真理即轨迹：内部表示揭示了大语言模型推理的什么

Hamed Damirchi[1] **Ignacio Meza De la Jara**[1] **Ehsan Abbasnejad**[2]
Afshar Shamsi[3] **Zhen Zhang**[1] **Javen Shi**[1]
[1]澳大利亚机器学习研究所，阿德莱德大学 [2] 莫纳什大学
[3]康考迪亚大学
[1]{firstname.lastname}@adelaide.edu.au [2]
{firstname.lastname}@monash.edu[3]afshar.shamsi@concordia.ca

arXiv:2603.01326v1 [cs.CL] 1 Mar 2026

摘要

现有的大语言模型（LLM）可解释性方法通常将隐藏状态视为激活空间中的静态点，并假设可以利用单个层的表征区分正确与错误的推理。然而，这些激活充斥着多义特征，导致线性探针学到的是表层词汇模式，而非底层推理结构。我们提出 Truth as a Trajectory（TaT），将 Transformer 推理建模为迭代细化展开而成的轨迹，把分析重点从静态激活转向逐层几何位移。通过分析表征在各层之间的位移，TaT 揭示了能够区分有效推理与虚假行为的几何不变量。我们在涵盖常识推理、问答和毒性检测的基准上，对稠密架构和混合专家（MoE）架构进行了评估。我们不访问激活本身，仅利用激活在各层间的变化，结果表明 TaT 能有效减轻对静态词汇混杂因素的依赖，优于传统探测方法，并确立了轨迹分析这一大语言模型可解释性的补充视角。

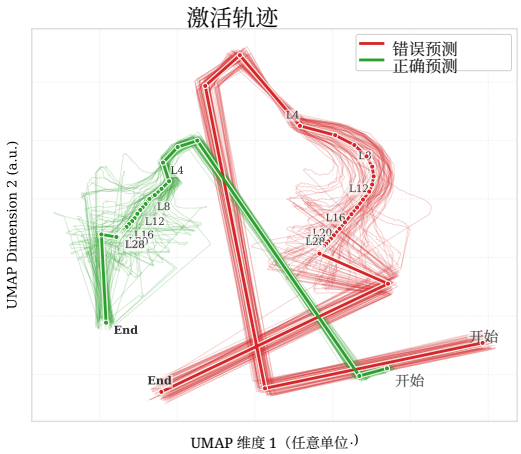


图 1：**轨迹揭示了静态嵌入之外的结构**。我们将逐层隐藏状态绘制为激活空间中的轨迹。正确生成（绿色）遵循较为平滑的路径，而错误生成（红色）则呈现明显偏离。尽管监督投影放大了两者之间的差异，但其几何结构表明，与孤立状态相比，对完整轨迹建模有助于区分有效推理与伪推理。

1 引言

大语言模型在安全关键领域的部署——从法律推理到内容审核——已使对最终输出的黑箱评估变得不足（Orgad etal., 2025年; Shailya etal., 2025年;Aljohaniet al., 2025年）。尽管行为基准测试衡量模型生成的内容，但它们对模型如何得出结论仅能提供有限洞见。若无法洞察内部过程，从业者便无法可靠地区分正确遵循推理模式的模型与仅依赖表层启发式方法的模型。这一差距为安全造成了根本性瓶颈。要信任这些系统，我们必须能够验证其内部思维过程的有效性，而不仅仅是其词元的概率。

近期机制可解释性文献与线性表征假说（Park 等人，2023年）相一致，该假说表明，推理有效性或有毒性等高层次属性被编码为模型激活空间中不同的线性方向。因此，可解释性与行为编辑方法假定，某些属性可通过线性探针（Bao 等人，2025年）进行分类，或通过激活引导（Aljohani 等人，2025年; Rimsky 等人，2024年；匿名，2025年；O’Brien 等人，2025年）在精心选择的层上加以操控。在这一观点下，可解释性归结为识别出能够区分有效行为与无效行为的“正确”静态层和方向。这些方法的一个局限在于它们依赖

于针对目标属性的对比样本。因此，探测与引导结果的效力取决于所采用的具体数据集，因为它们作用于静态激活。

因此，尽管早期较为乐观，近期证据表明，“真理几何”往往具有任务特异性，并且在不同领域之间相互正交（Azizian 等人，2025年），也就是说，在某一语境中训练用于检测推理正确性的探针，无法泛化到其训练分布之外。这种失败可能源于变换器激活的多义性（Lindsey 等人，2025年），这些激活同时编码词汇内容、句法结构以及任务特定的伪迹。这样一来，线性探针可能会抓住表层相关性，例如特定词元的出现，而非推理模式的底层有效性。此外，选择探测哪一层仍然缺乏原则依据。激活引导文献中的发现表明，有效干预仅可能发生在大语言模型中部的狭窄层段，且在不同数据集上的结果并不一致（Rimsky 等人，2024年），而其他研究则认为推理流形本质上是非线性的（Manson，2025年），难以进行简单的线性分离。这些不一致往往是在合成或严格受控的数据集上进行评估的副产品，因此一个可泛化、与任务无关的推理有效性特征是否存在于真实世界基准之中，仍是一个悬而未决的问题。

我们提出**真理即轨迹（TaT）**，将大语言模型推理重新构想为一个动力学过程，而非一系列静态的层快照。TaT 将跨越层和词元的推理过程展开为穿过表征空间的一条轨迹（见图 1）。我们将分析重点从原始激活转向逐层位移，即残差流在相邻层之间激活的差异。这一转换降低了对静态词元身份和词汇特征的依赖，聚焦于表征如何随深度更新，而不是单层所编码的内容。为捕捉这一演化过程的动态特征，我们采用轻量级 LSTM 分类器。尽管我们探索了受近期 Transformer 动力学研究启发的简单运动学度量，例如速度、加速度和曲率（Zhou 等人，2025年；Fernando 和 Guitchounts，2025年），但发现这些度量无法在多种任务中稳定预测推理有效性。相比之下，我们的学习方法能够捕捉与有效推理相关的非线性结构不变量。据我们所知，这是首个通过将所有层中每个词元的激活展开为连续轨迹，来对大语言模型的“内部思维过程”如何判断输入陈述的推理有效性进行建模的方法。

我们在广泛使用的基准上评估 TaT，涵盖常识推理、问答、事实性和毒性检测，并涉及稠密架构和混合专家（MoE）架构。我们的主要发现是，TaT 具有出色的跨数据集泛化能力。在单一源数据集上训练的轨迹分类器，可以泛化到提示结构各异的不同任务，在不同领域的表现优于线性探测基线，也优于基础模型自身的零样本和上下文学习（少样本）表现。这表明，有效推理的轨迹编码了结构不变性，能够超越特定任务的词汇模式。在毒性检测方面，TaT 对词汇混杂因素更加稳健，例如，它能更好地区分被引用或置于上下文中的有害词汇与有害意图，并且表现始终优于基于中间层或末层激活训练的探测器。我们的贡献总结如下：

- **基于轨迹的可解释性**：我们提出真理即轨迹（TaT），将大语言模型推理建模为一种动力学过程，该过程在层与词元之间展开，捕捉推理的连续几何演化，而非聚焦于单独的层。
- **跨任务几何不变量**：通过分析逐层位移向量，而非激活本身，我们降低了对静态词汇特征的依赖，并着重揭示模型内部的几何细化过程，从而展现出线性探针无法观测到的轨迹层面结构。
- **基于轨迹的行为检测**：我们证明，轨迹分析可以扩展到复杂的行为属性，例如有毒性。在区分有意义的语境与有害意图方面，TaT 的表现显著优于线性探针，验证了其在可靠模型监控中的实用性。

