩阵元素 $S_{ij}^l=\cos(e_i^l,d_j^l)$ 表示实体 i 的名称表示 e_i^l 与实体 j 的描述表示 d_j^l 之间的余弦相似度。对角元素 S_{ii}^l 为同一实体的名称和描述之间的自相似度，非对角元素 $S_{ij}^l(i\neq j)$ 为不同实体名称及描述间的跨实体相似度。

基于相似度矩阵，CERD 方法为每个 Transformer 层 l 计算 3 个统计量。

1）对角相似度均值 $diag_{mean_l}$：

$$diag_{mean_l}=(1/B)\sum_{i=1}^{B}S_{ii}^l。$$

2）非对角相似度均值 $offdiag_{mean_l}$：

$$offdiag_{mean_l}=\frac{1}{B(B-1)}\sum_{i=1}^{B}\sum_{j\neq i}S_{ij}^l。$$

3）混淆比 CR_l：

$$CR_l=offdiag_{mean_l}/diag_{mean_l}。$$

对于特定的 Transformer 层 l，其所对应的混淆比 CR_l 是 CERD 方法所关注的核心指标。$CR_l=1.0$ 表示不同实体之间的平均相似度与同一实体自身的平均相似度持平，即为表示混淆的临界点。$CR_l>1.0$ 则表示跨实体相似度超过了自相似度，意味着在该层中，不同实体的表示比同一实体的不同描述之间的表示更为相似，是严重的跨实体表示混淆。$CR_l<1.0$ 则表示该层具有良好的实体分离能力，即同一实体的不同表示紧密聚集，而不同实体的表示被有效推离。

2.2 诊断结果

如表 1 所示，本文在聚焦场景认知任务的 SBD 数据集上对 Qwen3.5 系列 0.8B、2B、4B 和 9B 四种不

Table 1 Cross-Entity Representation Diagnosis Results

表 1 跨实体表示诊断结果

模型	总层数	$CR_l>1$的层数（占比）	CR_l均值	CR_l最大值（层l）	CR_l
Qwen3.5-0.8B	24	20（83.3%）	1.032 9	1.070 6（l=0）	0.979 1~1.070 6
Qwen3.5-2B	24	16（66.7%）	1.004 6	1.033 7（l=1）	0.967 9~1.033 7
Qwen3.5-4B	32	12（37.5%）	0.980 7	1.016 6（l=17）	0.927 9~1.016 6
Qwen3.5-9B	32	12（37.5%）	0.983 8	1.005 6（l=17）	0.942 1~1.005 6

注：括号内的百分数指的是混淆比 CR_l>1 的层数占总层数的百分比。

同参数规模的模型进行了跨实体表示诊断，诊断结果揭示了一组跨规模一致的模式，也呈现出显著的规模间差异。

整体而言，在 0.8B~9B 参数规模范围内的模型均存在混淆比$CR_l > 1.0$的 Transformer 层，且其整体的混淆比范围均聚集在 1.0 附近。这表明模型内部并未形成类似人脑中齿状回结构的充分可用的模式分离能力，进而导致了跨实体表示混淆，在一定程度上验证了本文所提出的假设。

具体地，从模型规模上看，其混淆比的峰值随规模单调递减，这表明随着模型参数规模的增大，其内参数化知识能够被更加清晰地组织。但同时可以观察到，单纯靠增大模型规模来获得更好的固有表示分离的边际收益在快速衰减：从 0.8B 到 2B，峰值下降了 0.037；从 4B 到 9B，峰值仅下降了 0.011，这表明仅靠堆叠参数量并不能使模型获得充分可用的模式分离能力，因此引入一个显式的模式分离约束进行干预是非常有必要的。

更进一步地，如图 3 和图 4 所示，本文在 SBD 的训练集上对上述大语言模型进行了标准自回归训练以使其记忆相关知识，并可视化了其训练前后的逐层跨实体余弦相似度及混淆比曲线。结果表明，标准自回归训练能够在较深的 Transformer 层中改善原始模型内部的知识表示分布情况，显著提升其跨实体表示间的余弦相似度差异并降低其混淆比，但其在较浅层的作用效果相对有限。同时，结合大语言模型在标准自回归训练后的各模型在场景认知任务上差强人意的表现可知，仅通过标准自回归训练所形成的表示空间改进是不充分的，仍需引入更加显式的干预手段。

值得注意的是，这一可视化结果同时揭示了训练前原始模型混淆比的峰值分布模式随参数规模发生的系统性变化。结合表 2 所示的 Top-5 高混淆层分布情况可见，在 0.8B 模型中，混淆比最高的 5 层集中于浅层（0、1、3、4、8 层），且这一模式在 2B 模型中基本保持一致。2 个规模相对较小的模型的高混淆层均集中在较浅层，对此本文可以合理地推测，当模型容量较小时，浅层的自回归训练信号不足以产生有效的表示分离，而深层由于积累了更多层的非线性变换，分离能力相对较强。然而在 4B 和 9B 两个规模相对较大的模型中，混淆比峰位发生了显著的右移。4B 模型中高混淆比的 Transformer 层转移至 17、16、18、2、5 层，其中相对较深的层占比明显提升，这一现象在 9B 模型中同样存在。这一峰位右移意味着随着模型容量增大，网络能够在更早的层完成初步的语义组织，而真正需要强制分离的瓶颈也随之转移至模型的中部各层。

