通过布鲁姆分类法和线性探测，解析大语言模型中认知复杂度的机制可解释性

Bianca Raimondi意大利博洛尼亚大学
bianca.raimondi3@unibo.it

Maurizio Gabbrielli意大利博洛尼亚大学
maurizio.gabbrielli@unibo.it

arXiv:2602.17229v1 [cs.AI] 19 Feb 2026

摘要

大语言模型的黑箱特性，使我们亟须开发超越表层性能指标的新型评估框架。本研究以布鲁姆分类法为层次化视角，考察认知复杂度的内部神经表征。通过分析不同大语言模型的高维激活向量，我们探测不同认知层级（从基础召回（记忆）到抽象综合（创造））是否能在模型的残差流中线性分离。结果表明，线性分类器在所有布鲁姆层级上的平均准确率约为 95%，有力证明认知层级编码于模型表征的一个线性可访问子空间中。这些发现表明，模型会在前向传播的早期阶段解析提示的认知难度，且随着层数加深，表征的可分性逐渐增强。

1 引言

基于 Transformer 的大语言模型（LLM）的兴起，从根本上改变了人工智能领域，催生出一些系统，它们展现出的行为曾一度被认为是人类认知所独有的（Vaswani 等人，2017年；Brown 等人，2020年）。这些模型展现出多样的能力，挑战了我们对机器理解的定义：从生成连贯的散文到求解复杂的数学定理。然而，这种功能性能力本质上仍是一个黑箱：促成如此高阶推理的具体内部机制依然不为人所知（Olah 等人，2020年）。

这一可解释性挑战中一个至关重要却尚未得到充分探索的维度，是对认知深度的评估。当前基准测试通常依据答案正确性来评估 Hendrycks 等人，2020年）。尽管这类指标很有用，却无法区分正确答案究竟是源于死记硬背，还是来自复杂的多步推理。在人类认知中，这一区别具有基础性意义。能够正确背出某个历史日期的学生，与分析某个历史事件的社会政治原因的学生，运用的是不同的认知能力。布鲁姆分类法将这种认知技能层级系统化为一个框架，把教育目标划分为六个层次：记忆，理解，应用，分析，评价和创造（Krathwohl，2002年）。

如果大语言模型是在近似通用智能，那么其内部架构至少应在一定程度上反映不同程度的认知要求。当 Transformer 执行分析或评估等高阶操作时，预期会出现不同形式的神经活动；这与事实回忆等低阶过程截然不同。若存在与不同认知层级相关的专门内部表征，就表明模型已学会一种关于问题难度和任务类型的泛化概念，并将其抽象于特定领域内容之外。

本文从布鲁姆分类法的视角，对大语言模型的内部表征开展机制性研究，并探讨以下研究问题及相应发现：

- **研究问题1.** 大语言模型在处理与不同布鲁姆层级相关的提示时，其内部激活是否表现出差异?
- **研究问题2.** 认知复杂度的表征在网络的何种深度出现?
- **研究问题3.** 布鲁姆层级信息能否从隐藏状态中线性恢复，且由此产生的分类错误是否反映该分类法的序结构与语义结构?

为了解答这些问题，我们从四个最先进的开放权重大语言模型的残差流中提取激活向量，



图1：实验流程概览，从数据集构建和激活提取到逐层线性探测。

并在每一层训练线性探针，以对布鲁姆层级进行分类。

主要发现可概括如下。首先，线性分类器达到了 ∼ 95% 的准确率，表明认知层级被编码在模型内部表征的线性可访问子空间中。其次，认知复杂度在处理的早期至中期阶段出现，而不是延迟到网络最深层。最后，混淆矩阵分析显示，误分类主要发生在相邻的布鲁姆层级之间。这种结构化的错误模式表明，学习到的表征反映了人类教育理论的序数组织结构，而非任意的标签边界。

2 相关工作

尽管基于 Transformer 的大语言模型在经验上取得了成功（Vaswani 等人，2017年；Brown 等人，2020年），它们仍然难以解释，尤其是在支撑复杂推理的内部机制方面（Olah 等人，2020年）。理解这些内部机制已成为机制可解释性（MI）的核心关注点，其目标是通过分析神经网络的内部表征和计算结构来对其进行逆向工程（Elhage 等人，2021）。近期的综合性综述（Rai 等人，2024年；Bereska 和 Gavves，2024年；Shu 等人，2025年；Zhao 等人，2024年；Gantla，2025年）已将机制可解释性确立为人工智能安全的关键框架，强调需要超越表层性能指标，进而理解支撑模型行为的实际计算过程。

一些研究已将机制可解释性的研究范围拓展至涵盖语义概念（Cywi´nski 等人，2025年）和偏差（Raimondi 等人，2025年；Simbeck 和 Mahran，2025年）。这些研究表明，现代机制可解释性能够有效探究高层次认知构念，这也支持了我们将布鲁姆分类法用作评估框架的做法。

此前对大语言模型的评估主要依赖基准测试准确率，包括大规模多任务语言理解等数据集（Hendrycks 等人，2020年）。然而，此类基准测试主要评估任务表现，并不能直接探究底层认知表征的结构。这一局限促使研究者采用布鲁姆分类法（Krathwohl，2002年），将其作为评估人工智能系统中认知复杂度的层级框架。近期研究已证明该分类法在教育应用中的效用（Herrmann-Werner 等人，2024年；Elkins 等人，2024年；Kumar 等人，2025年），并提出了面向人工智能时代的扩展（Hmoud 和 Ali，2024年；Luo 等人，2025b年）。与此最直接相关的是通过布鲁姆视角对大语言模型进行系统性认知评估的研究（Huber 和 Niklaus，2025年；Yu 等人，2025年），不过其关注的是行为输出而非内部表征。