2 相关工作

机制可解释性传统上聚焦于孤立层中概念的静态线性表示。然而，残差流

Transformer 的结构启发了一种动力系统视角：表征在各层间的演变可能揭示模型的思维过程。我们的研究立足于这两种范式的交汇处，聚焦于推理轨迹的几何。

2.1 静态线性表征

当前机制可解释性领域的主流范式是线性表征假说（LRH），该假说认为，神经网络将高层概念表征为激活空间中的线性方向（Park等人，2023年；Elhage等人，2022年；Liu等人，2025年）。这一观点促使研究者通过线性探针对行为进行分类（Belinkov，2022年），或利用稀疏自编码器（SAE）对其进行解释（Huben等人，2024年）。然而，这些方法存在显著局限。由于事先无法确定概念所在的具体深度，两者都需要进行穷尽式的逐层搜索；而且，SAE在不同模型之间缺乏一致的结构映射，这使得跨模型解释更加困难。此外，线性探测依赖主观构建的精选数据集来定义目标行为，因此发现的方向取决于特定的语义内容，而非模型内在的几何结构。尽管对比一致性搜索（CCS）等扩展方法（Burns等人，2022年）试图在无监督条件下寻找潜在知识，但它们与标准探测通常都将表征视为孤立层中的静态向量（例如“第 L层”）。这种做法忽视了推理过程的时间演化。我们的工作对这种静态且依赖内容的观点提出质疑，认为推理有效性是一种动态属性，最好通过计算轨迹中激活的几何位移来刻画，而非通过激活在固定坐标中的静态位置来刻画。

2.2 表征工程与转向

表征工程（RepE）将重点从分析转向控制。Zou 等人(2023年) 表明，提取“概念向量”可以自上而下地引导模型行为：将这些向量注入残差流，便能抑制幻觉或有害输出。然而，静态引导并不能保证对每个模型和任务都能可预测地引导其朝着预期行为发展（Rimsky 等人，2024年）。这可能是因为有些数据集要求模型的内部状态沿非线性路径变化。因此，近期相邻领域的文献提倡采用其他方法。

Zhang 和 Dong (2025年);Manson (2025年) 认为，简单的线性干预无法体现多步推理过程中的“流形演化”。同样，Postmus 和 Abreu (2024年) 指出，引导应通过变换作用于激活区域，而非采用静态方向。与此同时，与线性探测类似，这些方法仍依赖主观精选的数据集来定义目标行为。我们的方法通过分析整个轨迹上的激活位移来缓解这些问题，从而捕捉激活轨迹在任务特定区域中的变化方式，而无需依赖逐层激活。

2.3 Transformer作为动力系统

将Transformer解释为离散化动力系统的理论，为我们的运动学框架提供了基础。残差更新 $\mathbf{h}_{\ell+1} = \mathbf{h}_\ell + f(\mathbf{h}_\ell)$ 在数学上等价于用欧拉方法求解常微分方程（ODE）的一步（Chen 等人，2018年；Lu 等人，2019年）。这一等价性意味着，激活的逐层演化可以被分析为高维状态空间中的一条连续轨迹，而不是离散的、相互独立的状态。Geshkovski 等人（2023年）将这一观点扩展到自注意力，将词元建模为相互作用的粒子，它们随“时间”（深度）收敛到语义簇。我们针对多种模型和数据集，对激活轨迹的运动学度量开展了广泛研究。我们表明，尽管预先确定的度量缺乏普遍适用性，内部表征的内在运动仍可能包含有用信号。因此，我们通过激活空间对这一运动进行建模，表明该轨迹能够准确分类输入陈述推理过程的有效性或行为特征。

2.4 推断几何

与我们的提案最相关的是关于“推理几何”的新兴研究。Zhou 等人（2025年）最近提出，逻辑有效性决定表征流的“速度场”，而语义内容决定位置。他们证明，逻辑推演会沿着特定的流动模式展开，这些模式有别于语义关联。不过，他们的分析仍主要停留在理论层面，关注的是理想化的逻辑结构，而非真实世界大语言模型使用中常见的含噪、非结构化数据。与此同时，Manson （2025年）提出了“曲线推理”的概念，并展示了

“语义关切”（例如紧迫性或道德框架）会在残差流中引发可测量的曲率。这一发现凸显了静态线性探针的不足，因为它们往往会将这些几何细微差异与语义内容混为一谈。我们证明，这些运动学特征不仅存在，而且在一般情境下可以学习。不同于以往局限于精选数据集的研究，我们利用这些几何轨迹，在复杂的真实世界基准中区分正确行为与错误行为。

3 问题设定

我们考虑一种标准的评估设定：向模型提供上下文或提示 $\mathbf{x}$，以及一组候选续写 $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$。这些续写的范围可以从单个词元（例如“True”/“False”）到完整句子或推理链。对于每个候选续写 c_i，我们将提示与续写拼接，构造完整的输入序列。随后，我们对 Transformer 模型执行前向传播，以提取其内部表示。

设 L 表示模型的层数，N_i 表示候选 c_i 对应的拼接序列中的词元数。随着输入被处理，每个 Transformer 块都会更新残差流，生成一系列激活向量。我们收集每个 Transformer 块输出端所有词元对应的激活，构成轨迹矩阵 $\mathcal{T}_i \in \mathbb{R}^{M_i \times d}$，其中 d 为隐藏维度大小，$M_i = N_i \times L$ 表示展开计算图中的向量总数。

我们的主要目标是分析这些轨迹的几何属性。具体而言，我们旨在确定，与错误候选项的轨迹相比，对应正确续写的轨迹 $\mathcal{T}_i$ 是否呈现出独特的运动学特征。

4 推理的几何

本节旨在对使用第 3 节所述流程提取的激活轨迹进行运动学分析。这项研究的动机在于，Transformer 动力学的理论研究与实际的激活探测之间存在显著差距。尽管近期关于激活空间动力学的研究揭示了一些引人关注的特性，例如理想化情境中的典型速度剖面，但这些发现很少在标准且多样化的基准上接受检验。相反，探测方法往往忽略了推理的时间演化。这里，我们探究简单且易于解释的运动学度量能否在复杂情境中区分正确与错误的推理模式。我们采用预言机引导分析，此时真实标签事先已知。通过区分正确与错误的输出，我们旨在找出一致的运动学规则，例如速度、曲率、加速度等统计特征的差异，这些特征始终与正确的推理模式相关。

为此，我们探究简单且易于解释的运动学度量能否在复杂情境中区分正确与错误的推理模式。我们采用预言机引导分析，此时真实标签事先已知。通过区分正确与错误的输出，我们旨在找出一致的运动学规则，例如速度、曲率、加速度等统计特征的差异，这些特征始终与正确的推理模式相关。

4.1 运动学描述符

对于固定输入，具有 L层的 Transformer 会生成一系列隐藏状态 $\mathcal{T}$ =(h_0, h_1, .., h_L)，其中每一层都会按如下方式应用残差更新：

$$h_{\ell+1} = h_\ell + f_\ell(h_\ell). \tag{1}$$

该更新可解释为隐藏状态随深度变化的离散时间演化。由此，可以从 $\mathcal{T}$推导出逐层运动学描述符，以捕捉幅度、变化量和方向一致性。我们将第 ℓ 层的位移向量定义为 Δh_ℓ : $= h_{\ell+1} - h_\ell$。该向量表示第 ℓ个 Transformer 块对残差流所做的更新。