特别地，为了避免本文所采用的基于序列位置

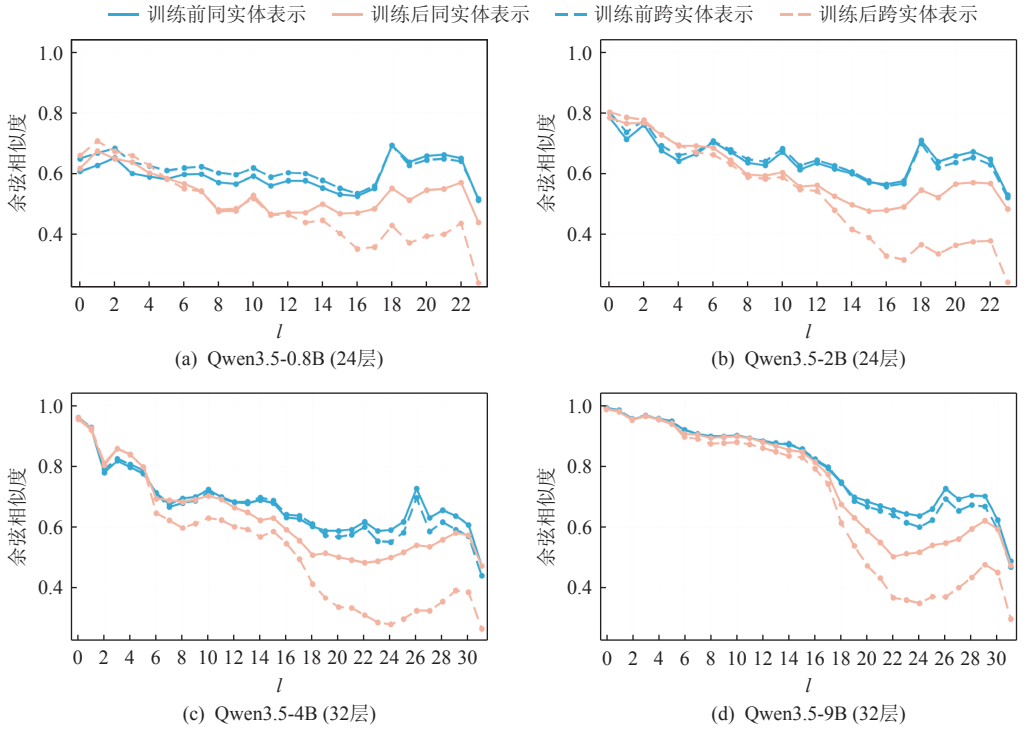


Fig. 3 Layer-wise cosine similarity in cross-entity representation diagnosis

图 3 跨实体表示诊断的逐层余弦相似度

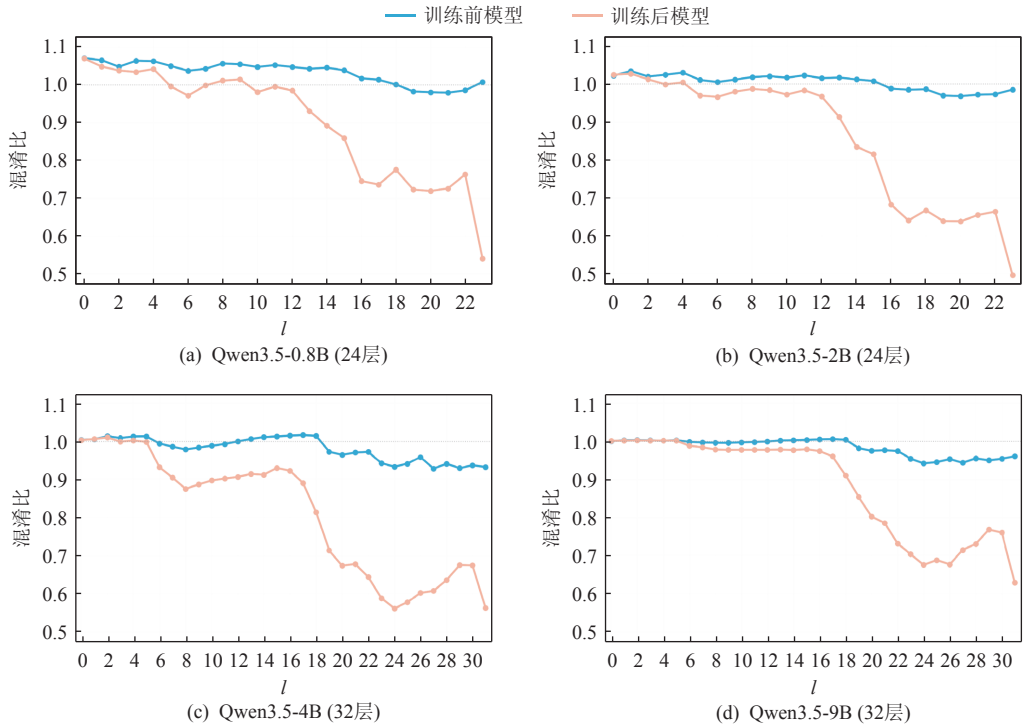


Fig. 4 Confusion ratio in cross-entity representation diagnosis

图 4 跨实体表示诊断的混淆比

Table 2 Distribution of the Top-5 High-Confusion Layers Under Cross-Entity Representation Diagnosis

表 2 跨实体表示诊断下的 Top-5 高混淆层分布

模型	Top-5 高混淆层 l
Qwen3.5-0.8B	0、1、3、4、8
Qwen3.5-2B	1、4、3、11、0
Qwen3.5-4B	17、16、18、2、5
Qwen3.5-9B	17、16、18、15、2

的无监督表示抽取策略对实验结论造成偶然性影响，本文进一步对不同实体表示与描述表示池化区间组合下的 CERD 结果进行了敏感性分析。通过调整实体段和描述段的采样范围，分析不同区域选择对跨实体表示混淆程度测量结果的影响，以验证 CERD 方法对于表示抽取位置变化的鲁棒性。

具体地，在实体表示采样范围方面，通过将起始位置从 20% 逐步右移至 25% 和 30%，并同步使其宽度从 20% 收窄至 15%，检验 CERD 方法对实体出现位置偏移的敏感程度；同时保持 40%~55% 的间隔区域，以排除实体末尾与描述起始之间的语义过渡区对表示提取的污染。此外，本文额外加入 25%~45% 的实体表示采样范围与 60%~100% 的描述表示采样范围组合，同时考察实体范围扩展与间隔区加宽对混淆度量的联合影响，以验证 CERD 在不同表示采样策略下结论的一致性，排除特定采样窗口的偶然性影响。

表 3 表明，CERD 方法对不同的表示抽取策略具有良好的鲁棒性，各模型的 Top-5 高混淆层集合构成在合理的采样区间变化下保持恒定，仅内部排序存在轻微波动。这一结果为 CERD 在无监督设定下使用固定位置比例的合理性提供了支撑，也排除了特定表示采样区间选择对后续 LPS 层消融实验结论的偶然性干扰，一定程度上验证了 CERD 能够无监督地捕获特定表示差异，而非由人为设定的池化区间所驱动。

综上所述，CERD 的诊断结论表明，不同参数规模的大语言模型在场景认知任务中均会在不同程度上表现出内部参数化知识记忆的相关性混淆，其模式分离能力存在一定局限。该结论在一定程度上验证了本文所提出的核心假设，并为后续干预方法的设计提供了支撑。

3 隐空间模式分离方法

基于 CERD 诊断结果，本文进一步设计了一种如图 5 所示的 LPS 方法，通过三元组对比损失对其自回归训练过程进行靶向性干预，以模拟人脑认知

Table 3 Sensitivity Analysis of CERD Under Different Sampling Ranges of Entity and Description Representations