我们的工作推动了这一领域的发展，将研究重点从行为输出转向内部机制分析。尽管线性探测已成为研究线性可解码特征的基础方法（Be-linkov，2022；Kim 等人，2025），近期进展开始探索更高层次的概念表征（Kim 等人，2018；Luo 等人，2025a）。新兴证据表明，大语言模型会自发地组织表征，以编码抽象的任务属性（Jankowski 等人，2025；Benito-Rodriguez

模型	L	d_m
Llama-3.1-8B-Instruct	32	4096
Qwen3-4B-Instruct-2507	36	2560
gemma-3-4b-it	35	2560
DeepSeek-R1-Distill-Llama-8B	32	4096

表 1：本研究使用的模型。

等，2025年；Jin 等，2025年）。

我们研究的一项关键进展，是发现模型内部表征中存在一个结构性的认知可分性起始点(CSO)。不同于以往以输出为中心的评估（Huber 和 Niklaus，2025年；Budagam 等人，2024年；Zoumpoulidi 等人，2025年），我们的方法直接提供了认知复杂度如何编码的机制证据。通过结合互信息技术（Rai 等人，2024年；Bereska 和 Gavves，2024年）与布鲁姆分类法，我们证明模型对认知复杂度的辨析远超表层词汇线索，在前向传播早期便实现近乎完美的探针准确率（$\sim$ 95%），并揭示大语言模型会自发地组织其潜在空间，以编码层级化的认知要求（Hewitt 和 Liang，2019）。

3 方法论

我们的实验框架[1] 旨在系统地提取并分类图 1 所示的四种不同大语言模型的内部激活1。

3.1 问题表述

设 $\mathcal{D} = \{(q_i,\ y_i)\}_{i=1}^{N}$ 是一个自然语言问题数据集，其中 q_i 表示提示，$y_i \in \{0, \ldots, 5\}$ 表示其对应的布鲁姆层级。给定一个具有 L 层和隐藏维度 d_m 的模型，我们研究认知层级 y_i 是否可从内部表征中线性解码。

3.2 模型

为确保观察到的结果并非特定于单一架构，我们评估了 HuggingFace 上提供的多种开放权重大语言模型[2]（参见图 1）。所选模型在参数数量、深度、架构设计和训练数据规模方面各不相同。

对于每个模型，我们在前向传播过程中提取每个Transformer层的残差流激活值。

[1]我们的代码位于：匿名评审期间省略链接[2]

[2] https://huggingface.co

布鲁姆层级	#样本
0（记忆）	188
1（理解）	188
2（应用）	188
3（分析）	188
4（评价）	188
5（创造）	188
总计	**1128**

表2：样本分布。

所有模型均在推理模式下进行评价，且未执行任何微调。

3.3 数据集

我们的分析以一个经过精心筛选、均衡的数据集为基础，其中包含 1,128 道标注了布鲁姆分类法等级的问题（表2）。每个提示都对应一个明确的认知复杂度等级。最终语料库由两个不同的教育数据集汇总而成：

- **计算机科学课程查询：** Zaman等人(2024年) 提供的数据集包含计算机科学课程（数据结构、计算机导论和研究）的学生查询。这些查询根据认知难度进行独特评分，从而细致地映射到布鲁姆分类法的各个等级。

- **EduQG：** 为增强任务的多样性，我们纳入了来自 *EduQG*(Hadi-far等人，2023）的样本。该数据集提供多种格式、专为教育领域设计的选择题，并附有经专家验证的布鲁姆分类法标注。

为避免类别不平衡导致线性探针产生架构偏差，我们对汇总数据进行了受控下采样。这确保探针的准确率反映的是模型的表征能力，而非标签分布的统计先验。此外，问题的长度从 3 到 39 个词不等（$\mu = 12.95, \sigma = 5.52$），确保了语言复杂性的多样性。

布鲁姆层级	d_0	...	d_m
0	0.0047	...	0.0007
1	0.0055	...	0.0029
5	0.0050	...	0.0019
3	0.0015	...	0.0075

表 3：特定层 i 的存储激活示例。

3.4 激活提取与残差流分析

我们分析的核心是从残差流中提取激活向量 x_l。为清晰起见，我们给出一个简化的公式，以突出层的主要计算组件。在一个包含 L 层的仅解码器 Transformer 中，第 $l+1$ 层的状态递归更新如下：

$$x_{l+1} = x_l + \text{Attn}(x_l) + \text{MLP}(x_l) \quad (1)$$

其中 $x_l \in \mathbb{R}^{d_m}$ 表示第l层的隐藏状态。对于数据集中的每个样本，我们都会执行一次前向传播，并捕获最终词元位置的激活向量。我们关注最终词元位置，因为它汇总了生成开始前的全部上下文信息。对于仅解码器模型，最终词元的隐藏状态是唯一一个关注过整个提示的位置，因此代表了模型在生成前对输入的完整编码（Elhage 等人，2021）。我们将均值池化表示留作未来工作的比较基准。