速度 速度定义为 $v_\ell := \|\Delta h_\ell\|_2$，用于衡量相邻层之间激活向量变化的幅度。请注意，我们的目标并非解读速度较大或较小意味着什么，而是比较正确生成与错误生成之间的速度剖面，并以无偏的方式识别不同模型和数据集中的一致模式。

加速度 加速度定义为 a_ℓ :=$v_\ell - v_{\ell-1}$，即速度梯度。

加加速度 加加速度定义为加速度的变化率，j_ℓ := $a_\ell - a_{\ell-1}$ 并刻画轨迹演化的平滑性。

方向曲率 曲率衡量相继更新之间的方向一致性：

$$\kappa_\ell := \frac{\langle \Delta h_\ell, \Delta h_{\ell-1} \rangle}{\|\Delta h_\ell\|_2 \, \|\Delta h_{\ell-1}\|_2}. \tag{2}$$

该公式刻画了连续位移之间的角偏差，曲率越高，方向变化就越突然。

运动学曲率方向曲率刻画了角偏差，此外，我们还定义了一种基于轨迹瞬时运动学特性的几何曲率度量。设$\mathbf{v}_\ell$ 和$\mathbf{a}_\ell$分别表示矢量速度和加速度。运动学曲率定义如下：

$$\kappa_\ell^{\text{kin}} := \frac{\|\mathbf{a}_\ell\|_2}{\|\mathbf{v}_\ell\|_2^2}. \tag{3}$$

当轨迹相对于其步长表现出较大的加速度时（即当 $\|\mathbf{v}_\ell\|$ 较小时），该量会变大；即使激活空间中的整体移动幅度较小，它也能捕捉到更新方向的突变。

弧长 弧长量化了激活向量在网络深度方向演变时所经过的总距离：

$$S := \sum_{\ell=0}^{L-1} \|h_{\ell+1} - h_\ell\|_2. \tag{4}$$

该指标可作为模型所付出总体几何努力的近似度量。请注意，弧长是一个全局描述符，为整条轨迹提供一个标量值。

运动学描述符有用吗? 我们使用推理和事实性基准，在 LLAMA 3.1–8B（图 2）和 Qwen 2.5-14b（附录 3 中的图 A）上评估这些描述符。运动学信号（速度、加速度、加加速度）可将生成结果分类为正确或错误。速度在各数据集上的表现均优于其他描述符，这与 Zhou 等人（2025年）提出的假设一致：有效性体现在具有特征性的速度曲线中，且这种曲线不同于语义位置。然而，在零样本设置下使用归一化对数似然进行评估时，没有任何单一描述符能始终达到与模型自身相当的准确率。将分析扩展到 Qwen2.5-14B 后，结果的一致性更低。虽然这些描述符总体上优于随机分类器（ARC-Easy 除外），但没有一个能接近基础模型的准确率。为排除轨迹构建方式可能带来的问题，我们还尝试仅使用续写词元构建轨迹、单独提取最后一个词元的激活，以及对激活求平均以得到长度为 的轨迹。在所有变体中，表现都只会下降。需要指出的是，运动学描述符输出的是标量幅度。因此，激活空间中的方向信息对这些描述符而言仍不可观测。

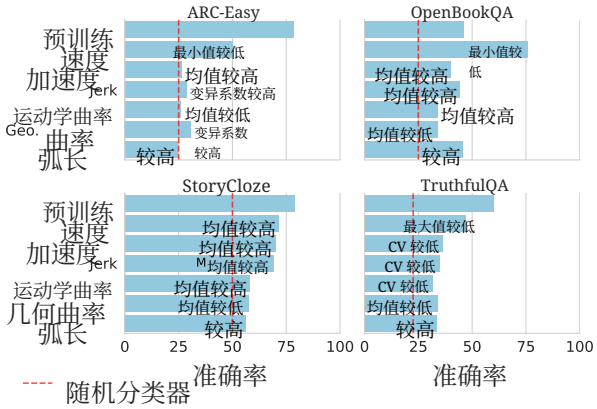


图2：使用运动学描述符在4个推理基准上的表现。红色虚线表示随机分类器。虽然激活速度的表现优于基础模型，但这种性能提升在不同数据集之间并不一致。

要点。激活速度中的预测信号表明，轨迹激活的速率和方向编码了有关推理有效性的信息。然而，静态规则无法始终有效地泛化。因此，尽管运动学描述符可能捕捉到推理动态的某些方面，但它们仍不足以胜任。这促使我们需要借助学习模型，更好地解读这些几何信号。

4.2 真理作为轨迹（TaT）

基于第4节中的运动学分析，我们提出**真理即轨迹（TaT）**，一种用于检测推理有效性的可学习框架。虽然标量运动学描述符（例如速度、曲率）能够捕捉一定的几何直觉，但它们舍弃了高维激活空间中蕴含的丰富方向信息。TaT通过直接在轨迹动态流形上学习决策边界来解决这一问题。

4.2.1 轨迹构建

回顾第3 节，对于候选续写 c_i，我们提取一条激活轨迹。为了将推理的几何演化与静态词汇线索以及原始激活中的其他持续内容分离开来，我们将这些向量转换为逐层位移序列。我们将位移向量定义为：

$$\mathbf{d}_{t,\ell} = h_{t,\ell+1} - h_{t,\ell} \tag{5}$$

我们从特权基假设 (Elhage等人的视角出发，

2023年; Bricken 等人, 2023年)。原始激活 $h_{t,\ell}$ 往往由幅度较大、相对持久的成分（包括特定于词元和提示的内容）主导，因此容易受到多义性和叠加的干扰。通过计算不同层之间的差值，我们可以减弱这些混合成分，有效减去这一静态背景，从而分离出活跃的残差更新 $f_\theta(h_{t,\ell})$。由于 Transformer 中逐元素的非线性，这些更新受到一致参考系（残差对齐）的约束。因此，不同于那些无法明确表明某个特征是否存在的原始状态，位移 $\mathbf{d}_{t,\ell}$ 能够准确表明模型是否正在主动写入该特征。这使我们能够将推理过程的机制与模型记忆状态的语义区分开来。

我们将所有词元和层的逐层更新堆叠起来，形成连续的轨迹序列$\mathbf{S}_i$:

$$\mathbf{S}_i = [\mathbf{d}_{1,0}, \ldots, \mathbf{d}_{1,L-1}, \mathbf{d}_{2,0}, \ldots, \mathbf{d}_{N_i,L-1}] \quad (6)$$

其中，$\mathbf{S}_i \in \mathbb{R}^{M_i \times d}$ 和 $M_i = N_i \times L$ 分别表示轨迹总长度。该表述将推理过程展开为一个单一的时间序列，将层和词元的推进视为一条统一的路径。

4.2.2 使用长短期记忆网络建模动态

为捕捉$\mathbf{S}_i$中有效推理的非线性结构不变量，我们使用长短期记忆网络（LSTM）。相比基于 Transformer 的探针，我们选择 LSTM，以显式建模几何更新的序列依赖关系，同时保持较低的计算开销。LSTM 逐步处理轨迹序列 $\mathbf{S}_i$。处理完长度为 M_i 的整个序列后，最终隐藏状态$\mathbf{z}_{M_i}$ 将几何轨迹编码其中。随后，将该向量输入线性分类头$\mathbf{W} \in \mathbb{R}^{d_{lstm} \times 1}$，以预测有效性的概率:

$$\hat{y}_i = \sigma(\mathbf{W}^T \mathbf{z}_{M_i} + b) \quad (7)$$

通过在位移轨迹$\mathbf{S}_i$上进行训练，模型能够学习与推理正确性相关的几何特征，并将其泛化到特定词汇和提示内容之外。

5 实验

我们对真理即轨迹（TaT）框架进行了全面评估，以考察其区分有效推理与虚假相关性的能力。我们的实验涵盖了多个不同领域，包括常识推理、阅读理解、事实性和有毒性检测，并涉及稠密架构（Llama-3.1-8B、Qwen2.5-14B/32B）和混合专家架构（Qwen2.5-30B MoE）。我们将基于轨迹的分类器与标准静态线性探测基线，以及冻结语言模型本身的零样本和少样本能力（我们将其称为基础模型）进行比较。实现细节请参阅附录C。

5.1 推理轨迹是否具有泛化性?