表 3 不同采样范围下实体与描述表示的 CERD 敏感性分析

模型	实体表示采样范围	描述表示采样范围	Top-5 高混淆层 l
Qwen3.5-0.8B	20%~40%	55%~100%	0、1、3、4、8
	25%~40%	55%~100%	0、3、1、4、8
	30%~40%	55%~100%	0、3、1、4、8
	25%~45%	60%~100%	0、1、3、4、8
Qwen3.5-2B	20%~40%	55%~100%	1、4、3、11、0
	25%~40%	55%~100%	1、4、3、0、11
	30%~40%	55%~100%	1、4、3、11、0
	25%~45%	60%~100%	1、4、0、11、3
Qwen3.5-4B	20%~40%	55%~100%	17、16、18、2、5
	25%~40%	55%~100%	17、16、18、2、5
	30%~40%	55%~100%	17、18、16、2、5
	25%~45%	60%~100%	17、16、18、5、2
Qwen3.5-9B	20%~40%	55%~100%	17、16、18、15、2
	25%~40%	55%~100%	17、18、16、15、2
	30%~40%	55%~100%	17、18、16、15、2
	25%~45%	60%~100%	17、16、18、15、2

中的模式分离过程，进而验证该过程是否能够改善大语言模型的场景认知能力。

3.1 三元组对比损失

LPS 方法在标准自回归训练的基础上向目标层集合 L 施加一个局部的三元组对比正则化信号以实现其隐空间中的模式分离。对于训练批次中的每条样本 $i(i=1,\cdots,B)$，LPS 方法在每个目标层 $l\in L$ 上构造对应的三元组。

具体地，三元组的 3 个构成元素均在当前批次内部产生，无需任何外部数据。锚点 a_i^l 为样本 i 在层 l 的实体段表示 e_i^l，通过序列 25%~40% 位置区间的平均池化获得。正样本 p_i^l 为样本 i 在层 l 的描述段表示 d_i^l，通过序列 55%~100% 位置区间的平均池化获得。其区间定义与跨实体表示诊断保持一致。

特别地，由于在同一训练批次中，不同实体可能共享相似的语义角色，因此它们未经正则化的描述段表示可能在隐空间中高度相邻。因此，三元组中的负样本采用了最难负样本挖掘策略进行构建。在同一批次的所有样本中，除样本 i 自身以外，选择描述段表示与锚点 a_i^l 余弦相似度最高的那个描述段表示作为负样本 n_i^l：

$$n_i^l = d_{j^*}^l ,$$

其中 $j^* = \arg\max_{j\neq i} \cos(e_i^l, d_j^l)$。最难负样本挖掘策略是为了避免随机负样本选中已经具有充分分离度的实体，导致其对比信号不产生有效的梯度。最难负样本精确指向当前表示空间中模型最容易混淆的跨实体对，通过在每次训练步骤中对其施加分离约束，将模型有限的容量集中用于处理自回归训练最难以解决的表示混淆。

综上，在每个目标层 l 上，三元组对比损失 L_{trip}^l 定义为：

$$L_{\text{trip}}^l = \frac{1}{B}\sum_{i=1}^{B} \max\left(0, \cos(e_i^l, n_i^l) - \cos(e_i^l, d_i^l) + 0.2\right) ,$$

其中包含了值为 0.2 的相似度阈值。该损失要求锚点实体表示 e_i^l 与其自身描述 d_i^l 间相似度必须比锚点与最难负样本 n_i^l 间相似度高出至少 0.2，这使得该

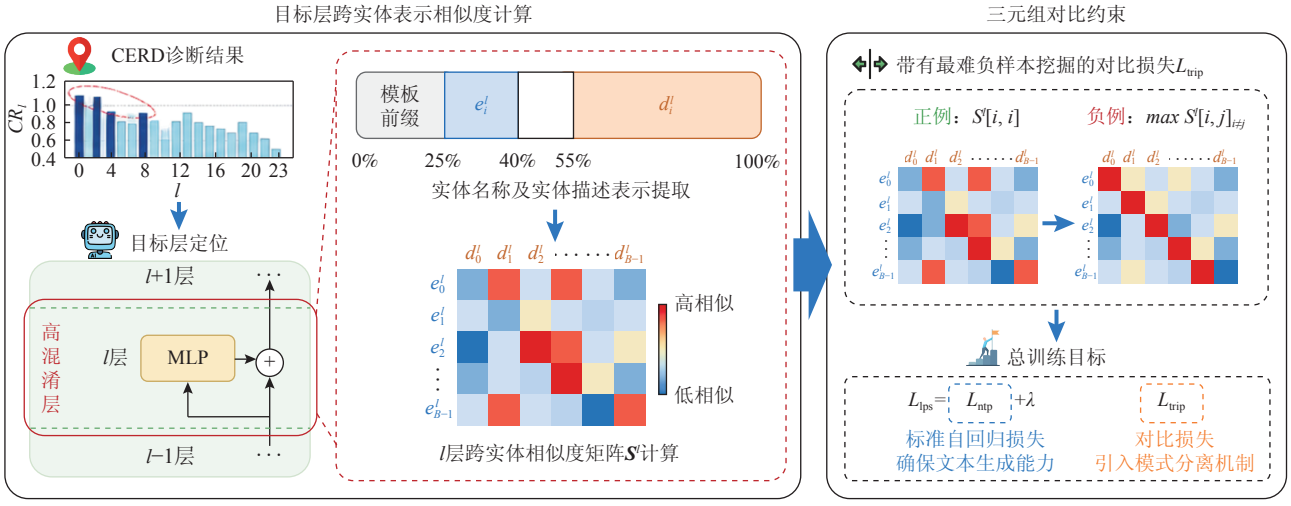


Fig. 5 Illustration of the latent-space pattern separation method

图 5 隐空间模式分离方法示意图

损失函数仅在某个负例实体的描述与锚点的相似度过高且超过阈值时才产生惩罚，以防止对模型内部知识表示的过度干预。

各目标层的损失取平均后作为整体的三元组对比损失 L_{trip}：

$$L_{\text{trip}} = \frac{1}{|L|}\sum_{l\in L} L_{\text{trip}}^{l},$$

其中 L 为选定的目标层数量。

3.2 总体训练目标

LPS 方法的总训练损失 L_{lps} 为标准自回归语言建模损失 L_{ntp} 与三元组对比损失 L_{trip} 的加权和：

$$L_{\text{lps}} = L_{\text{ntp}} + \lambda \cdot L_{\text{trip}},$$

其中 L_{ntp} 为下一词预测的交叉熵损失，λ 为控制三元组对比损失权重系数的超参数，以在隐空间模式分离能力与语言建模质量之间获得平衡。

特别地，本文所设计的 LPS 方法具备以下特点：

1）不替代标准自回归训练。总损失函数为自回归语言建模损失与三元组对比损失之和，两者之间通过一个权重系数进行平衡。在标准前向传播完成后，LPS 方法从目标层的隐藏状态中提取实体段和描述段表示，计算三元组对比损失，将其加权后加到自回归损失上，然后进行反向传播。