最终得到的数据结构是一个张量，其中包含数据集中所有样本的逐层隐藏状态。表 3 展示了数据集中几个示例问题的激活示意。数据集中的每一行对应于网络中特定层提取的一个独特激活向量。列定义如下：

- 布鲁姆层级：根据布鲁姆分类法表示问题的认知复杂度。在我们的实验设置中，这些层级以整数值编码（例如，0 表示记忆，1 表示理解等），使我们能够分析模型的内部状态如何随不同抽象层级而变化。

- $d_0 \ldots d_m$：表示特定层。如果模型的隐藏维度为d_m（例如，Llama 模型为 4096），每个值d_i 表示特定层中第 i个维度的激活水平。

3.5 线性探测

为了确定布鲁姆层级是否以线性方式编码，我们为每一层采用一个逻辑回归（LR）探针。给定来自第l层的激活向量 x_l，探针会预测提示属于特定布鲁姆层级 $k \in \{0, \ldots, 5\}$的概率。探针在特定层上的成功表明，认知信号已被解耦，并可由网络的后续组件读取。

线性探针采用多分类逻辑回归模型实现。这一选择可确保分类成功反映的是表征中可通过线性方式获取的信息，而非非线性分类器的表达能力。模型使用默认的 ℓ_2 正则化并进行特征归一化。数据集采用分层抽样策略划分为训练集和测试集，比例为$80/20$，以保持各布鲁姆层级的类别分布。模型性能通过标准多分类指标进行评估，包括准确率、精确率和召回率。此设置可确保探针严格保持线性且容量受限，因此，预测成功可归因于表征中已存在的线性可解码信息，而非探针的表达能力。

4 结果

本节全面分析了所有受评估模型在各层的探测结果。我们不聚焦于单个代表性模型，而是强调不同模型之间的一致性，以评估认知深度表征的涌现是否是一种稳健的架构现象。

为了分析认知深度表征如何在网络中涌现，我们针对所有受评估的架构，在每一层的残差流上分别训练了独立的线性探针。图2总结了由此得到的逐层准确率变化轨迹。

尽管架构各不相同，仍呈现出若干清晰且一致的模式。首先，探针准确率在最初几层较低，并随着深度增加而上升，这表明认知深度并未在输入或嵌入

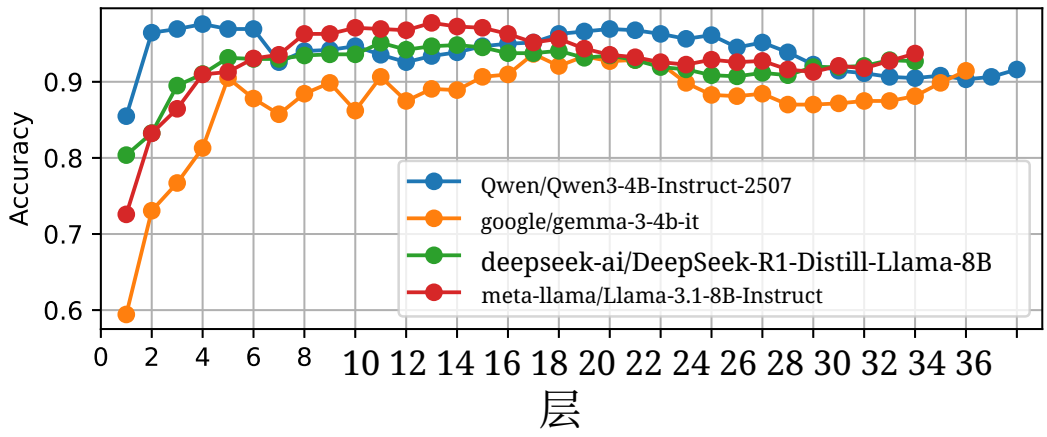


图 2：所有受评估模型各层的探针准确率。

层级上得到显式表征，而是通过计算逐步构建出来的。我们将这一现象称为前文提到的 CSO，指表征开始迅速且稳定地按认知层级实现线性可分的那一层。

其次，一旦达到较高的可分性，探针准确率便会在后续各层保持稳定。这种持续性表明，模型一旦确定了任务的认知性质，相应的表征就会保留在残差流中，而不是被反复重新计算。因此，更深层似乎表征的是已经确立的认知复杂度估计。该表征在生成过程中是否具有因果作用，而不只是相关性，仍有待通过干预研究加以验证。具体布鲁姆层级的准确率结果见附录A。

这些发现支持这样一种观点：认知深度是大语言模型内部表征中一种明确且结构稳定的特征，它在前向传播早期就已显现，并在后续层中保持稳定（Elhage 等人，2021）。较高的分类准确率表明，存在能够在表征空间中区分不同布鲁姆层级的仿射决策边界。形式上，某一层的线性可分性意味着，对应不同布鲁姆层级的表征可以由一组线性决策边界分开。具体而言，对于每个布鲁姆层级，表征空间中都存在一个超平面，使与该层级相关的激活向量位于超平面的正确一侧，并且其得分高于与任何其他层级相关的激活向量。换言之，每个提示的内部表征都更接近与其真实认知层级对应的区域，而非与其余层级对应的区域。因此，探测取得成功表明，认知复杂度编码在残差流中一个线性可访问的子空间内，而不是以高度纠缠或非线性的方式分布。