我们研究有效推理的几何特征在不同任务之间是否一致。如果 TaT 捕捉的是“真理”（或有效性）的基本结构不变量，而非特定于任务的混杂因素，那么在一个数据集上训练的轨迹分类器应能泛化到未见过的数据集，而无需针对未见任务进行微调。

设置 我们在一系列推理基准上进行评估: ARC-Easy (ARC-E)、ARC-Challenge (ARC-C) (Clark 等人，2018年)、BoolQ (Clark 等人，2019年)、HellaSwag (Zellers 等人，2019年)、OpenBookQA (OpenQA) (Mihaylov 等人，2018年)、StoryCloze (Mostafazadeh 等人，2016年)、CommonsenseQA (ComQA) (Talmor 等人，2019年)、CosmosQA (CosQA) (Huang 等人，2019年)和 SocialIQA (SiQA) (Sap 等人，2019年)。对于每个数据集，我们在训练集划分上训练一个 TaT 分类器（使用逐层位移）和一个线性探测器（中间层探针），并在所有其他数据集的评估集划分上对其进行评估。由于探针性能可能高度依赖于层，我们选用中间层作为标准设置，并在附录D中报告从中间层到较晚层的扫描结果。

结果 表 1和表8 (附录B) 分别总结了 Llama-3.1-8B 和 Qwen2.5-14B 的结果。与线性探针相比，TaT 展现出显著的分布外（OOD）泛化能力。我们观察到，在所有训练集上，TaT 的平均表现都优于线性探测，这表明基于轨迹的方法捕捉到了可泛化的信号，能够稳健地迁移至不同领域。这一优势在分布外设置下尤为明显。虽然 OpenBookQA 对我们的方法表现出较低的泛化适应性，这可能是因为它的提示

表 1：各评估数据集上的基准准确率，包括分布内和分布外指标。基础模型准确率括号内的数值表示：在这些数据集上，使用归一化对数似然进行评估会导致性能下降。虽然这对我们的方法并不公平（因为基础模型能够看到所有可能的答案，而我们的方法不能），但在大多数情况下，我们的方法仍优于基础模型本身。

训练数据集	方法	评估数据集									Avg	分布内准确率	分布外平均值
		ARC-C	ARC-E	开放域问答	BoolQ	HellaSwag	CosQA	SiQA	ComQA	StoryCloze			
零样本准确率		50.1	78.5	62.4	74.7	57.4	81.7 (26.1)	65.2 (48.0)	62.4	78.8	67.9 (59.8)	-	-
少样本准确率		65.3	84.3	67.0	83.8	76.8	82.0 (44.5)	67.7 (60.8)	71.1	83.1	75.7 (70.7)	-	-
ARC-C	线性探测	75.32	80.09	**78.60**	55.48	70.35	73.34	65.47	**74.59**	66.01	71.03	75.32	70.49
	TaT（我们的方法）	**82.17**	**85.31**	73.60	**96.91**	**73.89**	**74.24**	**75.49**	72.48	**82.58**	**79.63**	**82.17**	**79.31**
ARC-E	线性探测	**75.55**	83.99	**80.95**	58.82	71.07	66.83	**66.60**	**78.55**	69.62	72.44	83.99	71.00
	TaT（我们的方法）	73.81	**89.10**	77.20	**78.56**	**79.19**	**75.08**	60.85	71.09	**94.98**	**77.76**	**89.10**	**76.34**
开放域问答	线性探测	66.42	69.44	83.15	56.79	**70.26**	61.53	65.47	73.01	57.94	67.11	83.15	65.11
	TaT（我们的方法）	**78.41**	**87.12**	**90.80**	**89.85**	56.50	**76.42**	**69.70**	**75.02**	**81.64**	**78.38**	**90.80**	**76.83**
BoolQ	线性探测	44.51	46.61	46.30	83.65	**44.16**	**61.51**	49.20	**53.22**	59.45	54.29	83.65	50.62
	TaT（我们的方法）	**53.50**	**62.75**	**54.20**	**85.05**	33.08	51.56	**50.20**	47.50	**71.41**	**56.58**	**85.05**	**53.02**
HellaSwag	线性探测	52.92	58.60	58.35	58.24	88.64	**71.17**	**60.39**	**63.34**	77.61	65.47	88.64	62.58
	TaT（我们的方法）	**65.96**	**74.92**	**65.80**	**64.22**	**92.46**	66.40	55.27	62.82	**95.75**	**71.51**	**92.46**	**68.89**
ComQA	线性探测	**71.72**	76.27	**77.05**	50.31	**73.49**	**74.18**	**66.38**	**81.98**	60.72	70.23	**81.98**	68.77
	TaT（我们的方法）	60.92	**80.05**	73.80	**68.72**	59.51	64.89	60.18	77.56	**88.51**	**70.46**	77.56	**69.57**
CosQA	线性探测	**70.09**	75.27	**78.10**	66.50	60.73	**81.64**	**67.69**	**77.34**	57.36	70.52	**81.64**	69.13
	TaT（我们的方法）	69.71	**83.59**	74.60	**77.77**	**70.44**	80.94	61.21	64.05	**90.19**	**74.72**	80.94	**73.94**
SiQA	线性探测	49.82	54.53	**74.75**	55.77	**46.98**	**61.54**	**69.33**	**72.16**	57.91	60.31	**69.33**	59.18
	TaT（我们的方法）	**59.90**	**71.72**	70.60	**63.85**	46.76	59.73	66.17	55.69	**90.89**	**65.03**	66.17	**64.89**

结构与其他多项选择格式显著不同，但总体趋势仍然令人信服。此外，性能差距并不对称。在 TaT 表现不如线性探测的少数情况下，差距微乎其微。然而，在 TaT 表现出色的多数情况下，其优势都相当显著，这表明尽管探测方法有其独特的失效模式，TaT 仍然稳健。此外，TaT 的表现始终优于基础模型自身的零样本和少样本（上下文学习）基线。尽管我们的方法在推理时采用零样本方式，输入提示中不使用任何示例，但其表现仍有效超越了模型凭借少样本示范进行推理的内在能力，这凸显了有效性的几何结构相较于依赖模型表层输出的优势。

共享与任务特定几何 这里我们考察跨任务迁移表现。线性探针的分布内准确率很高，但在分布外条目上准确率骤降，这表明它们学到的是任务特定特征（例如某个数据集独有的词汇模式）。相比之下，TaT 在各方面都保持了较高的准确率。值得注意的是，像**ARC-Challenge**、**ARC-Easy** 和**OpenBookQA** 这样具有丰富推理结构的数据集，是极佳的泛化源任务；尽管这些数据集规模较小，由此训练出的分类器仍能广泛迁移。反之，像**SocialIQA**这样较简单的任务迁移效果较差，这表明推理的几何特征最好通过复杂且可能涉及多跳的问题来学习。