2）不引入额外参数。训练完成后，模型结构没有任何改变，无需特殊解码逻辑或额外的推理步骤。LPS 方法在训练阶段注入的分离压力已经通过参数更新固化到了模型的权重中，推理时不再需要任何额外计算，使其与现有的推理框架完全兼容。

3）不依赖外部数据。三元组完全由训练批次内部的样本构造，其锚点和正样本来自同一条数据，负样本来自同一批次中的其他数据，无需额外构造或增强训练数据集。

4 实　　验

与 CERD 诊断阶段的设定相同，本文在场景认知数据集 SBD 上对 Qwen3.5 系列的 0.8B、2B、4B、9B 四个不同规模的大语言模型施加了 LPS 干预，以验证该方法的有效性。实验环境及核心超参数设定如表 4 所示。

具体地，在训练配置方面，所有模型均基于其原始开放权重进行微调，训练数据为 SBD 数据集中的 5 000 条记忆集样本。训练使用 DeepSpeed 并行框架并采用其 ZeRO-3 优化策略，在 bfloat16 混合精度下执行训练。优化器采用线性预热与线性衰减的学习率调度策略，预热比为 0.1，梯度裁剪阈值设为 1.0。

Table 4 Experimental Environment and Core Hyperparameter Settings

表 4 实验环境及核心超参数设定

参数	配置
GPU	4 * NVIDIA A800
批大小	64
学习率	2E−5
预热比	0.1
对比损失权重 λ	0.1

在层选择策略方面，实验过程中共设计了 Top-1、Top-3、Top-5、Bottom-5 及均匀采样 5 种不同的层选择策略。其中，Top-1、Top-3、Top-5 及 Bottom-5 的干预范围由 CERD 诊断结果决定，分别对应混淆比最高的 1 层、3 层、5 层及混淆比最低的 5 层。而均匀采样则完全忽略 CERD 诊断结果，仅依据模型总层数的不同以相同间隔均匀采样 5 层作为对比。

在性能评估方面，训练完成后，本文在 SBD 数据集的 1 582 条理解集样本上评估模型的场景认知能力。评估使用贪婪解码策略，评估指标包括精确匹配（exact match, EM）、包含匹配（contains match, CM）及 BLEU-1 三项。

4.1 主实验结果

标准自回归基线及本文所提出的 LPS 方法在场景认知数据集 SBD 上的实验结果如表 5 所示。其中 EM 用于衡量生成文本与参考文本间的完全精确匹配率，CM 用于衡量二者间的子串包含匹配率，BLEU-1 用来衡量生成文本与参考文本之间的 n-gram 匹配程度。此外，ΔEM、ΔCM 及 ΔBLEU-1 则用来衡量施加 LPS 干预后各项指标相比自回归基线方法的相对增量百分比。

整体而言，本文提出的 LPS 方法在多数的层选择策略及模型规模中都表现出了远超自回归基线的场景认知能力，特别是 Top-1 层选择策略在 0.8B 模型的 EM 指标下获得了 101.5% 的相对提升，一定程度上验证了该方法的有效性。

具体地，从模型规模上看，LPS 方法的收益随模型规模上升而逐渐降低。这一现象并非偶然，而是与本文在 CERD 诊断阶段的结论相互照应。在 CERD 诊断阶段，各模型内部知识混淆比的峰值同样随规模单调递减，这表明随着模型参数规模的增大，其内参数化知识天然能够被更加清晰地组织，因此 LPS

Table 5 Experimental Results of the Standard Autoregressive Baseline and LPS Method on the Scenario Cognition Dataset

表 5 标准自回归基线及 LPS 方法在场景认知数据集上的实验结果

模型名称	干预层	EM/%	CM/%	BLEU-1	ΔEM/%	ΔCM/%	ΔBLEU-1/%
Qwen3.5-0.8B	Top-1	**9.29**	**20.21**	0.165 0	**+101.5**	**+106.2**	+105.7
	Top-3	6.51	15.99	0.131 7	+41.1	+63.2	+64.2
	Top-5	7.71	16.81	0.148 1	+67.1	+71.6	+84.7
	Bottom-5	6.13	13.84	0.115 7	+32.9	+41.3	+44.3
	均匀采样	8.66	20.10	**0.169 0**	+87.9	+105.2	**+110.7**
	自回归基线	4.61	9.80	0.080 2			
Qwen3.5-2B	Top-1	**11.06**	**24.46**	**0.191 7**	**+101.1**	**+95.5**	**+99.9**
	Top-3	9.36	19.34	0.161 3	+70.1	+54.5	+68.2
	Top-5	7.02	14.79	0.121 5	+27.6	+18.2	+26.8
	Bottom-5	6.32	13.78	0.111 5	+14.9	+10.1	+16.3
	均匀采样	8.34	16.81	0.137 0	+51.7	+34.3	+42.9
	自回归基线	5.50	12.52	0.095 9			
Qwen3.5-4B	Top-1	**18.02**	**37.61**	**0.334 5**	**+75.0**	**+21.2**	**+33.9**
	Top-3	14.29	33.19	0.284 4	+38.7	+6.9	+13.9
	Top-5	9.73	32.36	0.247 9	−5.5	+4.3	−0.7
	Bottom-5	4.17	18.20	0.149 6	−59.5	−41.3	−40.1
	均匀采样	12.33	30.66	0.252 5	+19.6	−1.2	+1.1
	自回归基线	10.30	31.04	0.249 8			
Qwen3.5-9B	Top-1	**21.49**	**45.58**	**0.340 6**	**+43.5**	**+20.0**	**+17.5**
	Top-3	19.91	39.70	0.313 2	+32.9	+4.5	+8.1
	Top-5	13.78	34.20	0.274 9	−8.0	−10.0	−5.1
	Bottom-5	7.40	20.10	0.181 8	−50.6	−47.1	−37.3
	均匀采样	14.66	37.23	0.288 0	−2.1	−2.0	−0.6
	自回归基线	14.98	37.99	0.289 8			

注：黑体数值表示最优值。

干预所表现出的收益也会随之下降。但即使在 9B 参数量规模的模型中，相比于标准自回归基线，LPS 方法仍能在 EM 指标上达到 43.5% 的相对提升，这更进一步表明了仅依靠标准自回归训练并不足以充分建立模型内部知识的模式分离，印证了本文所提出的 LPS 方法的必要性。

而从干预层上看，本文所提出的 LPS 方法展现出了一种“少即是优”的整体趋势，仅选择混淆比最高的一层进行干预即可达到最优效果。涉及层数较多的 Top-5、Bottom-5 等层选择策略会引入过多的模式分离约束，在参数化知识容量相对较小的 0.8B 及 2B 模型中表现为场景认知能力的有限提升，而在规模较大的 4B、9B 模型中甚至造成了场景认知能力的损伤。

