

大语言模型空间推理能力的内部表征机制研究

安纪元^{1,2}, 杨麟儿^{1,2}, 王梦焰^{1,2}, 鲁鹿鸣³, 安维华², 杨尔弘^{1,2}

(1. 北京语言大学国家语言资源监测与研究平面媒体中心, 北京 100083;

2. 北京语言大学信息科学学院, 北京 100083;

3. 北京师范大学人工智能学院, 北京 100875)

摘要: 空间推理是大语言模型 (LLMs) 迈向规划、导航与具身交互的重要能力基础。然而, 模型在空间推理基准上给出正确答案, 并不必然表明其内部已经形成结构化、可组合的空间表征; 相应行为也可能源于语言启发式、选项偏置或对任务格式的适配。为澄清这一问题, 本文从机制分析视角出发, 将空间推理分解为关系空间推理、视角旋转变换与空间序列执行三类核心计算基元, 并在英语、汉语和阿拉伯语三种语言环境下, 结合行为评估、线性探测、稀疏自编码器特征分析与因果干预, 系统考察 LLMs 中空间信息的编码位置、功能可用性及语言依赖性。实验结果表明, 任务相关空间变量主要分布于模型中间层, 且部分空间选择性特征会对模型输出产生因果影响; 空间信息的可解码性并不等同于可使用性, 基础模型中存在明显的空间编码与任务行为脱耦现象; 相近的外部准确率可能由不同语言条件下的差异化内部特征所支撑, 反映出跨语言机制退化与任务特异性。上述发现表明, 当前 LLMs 并非完全依赖浅层语言线索完成空间推理, 但也尚未形成统一、稳定且可迁移的空间认知机制。本文为超越基准准确率、从内部表征与因果机制层面评估 LLMs 的空间能力提供了新的实证视角。¹

关键词: 大语言模型; 空间推理; 机制可解释性; 线性探测; 稀疏自编码器; 因果干预; 多语言表征

中图分类号: TP391 **文献标识码:** A

A Study on the Internal Representational Mechanisms Underlying Spatial Reasoning in Large Language Models

Jiyuan An^{1,2}, Liner Yang^{1,2}, Mengyan Wang^{1,2}, Luming Lu³, Weihua An², Erhong Yang^{1,2}

(1. National Language Resources Monitoring and Research Center for Print Media, Beijing Language and Culture University, Beijing, 100083, China;

2. School of Information Science, Beijing Language and Culture University, Beijing, 100083, China;

3. School of artificial intelligence, Beijing normal university, Beijing, 100875, China)

Abstract: Spatial reasoning is a fundamental capability for large language models (LLMs) to support planning, navigation, and embodied interaction. However, producing correct answers on spatial reasoning benchmarks does not necessarily indicate that a model has developed structured and compositional spatial representations; such behavior may instead arise from linguistic heuristics, option biases, or adaptation to task formats. To clarify this issue, this paper adopts a mechanistic analysis perspective and decomposes spatial reasoning into three core computational primitives: relational spatial reasoning, perspective rotation, and spatial sequence execution. Across English, Chinese, and Arabic settings, we combine behavioral evaluation, linear probing, sparse autoencoder feature analysis, and causal intervention

¹<https://github.com/JiyuanAn/LLM-Spatial-Mechanisms>

to systematically examine where spatial information is encoded in LLMs, whether it is functionally usable, and how it depends on language. Experimental results show that task-relevant spatial variables are primarily represented in the middle layers of the model, and that some spatially selective features exert a causal influence on model outputs. Moreover, the decodability of spatial information does not necessarily imply its usability, revealing a clear dissociation between spatial encoding and task behavior in base models. Similar external accuracies may also be supported by different internal features across languages, suggesting cross-lingual mechanistic degradation and task specificity. These findings indicate that current LLMs do not rely solely on shallow linguistic cues for spatial reasoning, yet they have not developed a unified, stable, and transferable spatial cognitive mechanism. This work provides a new empirical perspective for evaluating the spatial abilities of LLMs beyond benchmark accuracy, focusing instead on internal representations and causal mechanisms.

Keywords: large language models; spatial reasoning; mechanistic interpretability; linear probing; sparse autoencoders; causal intervention; multilingual representations

0 引言

大语言模型 (LLMs) 与视觉语言模型 (VLMs) 正在从文本生成工具走向具备规划、决策和交互能力的通用系统, 并被用于导航指令理解、具身交互和机器人控制等高度依赖空间理解的场景 [1], [2], [3], [4]。世界模型研究同样指出, 结构化内部表征是预测、规划与控制的基础 [5]。因此, 一个比“模型在空间推理问题上能答对多少”更关键的问题是: LLMs 是否真正形成并调用了空间表征, 还是主要借助语言共现、答案偏置和任务格式获得表面成功 [6], [7], [8]?