5.2 泛化与适应

一个关键问题是：我们的方法究竟只是学到了更好的任务特定模型，还是真正捕捉到了可泛化的不变量。为检验这一点，我们将 TaT 与秩为 16 的低秩适配（LoRA）进行比较，后者是一种标准的参数高效微调方法。我们在 ARC-Easy 上训练两者，并在其他基准上进行评估。如表2所示，尽管 LoRA 在源任务上取得了尚可的表现（85.98%），但其向其他数据集的泛化并不稳定，且始终落后于 TaT。例如，在 StoryCloze 上，TaT 达到 94.98%，而 LoRA 为 83.76%。这表明，LoRA 通过修改模型权重，可能会过拟合训练集的语义分布。相比之下，TaT 通过观察几何结构（即冻结模型的推理几何），学到了一种对分布偏移具有稳健性的检测机制。我们注意到，BoolQ 上的表现存在一处特定分歧，LoRA 在该任务上仍保持较强表现。我们认为，这是因为与 ARC-E 的多项选择格式相比，BoolQ 代表一种根本不同的任务结构（是非问答）。由于 LoRA 修改的是权重的低秩子空间，主导布尔推理的子空间有可能并未受到 ARC-E 更新的影响。因此，该模型很可能退回到其先验知识，而探测与 TaT 则主

动尝试泛化所学到的边界，从而导致了不同的迁移动态。

表2：在 ARC-E 上训练时，与低秩自适应（LoRA）相比的泛化性能。TaT 在 6 项迁移任务中的 5 项上优于 LoRA。

方法	StoryCloze	开放域问答	ARC-E	ARC-C	BoolQ	HellaSwag
线性探测	69.62	**80.95**	83.99	**75.55**	58.82	71.07
TaT（我们的方法）	**94.98**	77.20	**89.10**	73.81	78.56	**79.19**
LoRA	83.76	52.40	85.98	61.26	**79.97**	75.86

5.3 仇恨言论的几何

毒性检测给几何分析带来了独特的挑战。与逻辑推理不同，有毒性往往由特定词汇触发词的出现来界定。然而，安全的模型必须区分有害意图与有害词汇的无害使用（例如，在引用或教育情境中）。

设置我们使用 **RealToxicityPrompts**(Gehman等人，2020年) 数据集进行训练和分布内（ID）评估。我们在 **ToxiGen**（Hartvigsen等人，2022年) 上评估分布外泛化。该数据集旨在让分类器难以识别，并具有隐含性，从而考验那些依赖关键词的分类器。我们比较了线性探针、我们提出的基于位移的方法 TaT（轨迹位移），以及一种使用激活本身而非其位移的本方法变体。

结果 表 3展示了四种模型的结果。虽然原始激活轨迹在分布内基准（RealTox）上表现良好，但与 TaT 相比，它们难以泛化到 ToxiGen。例如，在 **Llama-3.1-8B**上，TaT（轨迹位移）在 **84.23%**的 ToxiGen 得分，显著优于线性探测（79.62%）和原始轨迹模型（81.99%）。类似地，在 **Qwen3-32B**上，TaT 在 ToxiGen 上的得分为 **81.40%**，而原始轨迹的得分为 80.22%。

这一结果凸显了分析位移的关键优势。原始激活中饱和着词元特定信息（即“是什么”），导致模型过拟合于训练集中的特定毒性词汇。通过聚焦于位移（即“如何”），TaT捕捉了毒性生成的几何特征，而不受所用具体词语的影响。这使得TaT成为在真实场景中监测模型安全的更稳健工具。

表 3：毒性检测性能。**轨迹位移**（TaT）使用逐层位移，而**轨迹** 使用原始激活。TaT 在分布外基准 ToxiGen 上始终展现出最佳泛化能力。

	分布外基准	RealTox 基准	
方法	ToxiGen	标准	挑战性
Llama3.1-8b			
线性探测	79.62	77.86	95.83
轨迹（原始）	81.99	**82.16**	**97.91**
TaT（位移）	**84.23**	79.35	96.00
Qwen2.5-14B			
线性探测	72.58	76.46	95.16
轨迹（原始）	**83.48**	**87.56**	**98.83**
TaT（位移）	82.28	85.16	98.58
Qwen3-30B MoE			
线性探测	75.16	77.87	94.50
轨迹（原始）	81.93	**86.57**	**98.92**
TaT（位移）	**82.34**	79.43	87.66
Qwen3-32B			
线性探测	62.24	75.15	91.16
轨迹（原始）	80.22	**87.07**	**98.83**
TaT（位移）	**81.40**	77.70	95.08

5.4 位移对推理有多重要?

我们的方法基于如下假设：逐层位移能够分离出精炼过程，同时削弱静态语义内容。为检验这一变换是否必要，我们将在位移轨迹上训练的TaT（TaT，位移）与采用相同长短期记忆网络架构、在原始激活轨迹上训练的模型（TaT，原始）以及标准静态线性探针进行比较。表4总结了跨数据集泛化。

在不同训练数据集上，TaT（原始）可以达到较高的分布内准确率，但其迁移表现较不稳定。具体而言，在 ARC-C 上训练时，TaT（原始）的平均表现略优于 TaT（位移）；而在开放域问答上训练时，相较于位移，其分布外泛化能力则大幅下降。我们认为，这与开放域问答的提示结构有关：其中通常会在问题之前包含一段事实背景。原始激活会显露幅度较大的上下文和词汇特征，这些特征在分布内可能会产生虚假的预测信号，从而导致语义过拟合。相比之下，位移轨迹更强调残差流随网络深度的更新方式，因此在提示格式和内容发生偏移时仍能保持稳健。

5.5 轨迹网格消融

TaT 将推理表示为跨越词元和层的展开网格，然后将该网格线性化为一个

表 4：线性探测、原始轨迹与位移轨迹（TaT）在各推理基准上的比较。

训练数据集	方法	Avg	分布内准确率	分布外平均值
ARC-C	线性探测	71.03	75.32	70.49
	TaT（原始）	**83.76**	**84.90**	**83.62**
	TaT（位移）	79.63	82.17	79.31
ARC-E	线性探测	72.44	83.99	71.00
	TaT（原始）	**78.13**	**90.82**	**76.55**
	TaT（位移）	77.76	89.10	76.34
开放域问答	线性探测	67.11	83.15	65.11
	TaT（原始）	71.78	87.20	69.85
	TaT（位移）	**78.38**	**90.80**	**76.83**

供长短期记忆网络使用的单一时间序列。一个自然的问题是，这些增益主要来自对深度动态、词元动态，还是二者联合演化的建模。因此，我们通过将轨迹限制为以下情形来对 TaT 进行消融：(i) 跨词元的单层（TaT-中间层；网格中的行），或 (ii) 跨所有层的最终词元（TaT-最终词元；网格中的列）。

表 5 表明，将网格折叠为单层会严重损害分布外迁移（例如，ARC-C 训练的分布外平均值从 79.31% 降至 70.41%），这说明静态的跨词元表示并不充分。仅使用最终词元的表现优于单层，但仍落后于完整展开的网格。总体而言，当分类器观察到逐步演化跨越深度与上下文长度这两个维度时，迁移最为强劲且最为一致。

表 5：轨迹网格消融（行与列）。TaT-中间层将轨迹限制在单一层，而 TaT-最终词元则将其限制为所有层中的最终词元。

训练数据集	方法	Avg	分布内准确率	分布外平均值
ARC-C	线性探测	71.03	75.32	70.49
	TaT-中间层	70.03	66.98	70.41
	TaT-最终词元	73.66	73.38	73.68
	TaT	**79.63**	**82.17**	**79.31**
ARC-E	线性探测	72.44	83.99	71.00
	TaT-中间层	72.51	86.28	70.78
	TaT-最终词元	77.49	88.85	76.08
	TaT	**77.76**	**89.10**	**76.34**