值得注意的是，本文观察到与 CERD 诊断无关的均匀采样层选择策略在 0.8B~4B 规模的模型上同样带来了不同程度的正向增益，但在 9B 模型中反而引入了性能损失。这一现象表明，大语言模型内部的参数化知识需要被精准干预。这一发现进一步印证了本文所提出的“先诊断、再干预”的两阶段框架是非常有必要的，CERD 诊断能够为 LPS 方法精确定位亟需干预的 Transformer 层，进而避免均匀采样等层选择策略略引入模糊且不可解释的干预位置，防止由于改动不够精准而损害模型自身能力。

特别地，本文在 0、0.01、0.05、0.1、0.5 及 0.8 六项不同取值上对比了不同的对比损失权重 λ 下 LPS 方法在干预 Top-1 层时的性能表现，其中 $\lambda=0$ 时 LPS 方法退化为标准自回归训练。表 6 表明，以任何 $\lambda>0$ 权重引入三元组对比损失都能够在一定程度上提升其相对于标准自回归训练的性能，且当 $\lambda=0.1$ 时能够达到相对最高的性能收益。具体而言，当 λ 相对较小时，其约束信号会被严重稀释，无法有效在

Table 6 Performance Comparison of LPS Method Under Different Contrastive Loss Weight λ

表 6 LPS 方法在不同对比损失权重 λ 下的性能对比

λ	EM/%	CM/%	BLEU-1
0	4.61	9.80	0.080 2
0.01	4.67	11.06	0.091 4
0.05	7.40	16.37	0.115 1
0.1	**9.29**	**20.21**	**0.165 0**
0.5	8.97	20.10	0.161 6
0.8	8.08	19.26	0.154 7

注：黑体数值表示最优值。

模型内部建立模式分离，其收益相对较小；当 λ 过大时，其约束信号会对模型内部的参数化知识表示产生过度影响以降低其收益。

4.2 混淆比分析

为验证 LPS 方法所带来的场景认知性能增益与其内部知识表示混淆比间的相关性，本文在 LPS 方法干预后的模型上进行了与 CERD 诊断阶段设定相同的混淆比分析，并与标准自回归基线训练方法进行了对比。如图 6 所示，相较于标准自回归方法，受益于大语言模型内部各层间的残差连接，LPS 方法在各种局部层采样策略下均可降低大语言模型的全局混淆比，这一整体趋势表明本文所提出的 LPS 方法在一定程度上提升了大语言模型对其内部参数化知识的模式分离能力。

具体地，如图 7 所示，对比 LPS 方法在不同层选择策略下相比标准自回归方法的混淆比下降量可以发现，相较于标准自回归训练方法，在 Top-1 层施加干预的 LPS 方法能够在 0.8B、2B 模型的较浅层位置及 4B、9B 模型的较深层位置显著降低其混淆比，这些具有显著改善的 Transformer 层分布情况同样与 CERD 诊断结果中的高混淆层分布情况一致，即采用 Top-1 层选择策略的 LPS 方法能够有效对不同规模模型在 CERD 诊断中所发现的高混淆层进行干预并增强其模式分离能力。

特别地，本文对比了不同层选择策略下 LPS 方法的各层混淆比下降量均值与 EM 变化量间关联。如表 7 所示，在场景认知任务的核心 EM 指标上效果

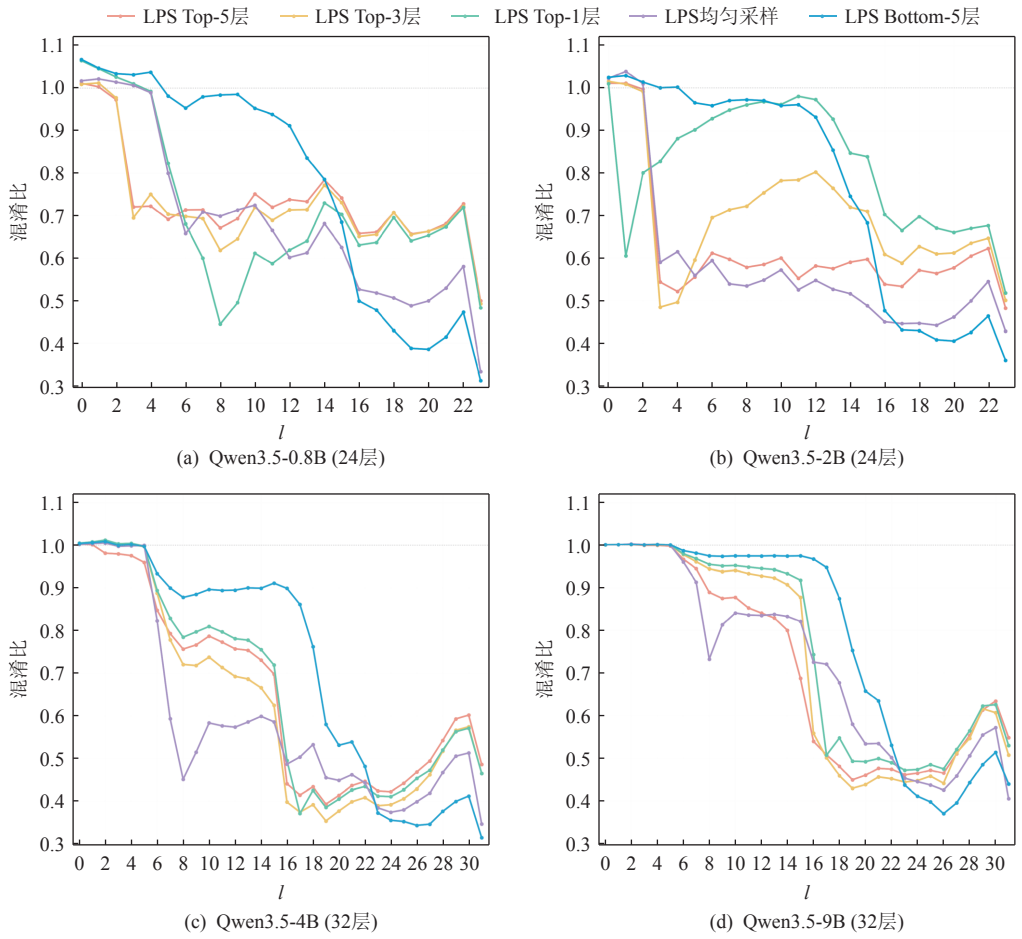


Fig. 6 Layer-wise internal knowledge confusion ratios of models trained with the standard autoregressive baseline and LPS on SBD