4.1 认知可分性起点

为了理解我们在此称为 CSO 的现象的内部机制，我们分析了第一个达到收敛的层。

令 A_l 表示第 l 层的探针准确率。我们将 CSO 层 $l^\star$ 定义为：

$$l^\star = \min\{l \mid A_l \geq \tau\} \tag{2}$$

其中，τ 是高准确率阈值。我们将 $\tau = 0.90$ 设定为一项保守标准：对于类别平衡问题，该标准显著高于多数类基线，同时对不同随机种子间的轻微波动保持稳健。

比较图2中的模型后，我们注意到，大多数模型的 CSO 都发生在第5层附近。

为了比较不同认知层级上的表现，我们使用雷达图（图$l^\star = 5$ 展示探针的分类准确率3）。每条轴代表六个布鲁姆层级中的一个，每条线对应特定模型上的探针。因此，雷达图全面展示了不同架构如何处理不同程度的认知复杂性。

为了进一步分析探针错误的结构，我们考察混淆矩阵。图4 展示了从 Llama-3.1-8B-Instruct 模型中获得的一个具有代表性的混淆矩阵。该混淆

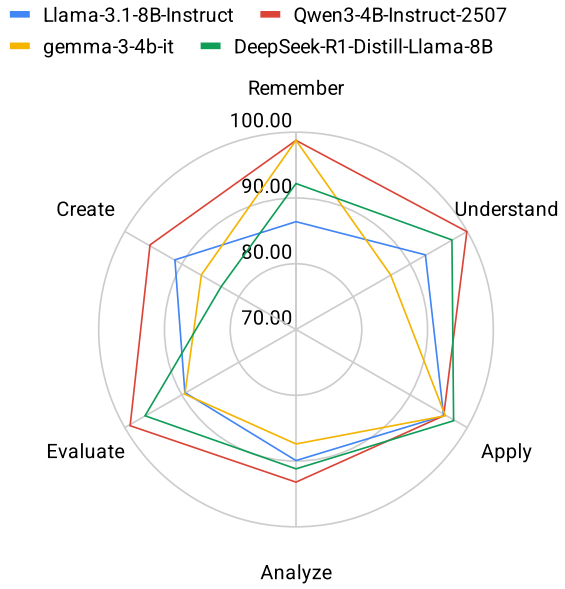


图3：四种模型在第5层各布鲁姆层级上的探针准确率。所有架构均表现出持续的高性能，细微差异反映了认知难度。

所有其他架构的混淆矩阵见附录B。

混淆矩阵揭示了高度结构化的错误模式。令 $\hat{y}_i$ 为预测的布鲁姆层级。我们分析误差距离 $|\hat{y}_i - y_i|$ 的分布。我们观察到：

$$\mathbb{E}\big[\,|\hat{y} - y|\,\big] \approx 1 \tag{3}$$

这表明误分类保留了序数邻近性，并发生在相邻布鲁姆层级之间。这种邻接效应表明，学习到的表征保留了认知复杂度的序数结构。

值得注意的是，这一模式与人工评估中已得到充分证实的歧义相一致。相邻层级之间的界限，例如记忆 和 理解，或分析 和评估，已知会因语境而异，且难以一致地进行操作化；相比之下，不同认知层级之间的区分则稳健得多（Krathwohl，2002年；Anderson 和 Krathwohl，2001）。

探针错误呈现出与这一结构相似的模式，这表明模型将认知深度编码为连续且具有语义意义的流形，而非一组任意的离散类别。这种邻接偏差错误的一致性

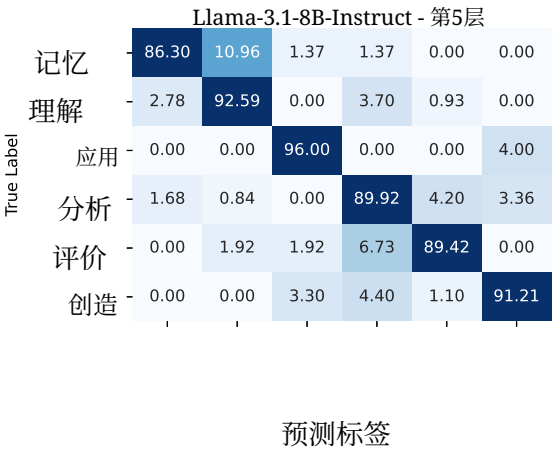


图4：Llama-3.1-8B-Instruct 在第5层的线性探针的代表性混淆矩阵。

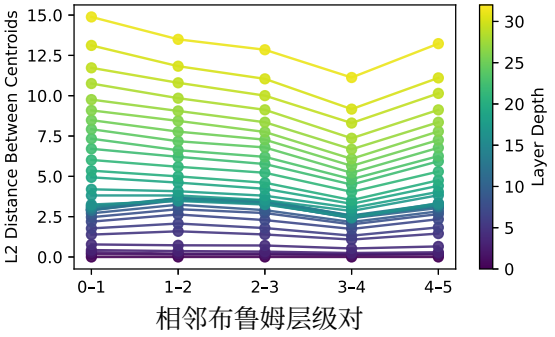


图5：Llama-3.1-8B-Instruct 模型各层之间相邻布鲁姆层级质心的欧几里得距离。早期层中的距离较小，并随着深度单调增加，表明认知层级逐步实现几何解缠。CSO 层 $l^\star = 5$ 标志着快速分离的开始。