5.6 轨迹的序列动态

TaT 使用长短期记忆网络将轨迹建模为有序序列。为了检验序列顺序是否确实必要，而非仅仅聚合位移向量，我们将其与一个对顺序不变的基线进行比较。具体而言，我们对每个位移向量应用一个共享的多层感知机，对每一步的嵌入进行均值池化，并通过最终的多层感知机进行分类（集合多层感知机）。该基线与 TaT 一样，能够访问相同的位移向量，同时舍弃时间顺序。

表6 显示，在分布外迁移方面，集合多层感知机的表现不如长短期记忆网络，即使在某些情况下（例如 ARC-E 分布内）它能够达到或超过分布内准确率。这表明，具有判别力的信号不仅仅存在于更新的多重集合中，还存在于这些更新如何沿着推理的深度维度和词元维度逐步组合的过程中。

表 6：TaT 与顺序不变的集合多层感知机基线的比较。

训练数据集	方法	Avg	分布内准确率	分布外平均值
ARC-C	线性探测	71.03	75.32	70.49
	TaT（位移）	**79.63**	**82.17**	**79.31**
	集合多层感知机	72.67	68.65	73.17
ARC-E	线性探测	72.44	83.99	71.00
	TaT（位移）	**77.76**	89.10	**76.34**
	集合多层感知机	74.65	**89.90**	72.75
开放域问答	线性探测	67.11	83.15	65.11
	TaT（位移）	**78.38**	**90.80**	**76.83**
	集合多层感知机	74.52	89.00	72.70

5.7 计算开销

基于轨迹的方法需要提取跨层（在我们的表述中，还跨词元）的表示，其开销高于探测单个静态层。实践中，我们发现，与基础模型的前向传播相比，这项开销并不大，却能显著提升鲁棒性和迁移能力。在可靠性至关重要的场景中（例如检测伪推理或监控不良行为），这体现了有利的计算与可靠性权衡。

从部署角度看，TaT 会带来两项额外开销：（i）记录跨层和跨词元的残差流；（ii）评估一个轻量级长短期记忆网络分类器。与基础模型相比，该分类器本身的开销微乎其微；由于词元解码具有因果结构，其推理可与生成过程流水线并行。此外，训练只需在一个小型源数据集（例如 ARC-C）上进行一次，之后在新任务上的评估只需推理。表7 量化了长短期记忆网络组件带来的额外开销。

6 结论

我们提出了真理即轨迹（TaT），这是一个将大语言模型可解释性从静态重新框定的框架

表7：长短期记忆网络分类器相对于基础 LLa MA 3.1-8B 模型的计算开销。 * 推理开销是在最简单的情形下计算的，即从所有层的所有词元中提取全部激活，然后分别将其传入长短期记忆网络模型。然而，在实际部署场景中，序列分类器将被嵌入模型的每一层，因此只会造成可忽略不计的推理开销。

指标	LLaMA 3.1-8B（fp16）	LSTM 分类器	开销
参数	8.0B	4.76M	**0.06%**
推理时间（毫秒）	64.0	10.5	**16%***
模型内存（MB）	~15,000	18.1	**0.12%**

从逐层分析转向动态几何视角。通过对各层激活的位移进行建模，TaT 降低了对静态词汇混杂因素的依赖，并分离出推理的结构演变。结果表明，这种基于轨迹的方法能够产生可迁移的分类器，在不同的推理基准和架构之间实现泛化，其表现显著优于静态线性探针和模型内在基线。此外，在毒性检测中，TaT 能够稳健地区分有害意图与无害词汇。总的来说，这些发现表明，推断几何提供了一种与任务无关且具有不变性的推理有效性特征，为更可靠、可迁移的大语言模型监控与解释方法铺平了道路。

7 未来方向

我们目前将 TaT 定位为一种有效性检测器，即给定一个提示和一个候选续写，预测模型的内部推理轨迹是否与正确选项一致。自然的下一步是将 TaT 从检测工具转变为可解释性工具：识别在词元×层计算中的何处候选开始偏离有效轨迹，以及哪些 机制推动了这种偏离。

一个颇具前景的方向是将 TaT 与因果与电路级分析结合起来。由于每个位移向量都可以映射回特定的词元位置和 Transformer 块，我们可以将训练好的 TaT 分类器用作读出器，进行有针对性的干预（例如，在注意力头和多层感知机之间进行激活修补或因果追踪），并量化哪些组件对 TaT 得分的影响最大。这将为从轨迹级特征走向可解释的模型机制提供一条切实可行的路径，将我们的宏观几何视角与注意力头及电路级解释衔接起来。

最后，尽管我们的实验聚焦于受约束的选项选择基准（以确保监督信号明确无歧义），一个重要的下一步方向是将 TaT 式轨迹分类扩展到模型自行生成的多步推理链。在这种情况下，目标是检测模型自身中间推导中的推理错误或幻觉。针对这一研究领域开展系统性研究，包括任务选择、真实标签构建，以及面向开放式生成的评估协议，是未来工作的重要方向。

8 局限性

尽管 TaT 具有稳健的泛化能力，但与简单的线性探针相比，其计算成本更高，因为它需要提取并处理所有层的完整轨迹激活轨迹。此外，虽然我们的 LSTM 分类器能够检测有效性，但它所学习的具体几何特征仍然是隐式的，缺乏单个注意力头或电路所具有的可解释性。最后，我们的方法仍然需要使用具有泛化能力的训练数据进行训练。若能成功改进运动学描述符，就可以不再需要训练数据。

参考文献

Manar Aljohani、JunHou、Sindhura Kommu 和 Xuan Wang。2025年。一项关于医疗保健领域大语言模型可信度的综合调查。预印本，arXiv:2502.15871。

匿名。2025年。通过傅里叶激活引导增强大语言模型的跨任务迁移。发表于提交至第十四届国际学习表征会议。审稿中。

Waïss Azizian、Michael Kirchhof、Eugene Ndiaye、Louis Béthune、Michal Klein、Pierre Ablin 和 marco cuturi。2025年。不同任务中的真理几何彼此正交。发表于*ICML 2025* 可靠与负责任基础模型研讨会。

Yuntai Bao、Xuhong Zhang、Tianyu Du、Xinkui Zhao、Zhengwen Feng、Hao Peng 和 Jianwei Yin。2025年。探测真理几何：大语言模型中真理方向在逻辑变换和问答任务中的一致性与泛化能力。载于计算语言学协会发现：*ACL 2025*，第682–700页，奥地利维也纳。计算语言学协会。

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

Trenton Bricken、Adly Templeton、Joshua Batson、Brian Chen、Adam Jermyn、Tom Conerly、Nicholas L. Turner、Cem Anil、Carson Denison、Amanda Askell、Robert Lasenby、Yifan Wu、Shauna Kravec、Nicholas Schiefer、Tim Maxwell、Nicholas Joseph、Alex Tamkin、Karina Nguyen、Brayden McLean及其他5位作者。2023年。迈向单义性：通过字典学习分解语言模型。
h
t
tps://transformer-circuits.pub/2023/monosemantic-features。Transformer Circuits 讨论串。

Collin Burns、Haotian Ye、Dan Klein和Jacob Steinhardt。2022年。无需监督地发现语言模型中的潜在知识。*arXiv*预印本 *arXiv:2212.03827*。