图 6 在场景认知数据集上通过标准自回归基线与 LPS 方法训练后各模型内部知识的逐层混淆比

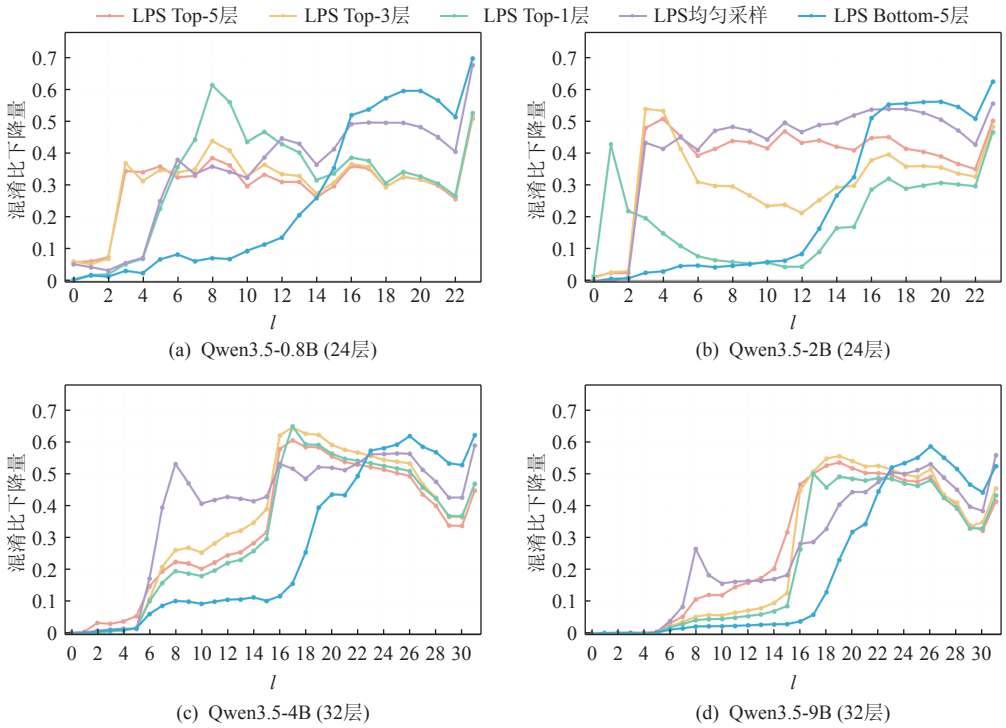


Fig. 7 Confusion ratio reduction of LPS relative to the standard autoregressive method under different layer selection strategies

图 7 LPS 方法在不同层选择策略下相比标准自回归方法的混淆比下降量

最佳的 Top-1 层选择策略并非总是在最大程度上降低了模型内部混淆比，而是引入了一种相对轻量的模式分离能力。本文认为，该现象在一定程度上解释了 LPS 方法展现出的“少即是优”的整体趋势：大语言模型内部的参数仅能接受相对适量的干预，若采用 Top-3、Top-5 等范围较大的层选择策略，则会引入过大的表示干预。过度提升不同知识间的区分度，虽然其可能引入更强的表示分离能力、大幅降低混淆比，但任务性能并不会随之单调提升，甚至可能破坏其本身应当具备的知识相关性并损害模型自身性能。该现象也进一步揭示了生物启发机制在大语言模型中的迁移并非简单复现人脑功能。尽管 LPS 的设计动机源于人脑齿状回所具备的近乎正交的强模式分离能力，但大语言模型内部知识表示具有不同于生物神经系统的组织方式，过强的模式分离反而可能削弱已有语义关联的保持能力。因此，对于大语言模型而言，理想的模式分离并非最大化地增强表示区分度，而是在知识保持与表示解耦之间寻求适度平衡。

5 相关工作

5.1 大语言模型的认知能力

近年来，大量工作致力于评估和理解大语言模型的认知能力边界。Mahowald 等人[17] 区分了形式语言能力与功能性语言能力，指出大语言模型虽在语言形式层面（如句法、形态）表现优异，但在依赖世界知识和逻辑推理的功能性语言使用中仍存在显著不足。Niu 等人[18] 的研究表明，大语言模型具有类似人类的语言处理能力，然而在面对新颖提示时其推理能力存在明显缺陷。Binz 和 Schulz[19] 发现 GPT-3 在因果推理与审慎思考任务上的表现有限，表明其泛化能力高度受限于训练数据分布。Ullman[20] 通过修改后的心理理论任务展示了大语言模型在心智建模方面的系统失败，在任务设置发生微小变化时模型表现急剧下降，进一步揭示了其认知模型的脆弱性。Blank[21] 强调了评估大语言模型认知能力时的方法论陷阱，倡导采用严格基于语言的任务设计。Berglund 等人[22] 发现了“逆转诅咒”现象，当大语言模型在训练中学习到“A 是 B”这一事实后，无法自动推出“B 是 A”，揭示了其高度依赖于从输入到输出的单向表面关联，而非对实体间语义关系的认知。与上述从通用认知能力或简单事实知识出发的研究不同，Ma 等人[5] 聚焦于场景认知这一特定任务，并发现了当前大语言模型存在的场景认知能力不足的缺陷，揭示了大语言模型主要依赖表层记忆而非深层语义理解，但并没有相关工作对该缺陷进行进一步分析和改进。

上述研究表明，大语言模型在知识记忆的组织和召回机制上存在根本性缺陷。然而，现有工作尚

Table 7 Correlation Between Confusion Ratio and EM Variation of the LPS Method Under Different Layer Selection Strategies

表 7 不同层选择策略下 LPS 方法的混淆比与 EM 变化量间关联

模型	LPS 层选择策略	ΔCR_{mean}	ΔEM/%
Qwen3.5-0.8B	Top-1	−0.183	**+101.5**
	Top-3	−0.176	+41.1
	Top-5	−0.165	+67.1
	Bottom-5	−0.126	+32.9
	均匀采样	**-0.212**	+87.9
Qwen3.5-2B	Top-1	−0.038	**+101.1**
	Top-3	−0.152	+70.1
	Top-5	−0.230	+27.6
	Bottom-5	−0.087	+14.9
	均匀采样	**-0.274**	+51.7
Qwen3.5-4B	Top-1	−0.147	**+75.0**
	Top-3	−0.180	+38.7
	Top-5	−0.153	−5.5
	Bottom-5	−0.093	−59.5
	均匀采样	**-0.215**	+19.6
Qwen3.5-9B	Top-1	−0.129	**+43.5**
	Top-3	−0.155	+32.9
	Top-5	**-0.172**	−8.0
	Bottom-5	−0.098	−50.6
	均匀采样	−0.163	−2.1