在所有模型中均呈现相同的结构（附录 B），这进一步证明，观察到的行为反映了内部表征的真实属性，而非特定架构或探测设置造成的假象。

4.1.1 CSO 的几何证据。

为补充基于探针的分析，我们考察跨层的表征几何组织。对于每个层 l 和布鲁姆层级 k，我们计算类别质心

$$\mu_{l,k} = \mathbb{E}[x_l \mid y = k] \tag{4}$$

其中，x_l 表示第 l 层的残差流激活。然后，我们测量相邻质心之间的欧几里得距离：

$$D_l^{(k)} = \|\mu_{l,k+1} - \mu_{l,k}\|_2 \tag{5}$$



图6：控制实验的混淆矩阵。

图 5 显示，Llama-3.1-8B-Instruct 模型呈现出明显的、随深度变化的分离模式（其他模型的进一步分析见附录 C）。在较浅层中，质心距离较小，表明不同认知层级的表征仍高度纠缠。从大约第 $l^\star = 5$层开始，距离急剧增大，并在更深层持续增长，显示出布鲁姆层级逐步实现几何解纠缠。

重要的是，这项分析完全与探针无关。质心间距离大幅且系统性地增大，表明 CSO 对应于潜在空间中与探针无关的几何分离，而非线性分类器造成的假象。经过过渡层后，更深层并非只是保留可分性，还会进一步增强这种可分性，这表明认知复杂度在表示空间中被编码为一个不断扩展且有序的流形。

4.2 控制实验

为验证探针准确率反映的确实是认知表征，而非表层词汇模式，我们开展了控制实验。

长度分析。 我们考察了不同布鲁姆层级的问题长度是否存在差异，以及问题长度是否可能构成潜在混杂因素。统计检验表明，各层级之间的长度差异并不显著（$p = 0.053$），邦弗朗尼校正后，所有两两比较也均未达到显著水平。这些结果表明，不同布卢姆类别的问题长度大致均衡，因此不太可能以驱动探针分类性能。

浅层文本基线。 我们使用问题文本的 TF-IDF 向量（Qaiser 和 Ali，2018）和句子嵌入（Reimers 和 Gurevych，2019）训练逻辑回归分类器，不访问大语言模型内部状态。TF-IDF 的准确率为 73%，句子嵌入的准确率为 61%，均明显低于我们的探针准确率（$\sim$ 95%），这表明简单的词汇特征无法完全解释这一信号。

图 6 中相应的混淆矩阵显示，与残差流探针相比，其错误模式明显更加嘈杂且缺乏结构性。误分类分布更加分散，并且经常出现在非相邻的布鲁姆层级之间。

这种对比表明，仅凭浅层词汇特征，无法像模型内部大语言模型表征那样，以同等的保真度编码认知复杂度的序数结构。这些结果进一步支持了这样的解释：认知深度是在模型内部构建并以几何方式组织的，而不是可以还原为表层文本线索。

5 讨论

我们的结果表明，布鲁姆层级信息可从多个大语言模型的早期层中线性解码，这说明提示会根据认知要求迅速组织起来。不同架构中一致的认知可分性起始点进一步表明，这种组织并非特定于某个模型，而是反映了一种稳健的

表征模式。结合质心分析，这些发现支持这样一种观点：认知复杂度在残差流中被编码为一个逐步解缠且有序的流形。

与此同时，线性可分并不意味着模型会因果性地使用这些信息：模型可能是根据表层策略或任务类型对输入进行聚类，而非依据完全抽象的认知深度。尽管浅层基线的表现不佳，残差流中的可分性仍可能在一定程度上反映了提示词构造中由训练诱导形成的规律。尽管如此，这些结构化的序数误差模式表明，内部表征保留了有意义的层级关系，而非任意的标签边界。通过有针对性的干预来确立因果相关性，仍是重要的后续工作。

6 结论与未来工作

我们的研究表明，大语言模型具有可测量且线性可分的认知复杂度内部表征，并且该表征与布鲁姆分类法高度吻合。通过系统地提取并分析各种最先进模型残差流中的高维激活向量，我们证明，不同认知要求层级之间的区别并非仅仅是语言层面的表象，而是模型潜在空间的一种结构属性。

我们对这些表征逐层动态的研究揭示，在前向传播的早期至中间阶段会出现显著的认知可分性起始 。这一发现表明，大语言模型早在生成最终词元之前便已判定任务的认知性质；在早期层中，表征会根据认知层级逐渐分化。此外，线性探针取得的高分类准确率表明，这些内部表征以连贯且具有几何结构的方式编码认知层级，同时保留了人类教育理论的序数结构。

这些结果表明，布鲁姆分类法不仅适用于人类学习者，也是一套严谨且切合实际的框架，可用于审查人工认知的内部机制。通过超越表层性能指标，转而理解认知深度的机制，我们能够更好地区分死记硬背与真正的推理。

我们的研究结果揭示了若干值得探索的方向。首先，因果干预研究可以验证我们识别出的认知表征是否被模型实际利用。通过有选择地编辑或消融对布鲁姆层级预测力最强的维度，我们可以检验下游生成质量是否会以符合认知简化或混淆的方式下降。

此外，通过将我们的方法论扩展到非英语语言和专业领域（如医学诊断、法律推理），开展跨语言和跨领域分析，将有助于确立 CSO 的普适性。

随着该领域迈向日益自主且复杂的人工智能智能体，通过其神经相关项监测、解读并可能调控这些认知状态的能力将至关重要。我们的工作为构建更加透明、稳健且对齐的系统迈出了一步，确保人工智能能够依据与我们用于评估人类智能的同等精细的理解标准接受评估。

局限性

我们的分析仅关注仅解码器 Transformer 架构和英语教育类问题。其对其他架构（如仅编码器或编码器-解码器模型）及其他语言的普适性仍有待验证。此外，尽管线性探针表明认知复杂度可通过线性方式获取，但这并不能证明模型在生成过程中会以线性方式使用这些信息。最后，我们的方法论仅考察残差流，注意力模式和多层感知机内部机制则有待未来研究。

参考文献

Lorin W Anderson 和 David R Krathwohl。2001。学习、教学与评估分类法：布卢姆分类法修订版——教育目标分类法完整版本。Addison Wesley Longman, Inc.