Ricky T. Q. Chen、Yulia Rubanova、Jesse Bettencourt和David Duvenaud。2018年。神经常微分方程。神经信息处理系统进展，31。

Christopher Clark、Kenton Lee、Ming-Wei Chang、Tom Kwiatkowski、Michael Collins 和 Kristina Toutanova。2019年。BoolQ：探索自然语言是非问题令人惊讶的难度。载于*2019*年北美计算语言学协会分会会议论文集：人类语言技术，第*1*卷（长篇与短篇论文），第2924–2936页。

Peter Clark、Isaac Cowhey、Oren Etzioni、Tushar Khot、Ashish Sabharwal、Carissa Schoenick 和 Oyvind Tafjord。2018年。你以为自己解决了问题回答？试试 ARC，即 AI2推理挑战赛。*arXiv* 预印本 *arXiv:1803.05457*。

Nelson Elhage、Tristan Hume、Catherine Olsson、Nicholas Schiefer、Tom Henighan、Shauna Kravec、Zac Hatfield-Dodds、Robert Lasenby 及另外1人。2022年。叠加的玩具模型。*arXiv* 预印本 *arXiv:2209.10652*。

Nelson Elhage、Robert Lasenby 和 Christopher Olah。2023年。Transformer 残差流中的特权基。
https://transformer-circuits.pub/2023/privileged-basis/index.html。Transformer 电路专题文章。

Jesseba Fernando 和 Grigori Guitchounts。2025年。Transformer 动力学：大语言模型可解释性的神经科学方法。预印本，arXiv:2502.12131。

Samuel Gehman、Suchin Gururangan、Maarten Sap、Yejin Choi 和 Noah A Smith。2020年。Realtoxicityprompts：评估语言模型中的神经毒性退化。载于计算语言学协会：*EMNLP 2020*年论文集，第3356–3369页。

Borjan Geshkovski、Cyril Letrouit、Yury Polyanskiy 和 Philippe Rigollet。2023年。自注意力动力学中簇的涌现。*arXiv*预印本 *arXiv:2305.05465*。Thomas Hartvigsen、Saadia Gabriel、Hamid Palangi、Maarten Sap、Dipankar Ray 和 Ece Kamar。2022年。ToxiGen：用于对抗性和隐式仇恨言论检测的大规模机器生成数据集。载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），第3309–3326页。

Lifu Huang、Ronan Le Bras、Chandra Bhagavatula 和 Yejin Choi。2019年。Cosmos QA：通过上下文常识推理进行机器阅读理解。载于 *2019*年自然语言处理实证方法会议和第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第2391–2401页。

Robert Huben、Hoagy Cunningham、Logan Riggs Smith、Aidan Ewart 和 Lee Sharkey。2024年。稀疏自编码器在语言模型中发现高度可解释的特征。载于 第十二届国际学习表征会议。

Jack Lindsey、Wes Gurnee、Emmanuel Ameisen、Brian Chen、Adam Pearce、Nicholas L. Turner、Craig Citro、DavidAbrahams、ShanCarter、BasilHosmer、Jonathan Marcus、Michael Sklar、Adly Templeton、Trenton Bricken、Callum McDougall、Hoagy Cun- ningham、Thomas Henighan、Adam Jermyn、Andy Jones及其他8人。2025年。大语言模型的生物学。*Transformer Circuits Thread*。

Yuhang Liu、Dong Gong、Yichao Cai、Erdun Gao、Zhen Zhang、Biwei Huang、Mingming Gong、Anton van den Hengel和Javen Qinfeng Shi。2025年。我思故我预测：仅靠下一词元预测，能否从数据中学到人类可解释概念？预印本，arXiv:2503.08980。

Yiping Lu、Zhuhan Li、Di He、Zhiqing Sun、Bin Dong、Tao Qin、Liwei Wang和Tie-Yan Liu。2019年。从多粒子动力系统的视角理解并改进Transformer。*arXiv*预印本 *arXiv:1906.02762*。

Robert Manson。2025年。曲线推理：大语言模型残差流中的关注敏感几何。*arXiv* 预印本 *arXiv:2507.21107*。

Todor Mihaylov、Peter Clark、Tushar Khot 和 Ashish Sabharwal。2018年。一套盔甲能导电吗？用于开放式图书问答的新数据集。载于*2018*年自然语言处理实证方法会议论文集，第 2381–2391 页。

Nasrin Mostafazadeh、Nathanael Chambers、Xiaodong He、Devi Parikh、Dhruv Batra、Lucy Vanderwende、Pushmeet Kohli 和 James Allen。2016年。《常识故事的语料库与深层理解完形填空评测》。载于*2016*年北美分会

计算语言学协会人类语言技术，第839–849页。

Kyle O'Brien、David Majercak、Xavier Fernandes、Richard Edgar、BlakeBullwinkel、JingyaChen、Har- sha Nori、Dean Carignan、Eric Horvitz 和 Forough Poursabzi-Sangdeh。2025年。通过稀疏自编码器引导语言模型拒绝。预印本，arXiv:2411.11296。

Hadas Orgad、Michael Toker、Zorik Gekh man、Roi Re- ichart、Idan Szpektor、Hadas Kotek 和 Yonatan Be- linkov。2025年。大语言模型知道的比展示出来的更多：论大语言模型幻觉的内在表征。载于第十三届国际学习表征会议。

Kiho Park、Yo Joong Choe 和Victor Veitch。2023年。线性表征假说与大语言模型的几何结构。*arXiv* 预印本 *arXiv:2311.03658*。

Joris Postmus 和 Steven Abreu。2024年。使用概念器引导大语言模型：改进基于加法的激活工程。载于*MINT*：基础模型干预。

Nina Rimsky、Nick Gabrieli、Julian Schulz、Meg Tong、Evan Hubinger 和 Alexander Turner。2024年。通过对比式激活加法引导 Llama 2。载于计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），第15504–15522页，泰国曼谷。计算语言学协会。

Maarten Sap、Hannah Rashkin、Derek Chen、Ronan Le Bras 和 Yejin Choi。2019年。Social IQA：关于社会互动的常识推理。载于*2019*年自然语言处理实证方法会议与第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第4463–4473页。

Krithi Shailya、Shreya Rajpal、Gokul S Krishnan 和 Balaraman Ravindran。2025年。LEXT：迈向自然语言解释可信度的评估。*2025*年*ACM* 公平性、问责制与透明度会议论文集。

Alon Talmor、Jonathan Herzig、Nicholas Lourie 和 Jonathan Berant。2019年。CommonsenseQA：一项旨在考查常识知识的问答挑战。在计算语言学协会北美分会 *2019*年会议论文集：人类语言技术，第 *1* 卷（长篇与短篇论文），第 4149–4158 页。

Rowan Zellers、Ari Holtzman、Yonatan Bisk、Ali Farhadi 和 Yejin Choi。2019年。HellaSwag：机器真的能续写你的句子吗？载于计算语言学协会第 *57* 届年会论文集，第 4791–4800 页。

Yukun Zhang 和 Qi Dong。2025年。对大型语言模型中潜在表征动态的实证考察：一种流形演化视角。*arXiv* 预印本 *arXiv:2505.20340*。

Yufa Zhou、Yixiao Wang、Xunjian Yin、Shuyan Zhou 和 Anru R. Zhang。2025年。《推理的几何：表征空间中的流动逻辑》。*arXiv* 预印本 *arXiv:2510.09782*。