注：黑体数值表示最优值。ΔCR 越小越好，ΔEM 越大越好。

未从底层表示机制的角度提出系统性的诊断框架和训练阶段的改进方法。本文正是在这一背景下，受人脑海马体齿状回模式分离机制的启发，探索大语言模型内部跨实体知识表示的混淆问题及其解决方案。

5.2 大语言模型内部知识表示分析

理解大语言模型内部如何存储和组织知识，是诊断和改进其认知能力的前提。近年来，研究者发展了一系列探针（probe）和表示分析方法来揭示 Transformer 模型内部的表示结构。Petroni 等人[23]通过 LAMA 探针系统地展示了预训练语言模型参数中存储了可被“查询”的事实性知识。Meng 等人[24]提出了 ROME 方法，通过因果中介分析定位了 GPT 模型中与特定事实性知识相关的中间层 MLP 模块，并展示了对这些层的精确编辑可以定向改变模型的事实性输出。Geva 等人[25]进一步揭示了 Transformer 的前馈网络层可以被理解为以键值记忆（key-value memory）的形式存储知识，前馈网络的第 1 层产生键向量用于模式匹配，第 2 层产生值向量用于输出分布。Dai 等人[26]提出了“知识神经元”的概念，识别出预训练 Transformer 中负责存储特定事实性知识的稀疏神经元集合。这些工作深刻揭示了大规模参数化知识存储的微观机制，表明大语言模型的知识虽具有可定位性和可编辑性，但其组织方式可能天然存在结构性问题。

与上述工作不同，本文提出的 CERD 方法并不关注单个知识片段的定位、编辑或不同模型间的表示比较，而是从模式分离的独特视角出发，逐层量化不同实体的名称表示与描述表示之间的跨实体混淆程度。CERD 通过在无需额外训练的条件下计算混淆比，为大规模参数化知识组织的整体质量提供了一个系统性、可逐层分析的诊断框架。这一诊断框架直接服务于本文的核心假设验证，并为后续靶向性干预方法的设计提供了定量依据。

5.3 神经科学启发的表示学习方法

本文在大语言模型训练过程中引入类似人脑齿状回的模式分离机制的这一核心思想受到计算神经科学中互补学习系统（CLS）理论[27]的直接启发。CLS 理论指出，哺乳动物大脑演化出了 2 个功能互补的记忆系统：海马体作为快速学习系统，通过稀疏且去相关的表征在少量经历后迅速绑定任意关联，有效避免了灾难性干扰；新皮层则作为慢速学习系统，通过渐进式的、交错的突触可塑性变化提取环境中的统计规律，形成稳定的语义知识。Kumaran 等人[11]在 2016 年更新了这一理论框架，将其与现代深度神经网络中的参数化学习与非参数化记忆进行了明确对接，并提出了海马体回放机制可以为新皮层提供目标依赖的经验重加权。在计算层面，Marr[6]提出了海马体通过正交化编码实现模式分离的理论构想，即相似的输入模式在进入海马体 CA3 进行存储之前，应当首先被转化为更正交、更少重叠的编码。O’Reilly 和 McClelland 等人[27]进一步揭示了内嗅皮层、齿状回、CA3 的多层前馈通路如何通过逐层变换增强模式分离能力，并指出了齿状回苔藓纤维的强输入方差在这一过程中的关键作用。在实验神经科学层面，Leutgeb 等人[13]通过大鼠在体记录直接展示了齿状回颗粒细胞在渐变环境中的速率重映射（rate remapping）编码，即使空间布局不变，DG 细胞仍对微小的感觉输入变化产生显著的活动模式变化。McHugh 等人[14]通过 DG-NR1 基因敲除小鼠提供了因果性证据：当齿状回颗粒细胞的 NMDA 受体被选择性敲除后，小鼠在相似环境区分任务中表现出显著缺陷，而

对差异明显的环境区分则不受影响。

将神经科学原理应用于人工学习系统的设计是一个长期的研究方向。在深度学习时代,这一交叉领域已产生了多种有影响力的方法,例如受海马体回放启发的经验重放(experience replay)[28]、受突触可塑性启发的元学习(meta-learning)方法[29]等。然而,目前尚没有工作将齿状回模式分离的概念系统性地应用于大语言模型的训练过程中。现有的大语言模型训练方法主要依赖标准自回归目标并通过最大化训练语料中词序列的似然来隐式地塑造其内部表示空间,但这一目标本身并不包含任何显式的机制来主动分离不同实体在高维隐空间中的表示[16]。因此,本文提出的LPS方法通过在训练过程中引入三元组对比正则化,在CERD诊断所定位的高混淆Transformer层上靶向性地模拟类似齿状回的模式分离过程,在保持语言建模能力的前提下,靶向性地向特定的Transformer层注入局部的模式分离压力,通过参数更新将其固化到模型的权重中以优化其内部表示分布。

6 结 论

本文聚焦于大语言模型在场景认知任务上表现出的系统性泛化失败这一核心问题,受人脑海马体中的齿状回结构所具备的模式分离机制的启发,提出了场景认知失败与模型内部模式分离能力缺失之间的相关性假设,并提出了一种“先诊断、后干预”的两阶段方法来验证该假设并提升大语言模型的场景认知能力。在诊断阶段,本文设计了CERD方法,通过计算模型内部跨实体知识表示间的混淆比来分析模型逐层的模式分离能力,进而验证了在标准自回归训练范式下大语言模型存在的模式分离能力受限问题。在干预阶段,本文设计了LPS方法,在经CERD诊断发现的高混淆比层上针对性地向标准自回归训练过程引入三元组对比损失来显式建立跨实体表示间的模式分离约束,在不同规模的大语言模型上均显著提升了其场景认知能力,并降低了模型内部跨实体表示间的混淆比,最终验证了本文所提出假设的合理性及所提出方法的有效性,为理解与改进大语言模型的场景认知能力及知识记忆机制提供了新的视角。

然而,本文所提出的方法仅在大语言模型的损失函数中显式引入了类似人脑中齿状回结构的模式分离约束,虽然并未引入额外参数,但同样也限制了该方法的约束强度,难以形成更加细致且灵活的模式分离过程,且实验所覆盖的数据集和模型架构相对有限。对此,未来工作将从以下若干方面展开:1)构建针对性的、用于形成模式分离的网络结构,以形成更加灵活的知识表示映射过程;2)在更广泛的数据集上评估并优化LPS方法,为大语言模型建立更加鲁棒的模式分离机制;3)扩大模型规模及模型系列的覆盖范围,并发现其在知识记忆过程中存在的共性问题,进而从机制上形成一种即插即用、不受模型架构及参数规模约束的知识记忆优化方法。