Yonatan Belinkov。2022。探测分类器：前景、局限与进展。计算语言学，48(1):207–219。

Éloïse Benito-Rodriguez、Einar Urdshals、Jasmina Nasufi 和 Nicky Pochinkov。2025年。超越语言模型中的词元：通过文本体裁片段解读激活。*arXiv*预印本 *arXiv:2511.16540*。

Leonard Bereska 和 Efstratios Gavves。2024年。面向人工智能安全的机制可解释性综述。*arXiv*预印本 *arXiv:2404.14082*。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020年。语言模型是少样本学习者。神经信息处理系统进展，33：1877–1901。

Devichand Budagam、Ashutosh Kumar、Mahsa Khosh- noodi、Sankalp KJ、Vinija Jain 和 Aman Chadha。2024 年。分层提示分类体系：一种遵循人类认知原则的大语言模型通用评估框架。*arXiv* 预印本 *arXiv:2406.12644*。

Bartosz Cywi´nski、Emil Ryd、Senthooran Rajamanoha- ran 和 Neel Nanda。2025年。利用机制可解释性从大语言模型中引出潜在知识。*arXiv* 预印本 *arXiv:2505.14352*。

Nelson Elhage、Neel Nanda、Catherine Olsson、Tom Henighan、Nicholas Joseph、Ben Mann、Amanda Askell、YuntaoBai、Anna Chen、Tom Conerly 等另外 1 位作者。2021。《Transformer 电路的数学框架》。*Transformer Circuits Thread*，1(1):12。

Sabina Elkins、Ekaterina Kochmar、Jackie CK Cheung 和 Iulian Serban。2024年。《教师如何运用大语言模型和布鲁姆分类法创造教育测验》。载于人工智能 *AAAI* 会议论文集，第 38 卷，第 23084–23091 页。

Sandeep Reddy Gantla。2025年。《探索大语言模型中的机制可解释性：挑战、方法与洞见》。载于*2025*年数据科学、智能体与人工智能国际会议（*ICDSAAI*），第 1–8 页。IEEE。

Amir Hadifar、Semere Kiros Bitew、Johannes Deleu、Chris Develder 和 Thomas Demeester。2023年。EduQG：面向教育领域的多格式多项选择题数据集。*IEEE Access*，11：20885–20896。

Dan Hendrycks、Collin Burns、Steven Basart、Andy Zou、Mantas Mazeika、Dawn Song 和 Jacob Steinhardt。2020年。衡量大规模多任务语言理解能力。*arXiv* 预印本 *arXiv:2009.03300*。

Anne Herrmann-Werner、Teresa Festl-Wietek、Friederike Holderried、Lea Herschbach、Jan Griewatz、Ken Masters、Stephan Zipfel 和 Moritz Mahling。2024年。使用心身医学考试题评估 ChatGPT 对布鲁姆分类法的掌握程度：混合方法研究。《医学互联网研究杂志》，26：e52113。

John Hewitt 和 Percy Liang。2019。使用控制任务设计和解释探针。*arXiv* 预印本 *arXiv:1909.03368*。

Mohammad Hmoud 和 Shaqour Ali。2024年。人工智能教育中的布鲁姆分类法：提升人工智能时代教育效率与成效的拟议模型。《国际学习技术期刊》，31(2)：111。Thomas Huber 和 Christina Niklaus。2025年。大语言模型遇上布鲁姆分类法：对大语言模型评估的认知视角。载于第*31*届国际计算语言学会议论文集，第5211–5246页。Robert Jankowski、Filippo Radicchi、M Serrano、Marián Boguñá 和 Santo Fortunato。2025年。任务复杂度塑造神经网络中的内部表征与鲁棒性。*arXiv* 预印本 *arXiv:2508.05463*。

Mingyu Jin、Qinkai Yu、Jingyuan Huang、Qingcheng Zeng、Zhenting Wang、Wenyue Hua、Haiyan Zhao、Kai Mei、Yanda Meng、Kaize Ding 等。2025年。探索概念深度：大语言模型如何在不同层中获取知识与概念？载于第*31*届国际计算语言学会议论文集，第558–573页。

Been Kim、Martin Wattenberg、Justin Gilmer、Carrie Cai、James Wexler、Fernanda Viegas 等。2018。超越特征归因的可解释性：使用概念激活向量（TCAV）进行定量测试。载于国际机器学习会议，第2668–2677页。PMLR。

Junsol Kim、James Evans 和 Aaron Schein。2025年。政治视角的线性表征在大语言模型中涌现。*arXiv* 预印本 *arXiv:2503.02080*。

David R Krathwohl。2002年。布鲁姆分类法的修订：概述。理论与实践，41(4):212–218。

Ramya Kumar、Dhruv Gulwani 和 Sonit Singh。2025年。基于布鲁姆分类法对学习成果和考试题进行自动分析。*arXiv* 预印本 *arXiv:2511.10903*。