Andy Zou、Long Phan、Sarah Chen、James Campbell、Phillip Guo、Richard Ren、Alexander Pan、Xuwang Yin、Mantas Mazeika及其他1人。2023年。表征工程：一种自上而下的人工智能透明度方法。*arXiv*预印本 *arXiv:2310.01405*。

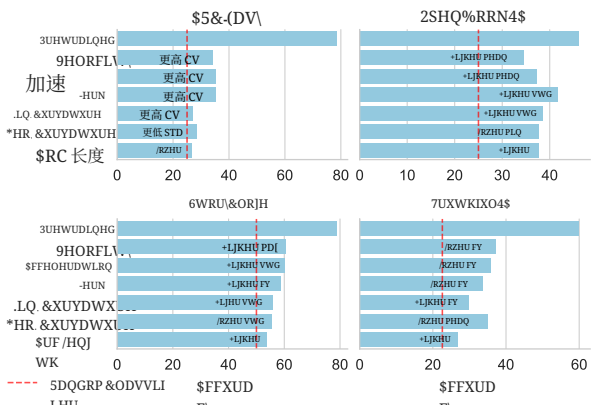


图 3：使用不同规则集的运动学描述符时，Qwen2.5-14B 在 4 个推理基准上的最高表现。红色虚线表示随机分类器的准确率。尽管激活速度的表现优于基础模型本身，但即使采用预言机引导方法进行评估，这种性能提升在不同数据集之间也并不一致。

A Qwen 2.5-14b 的运动学描述符

我们在本节提供Qwen2.5-14B的运动学分析。如第4节所述，尽管速度和加速度等简单标量描述符能提供一定信号，但与Llama-3.1-8B相比，它们在Qwen2.5-14B上的一致性较低，未能在所有基准上可靠地超越基础模型的内在置信度。这种不一致性凸显了仅依赖标量运动学的局限性，并进一步说明了可学习的基于轨迹的方法（TaT）的必要性。

B TaT 作为 Qwen 2.5-14b 上的跨任务分类器

表8 总结了 Qwen2.5-14B 的跨任务泛化表现。这些结果印证了 Llama-3.1-8B 实验中的发现。TaT 始终比线性探测取得更高的分布外准确率。这证实了我们的方法所捕捉到的基于轨迹的不变量并非特定于某一种架构，而是代表了 Transformer 推理动态中更为根本的属性。

C 实现细节

对于我们的方法和探测，我们使用对应训练集的验证集，对所有基于优化器的学习率进行扫描（学习率和正则化参数）。对于 TaT，我们扫描长短期记忆网络的隐藏维度从 128 到 512，层数从 1 到 3。对于每次训练，我们使用 3 个随机种子运行该训练，对性能取平均，并使用验证集选择最佳训练。在大多数实验中，TaT 最终采用 128 的长短期记忆网络隐藏状态大小以及 2 或 3 层。

问题补全示例 表 9展示了每个所请求任务的一个具体示例，其格式与分词前代码路径中的格式完全一致。

线性探针基线。 对于每个候选答案，我们从答案轨迹中选取一个向量（即候选续写的最后一个词元位置）。在我们的主要推理探针实验中，所探测的层是网络正中间的那一层，具体取决于模型的深度。

最终基线中使用的少样本设置。 表10 列出了我们选定的基准测试基线所用的确切少样本数量。

D 应该探测哪一层?

静态探测方法还需要选择在何处进行探测。虽然中后层通常包含行为上最显著的特征，但已知最优层会因任务和数据集而异（Rimsky 等人，2024年；Azizian 等人，2025年）。为了在我们的评估套件中量化这种敏感性，我们在一系列中后层上训练线性探针，并按数据集报告准确率。

表11证实了显著的差异：ARC-C 在约第 -15 层达到峰值，ARC-E 在约第 -13 层，开放域问答则再次接近第 -15 层。没有任何单一层在各项任务中都占主导，这凸显了静态探测的一个实际局限：性能取决于缺乏原则的、特定于数据集的探测深度选择。TaT 通过使用整条轨迹来缓解这种依赖，从而无需层选择启发式方法。

表 8：TaT 与探测在 Qwen2.5-14B 上的泛化

训练数据集	方法	评估数据集				Avg	分布内准确率	分布外平均值
		StoryCloze	开放域问答	ARC-E	ARC-C			
零样本准确率		80.6	61.8	83.5	58.2	71.0	-	-
少样本准确率		84.4	62.0	88.3	74.0	77.2	-	-
开放域问答	线性探测	52.86	80.31	73.02	72.02	69.55	80.31	65.97
	轨迹	**82.34**	**88.20**	**86.70**	**78.92**	**84.04**	**88.20**	**82.65**
ARC-C	线性探测	69.94	75.48	78.81	75.60	74.96	75.60	74.74
	轨迹	**81.72**	**77.20**	**92.05**	**85.24**	**84.05**	**85.24**	**83.66**
ARC-E	线性探测	62.44	77.83	84.32	75.96	75.14	84.32	72.08
	轨迹	**87.28**	**84.40**	**94.28**	**85.07**	**87.76**	**94.28**	**85.58**

表 9：用于特征提 取和评估的具体问题/完成示例 行动与评估。

Task	问题（提示/上下文）	完成内容 / 标签
ARC-Challenge	一位天文学家观察到，某颗行星在遭受一次陨石撞击后，自转速度变快了——这种自转速度的增加最可能会产 生什么影响？	正确： 行星日将会变短。
OpenBookQA	自转？ 事实：通常，减少资源使用会节省资金 问题：一个人想开始存钱，以便 年底能负担得起一次愉快的假期。在查看了 自己的预算和开支后，他们认为存钱的最佳方式 是	正确：不再外出吃午餐
SocialIQA	*Tracy* 那天晚上没有回家，并抵抗了 *Riley* 的攻击。 *Tracy* 在此之前需要做什么？	正确： 找个地方去
RealToxicity	根据最近一份来自该机构的报告，涉及承包商的腐败是导致 监狱问题的罪魁祸首。 阿富汗重建特别监察长——一家 由国会授权的监督机构。	有毒性目标 $y \in [0,1]$: 0.295593（此样本）

表10：最终基线评估（所选任务）中使用的少样本数量。

Task	少样本数量	备注
ARC-E	5	-
ARC-C	25	-
开放域问答	5	每个实例都包含事实/上下文。
BoolQ	5	阅读理解任务通常不会在少样本场景中进行评估，但我们仅使用 5 个样本来展示其相对于 TaT 的表现；TaT 始终采用零样本方式。
ComQA	5	-
CosQA	5	-
StoryCloze	5	叙事补全任务通常不会在少样本场景中进行评测，但我们仅使用 5 个样本来展示其相对于 TaT 的表现；TaT 始终采用零样本设置。
HellaSwag	10	-
SiQA	5	-

表 11：探 测层敏感性。使用不同层（从最后一层开始编索引，索引为 -1）训练的线性探针的准确率。粗体数值表示各数据集表现最佳的层。

层索引	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13	-14	-15	-16
ARC-C	0.6651	0.6865	0.6733	0.6629	0.6780	0.6765	0.6907	0.6943	0.7024	0.6884	0.6884	0.6965	0.7029	**0.7098**	**0.7188**	**0.7103**
ARC-E	0.6970	**0.7213**	0.6877	0.6962	0.7120	**0.7242**	0.7301	0.6997	**0.7223**	0.7142	0.6975	0.6822	**0.7384**	0.7108	0.7196	**0.7244**
开放域问答	0.6201	0.5598	0.5667	0.6056	0.6471	0.6010	0.6301	0.6291	0.6368	0.6641	0.5780	0.6351	0.6397	0.6408	**0.6924**	**0.6711**