作者贡献声明:马博翔负责方案细化、实验设计与实现及论文撰写;李茹负责方案整体设计,并从认知科学视角提供指导;郭少茹提供指导并负责论文修改与完善;李英豪负责实验结果分析;Víctor Gutiérrez-Basulto参与论文核心思路讨论并指导改进方案设计。

参 考 文 献

[1] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019, 1(8): 1−24

[2] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[C]//Proc of the 34th Advances in Neural Information Processing Systems (NeurIPS). Cambridge, MA: MIT, 2020: 1877−1901

[3] Chen Huimin, Liu Zhiyuan, Sun Maosong. The social opportunities and challenges in the era of large language models[J]. Journal of Computer Research and Development, 2024, 61(5): 1094−1103 (in Chinese)
(陈慧敏, 刘知远, 孙茂松. 大语言模型时代的社会机遇与挑战[J]. 计算机研究与发展, 2024, 61(5): 1094−1103)

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C/OL]//Proc of the 31st Advances in Neural Information Processing Systems (NeurIPS). Long Beach, CA: NIPS, 2017[2026-07-15]. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[5] Ma Boxiang, Li Ru, Wang Yuanlong, et al. Memorization ≠ understanding: Do large language models have the ability of scenario cognition?[C]//Proc of the Conf on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: ACL, 2025: 20747–20763

[6] Marr D. Simple memory: A theory for archicortex[J]. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 1971, 262(841): 23−81

[7] Yassa M A, Stark C E L. Pattern separation in the hippocampus[J]. Trends in Neurosciences, 2011, 34(10): 515−525

[8] McClelland J L, McNaughton B L, O'Reilly R C. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory[J]. Psychological Review, 1995, 102(3): 419−457

[9] Fillmore C J. Frame semantics[J]. Cognitive Linguistics: Basic Readings, 2006, 34: 373−400

[10] Su Xuefeng, Li Ru, Ma Boxiang, et al. Knowledge-guided frame-semantic parsing via LLM[J]. Journal of Chinese Information Processing, 2026, 40(3): 27−40 (in Chinese)
(苏雪峰, 李茹, 马博翔, 等. 知识指导的大语言模型框架语义分析方法[J]. 中文信息学报, 2026, 40(3): 27−40)

[11] Kumaran D, Hassabis D, McClelland J L. What learning systems do intelligent agents need? complementary learning systems theory updated[J]. Trends in Cognitive Sciences, 2016, 20(7): 512−534

[12] Amaral D G, Scharfman H E, Lavenex P. The dentate gyrus: Fundamental neuroanatomical organization (dentate gyrus for dummies)[J]. Progress in Brain Research, 2007, 163: 3−22

[13] Leutgeb J K, Leutgeb S, Moser M B, et al. Pattern separation in the dentate gyrus and CA3 of the hippocampus[J]. Science, 2007, 315(5814): 961−966

[14] McHugh T J, Jones M W, Quinn J J, et al. Dentate gyrus NMDA receptors mediate rapid pattern separation in the hippocampal network[J]. Science, 2007, 317(5834): 94−99

[15] Lacy J W, Yassa M A, Stark S M, et al. Distinct pattern separation related transfer functions in human CA3/dentate and ca1 revealed using high-resolution FMRI and variable mnemonic similarity[J]. Learning & Memory, 2011, 18(1): 15−18

[16] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J/OL]. OpenAI Technical Report, 2018 [2026-06-01]. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[17] Mahowald K, Ivanova A A, Blank I A, et al. Dissociating language and thought in large language models[J]. Trends in Cognitive Sciences, 2024, 28(6): 517−540

[18] Niu Qian, Liu Junyu, Bi Ziqian, et al. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges[J]. arXiv preprint, arXiv: 2409.02387, 2024

[19] Binz M, Schulz E. Using cognitive psychology to understand GPT-3[J]. Proceedings of the National Academy of Sciences, 2023, 120(6): e2218523120

[20] Ullman T. Large language models fail on trivial alterations to theory-of-mind tasks[J]. arXiv preprint, arXiv: 2302.08399, 2023

[21] Blank I A. What are large language models supposed to model?[J]. Trends in Cognitive Sciences, 2023, 27(11): 987−989

[22] Berglund L, Tong M, Kaufmann M, et al. The reversal curse: LLMs trained on “a is b” fail to learn “b is a”[C]//Proc of the 12th Int Conf on Learning Representations (ICLR 2024). Washington DC: ICLR, 2024: 18623−18642

[23] Petroni F, Rocktäschel T, Riedel S, et al. Language models as knowledge bases?[C]//Proc of the 2019 Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: ACL, 2019: 2463−2473

[24] Meng K, Bau D, Andonian A, et al. Locating and editing factual associations in GPT[C]//Proc of the 36th Advances in Neural Information Processing Systems (NeurIPS). Long Beach, CA: NIPS, 2022: 17359−17372

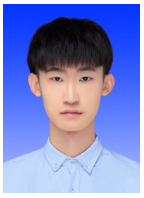
[25] Geva M, Schuster R, Berant J, et al. Transformer feed-forward layers are key-value memories[C]//Proc of the Conf on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: ACL, 2021: 5484−5495

[26] Dai Damai, Dong Li, Hao Yaru, et al. Knowledge neurons in pretrained transformers[C]//Proc of the 60th Annual Meeting of the Association for Computational Linguistics (ACL). Stroudsburg, PA: ACL, 2022: 8493−8502

[27] O’Reilly R C, McClelland J L. Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off[J]. Hippocampus, 1994, 4(6): 661−682

[28] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529−533

[29] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//Proc of the 34th Int Conf on Machine Learning (ICML). New York: ACM, 2017: 1126−1135



Ma Boxiang, born in 2000. PhD candidate. His main research interests include natural language processing and language cognition.

马博翔, 2000 年生。博士研究生。主要研究方向为自然语言处理、语言认知。



Li Ru, born in 1963. Professor, PhD supervisor. Her main research interests include natural language processing and information retrieval.

李　茹, 1963 年生。教授, 博士生导师。主要研究方向为自然语言处理、信息检索。



Guo Shaoru, born in 1990. Lecturer, master supervisor. Her main research interests include natural language processing and knowledge engineering.

郭少茹, 1990 年生。讲师, 硕士生导师。主要研究方向为自然语言处理、知识工程。



Li Yinghao, born in 1994. Lecturer. His main research interests include in-context learning and large language model reasoning.

李英豪, 1994 年生。讲师。主要研究方向为上下文学习、大模型推理。



Víctor Gutiérrez-Basulto, born in 1981. Associate professor, PhD supervisor. His main research interests include knowledge representation and reasoning, representation learning, natural language processing and data management.

Víctor Gutiérrez-Basulto, 1981 年生。副教授, 博士生导师。主要研究方向为知识表示与推理、表示学习、自然语言处理、数据管理。