Yifan Luo、Zhennan Zhou 和 Bin Dong。2025a。In- versescope：用于解释大语言模型的可扩展激活反演。*arXiv* 预印本 *arXiv:2506.07406*。

Yiming Luo、Ting Liu、Patrick Cheong-Iao Pang、Dana McKay、Ziqi Chen、George Buchanan 和 Shanton Chang。2025b。面向大语言模型时代、旨在培养信息素养的布鲁姆教育分类法增强版。*arXiv* 预印本 *arXiv:2503.19434*。

Chris Olah、Nick Cammarata、Ludwig Schubert、Gabriel Goh、Michael Petrov 和 Shan Carter。2020年。《放大：电路导论》。*Distill*，5(3):e00024– 001。

Shahzad Qaiser 和 Ramsha Ali。2018。文本挖掘：使用 TF-IDF 考察词语与文档的相关性。国际计算机应用期刊，181(1):25–29。

Daking Rai、Yilun Zhou、Shi Feng、Abulhair Saparov 和 Ziyu Yao。2024年。基于 Transformer 的语言模型机制可解释性实用综述。*arXiv* 预印本 *arXiv:2407.02646*。

Bianca Raimondi、Daniela Dalbagno 和 Maurizio Gabbrielli。2025年。通过机制可解释性分析微调后的大语言模型中的道德偏见。*arXiv* 预印本 *arXiv:2510.12229*。

Nils Reimers 和 Iryna Gurevych。2019年。Sentence-BERT：使用孪生 BERT 网络的句子嵌入。收录于*2019*年计算语言学实证方法会议论文集。计算语言学协会。

Dong Shu、Xuansheng Wu、Haiyan Zhao、Daking Rai、Ziyu Yao、Ninghao Liu 和 Mengnan Du。2025年。稀疏自编码器综述：解读大语言模型的内部机制。*arXiv* 预印本 *arXiv:2503.05613*。

Katharina Simbeck 和 Mariam Mahran。2025年。基于稀疏自编码器的机制可解释性：探测大语言模型中的宗教、暴力和地理。*arXiv* 预印本 *arXiv:2509.17665*。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。注意力就是你所需的一切。神经信息处理系统进展，30。

Yongan Yu、Mengqian Wu、Yiran Lin 和 Nikki G Lobczowski。2025年。思考：大语言模型能否出声思维？*arXiv* 预印本 *arXiv:2505.20184*。

Khandoker Ashik Uz Zaman、Ashraful Islam、Yusuf Mahbubul Islam 和 Md Abu Sayed。2024年。学生计算机科学课程查询数据集：依据布鲁姆分类法进行分类和评分。《数据简报》，53:110109。

Haiyan Zhao、Fan Yang、Bo Shen、Himabindu Lakkaraju 和 Mengnan Du。2024年。《探究大语言模型如何运作：一种可解释性视角》。*arXiv* 预印本 *arXiv:2402.10688*。

Maria-Eleni Zoumpoulidi、Georgios Paraskevopou- los 和 Alexandros Potamianos。2025年。Bloom-wise：使用受布鲁姆分类法启发的提示，增强大语言模型的问题解决能力。载于《第三届数学自然语言处理研讨会论文集（*MathNLP 2025*），第34–49 页。

附录

A 各布鲁姆层级的准确率

图7分别报告了各模型在不同层的每个布鲁姆层级上的探针准确率。不同架构呈现出一致的模式：到 CSO 层时，所有层级的表现都趋近于上限 $l^\star = 5$。

B 混淆矩阵

图8中的混淆矩阵证实，分类错误具有高度结构性，而非随机发生。在所有架构中，误分类主要出现在相邻的布鲁姆层级之间（例如，分析与评价），而远距离混淆（例如，记忆与创造）则不存在。

这种相邻偏差进一步证明，认知复杂度在表征空间中被编码为有序流形。因此，布鲁姆分类法的序数结构在几何上得以保留：相邻层级的表征彼此更接近，而相距较远的层级则被更大的间隔分开。这一模式在不同模型中保持一致，表明它反映的是基于 Transformer 的表征的一般属性，而非特定架构所造成的假象。

C 质心

图 9 展示了所有受评模型中，各层相邻布鲁姆层级质心之间的欧几里得距离。在较早的层中，质心间距离较小，表明不同认知层级的表征高度纠缠。

从 CSO 层开始，距离急剧增大，随后随着深度增加而持续单调扩展。这种渐进式分离表明，较深层不仅保留了线性可分性，还会主动放大不同认知类别之间的几何差异。

重要的是，这项分析与探针无关，因此证实认知可分性起始点对应的是潜在空间的内在重组，而非线性分类造成的假象。

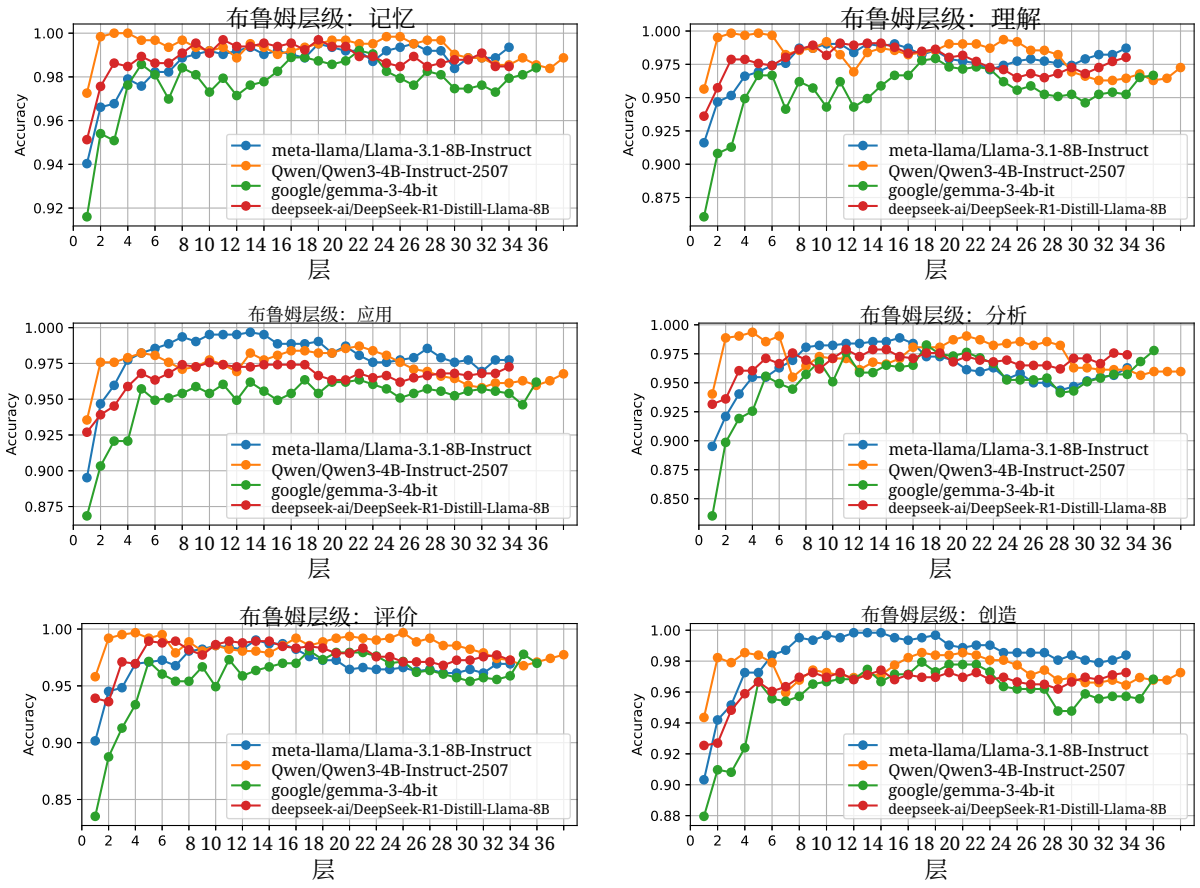
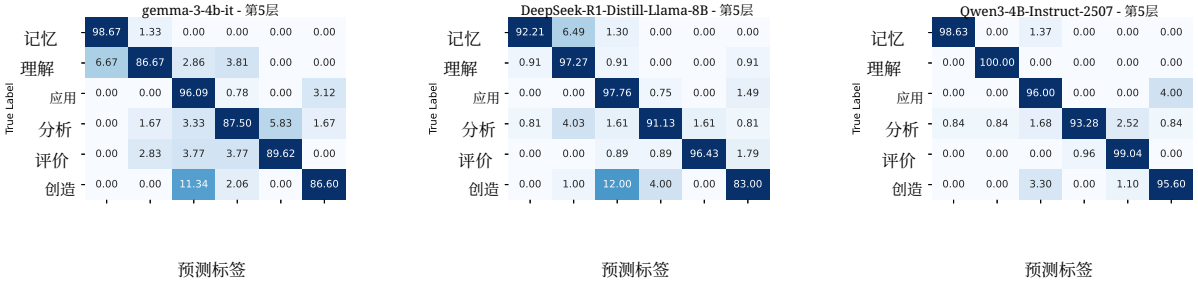
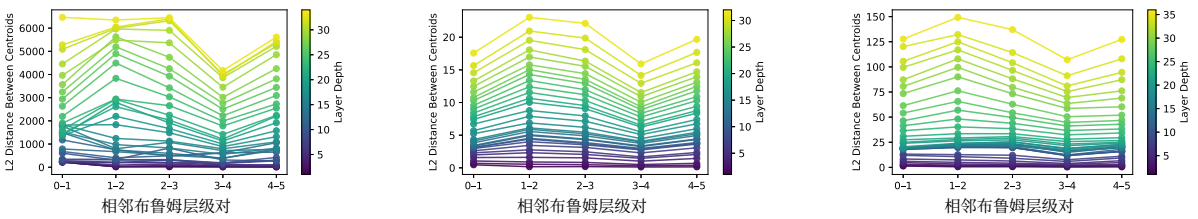


图7：各布鲁姆层级的逐层线性探测准确率。



gemma-3-4b-it DeepSeek-R1-Distill-Llama-8B Qwen3-4B-Instruct-2507

Fig图 8：四个代表性模型在 CSO 层的线性探针混淆矩阵 $l^\star = 5$.



gemma-3-4b-it DeepSeek-R1-Distill-Llama-8B Qwen3-4B-Instruct-2507

图 e 9：所有已评价模式中相邻布鲁姆层级质心之间的逐层欧几里得距离 ls.