近期空间推理基准显示, 模型规模扩大、指令微调和链式思维提示往往可以提升 LLMs 的外显表现 [6], [8], [9]。然而, 准确率只能说明模型“做出了什么选择”, 难以解释其依据。正确输出可能来自结构化空间计算, 也可能来自语言启发式、记忆、选项偏置或提示诱导; 错误输出也不必然说明模型完全没有编码空间信息 [10], [11], [12], [13]。换言之, 行为表现与内部表征可能脱耦, 这是基准评测难以揭示的盲区。

这一局限与人类空间认知研究形成对照。在认知科学中, 空间能力不仅由行为正确率刻画, 更通过认知地图、心理旋转、路径积分等内部机制加以解释; 这些机制性概念帮助研究者理解泛化、鲁棒性和系统性错误。相较之下, 关于大型神经模型空间推理的大多数研究仍主要报告任务表现或提示策略效果, 很少直接追踪隐藏状态中的空间变量及其功能作用 [6], [8], [14], [15]。因此, LLMs 空间行为背后的认知基础仍然不透明。

这种机制理解的缺失会同时影响系统评估与理论解释: 若不知道模型在何处、以何种形式保存空间状态, 就难以判断其在分布偏移、多语言输入或长程操作中的失效风险; 缺乏表征证据时, 模型与人类的行为相似性也只能停留在现象类比 [16], [17]。

本文认为, 基础模型空间推理研究需要从“测准不准”转向“看见并检验内部机制”。本文不仅考察模型是否答对, 更追问两个机制性问题: 隐藏状态中是否存在可解码的空间变量; 这些变量是否参与输出生成。受认知科学启发, 我们将空间能力分解为核心计算基元, 并在标准单次前向推理设置下构建受控任务族, 以更直接地观察隐式空间计算。

本文提出三类空间任务族: 关系空间推理、视角旋转变换和空间序列执行 (如图 1)。为分离空间计算与语言形式的影响, 本研究构建英语、汉语和阿拉伯语的平行任务, 在保持计算等价的前提下允许表层表达自然变化。我们进一步结合行为评估、线性探测、基于稀疏自编码器的特征分析以及因果干预, 定位空间信息在模型内部的出现位置, 检验其是否具有行为功能, 并比较不同语言下的实现路径。本文的主要贡献如下:

- 1) 构建面向机制分析的三类跨语言文本空间推理任务, 覆盖关系整合、参考系变换和状态更新。

- 2) 结合线性探测、SAE 特征分析与因果干预，检验空间信息从可解码表征到行为功能的转化。
- 3) 揭示 LLMs 空间能力的关键机制边界：空间表征主要形成于中间层，却未稳定接入输出决策，并在跨任务、跨语言条件下呈现碎片化实现。

通过结合认知理论与机制可解释性，本文澄清了基础模型空间能力的性质与边界，并为超越基准准确率的空间推理评测提供实证依据。

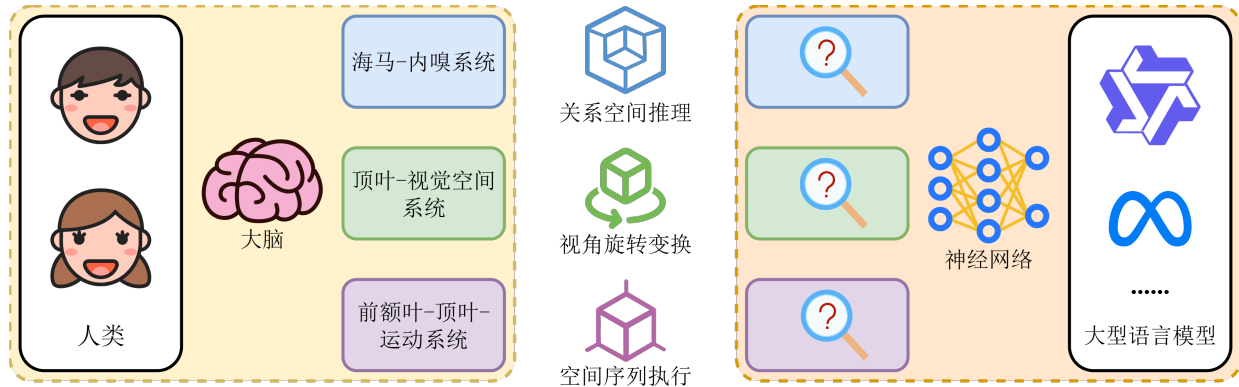


图 1: 从人类的空间认知到大型语言模型中的空间表征

1 相关工作

1.1 空间认知与计算机制

空间认知长期被认知心理学与神经科学视为人类智能的核心组成部分。经典心理测量框架区分了空间知觉、心理旋转和可视化等能力 [18], [19]; 后续研究进一步指出，空间行为由多种内部表征与计算机制协同产生 [20], [21]。认知地图支持超越即时感知的关系推断 [22]，心理旋转与视角采择说明空间表征可经历连续几何变换 [19]，导航研究则强调空间状态的维持和更新 [23]。这些发现提示，空间能力更适合被理解为若干计算基元的组合。本文采纳这一计算视角指导任务设计，但不预设 LLMs 复现了人类空间认知的神经实现。

1.2 大语言模型的空间推理与评测基准

随着大语言模型在复杂推理和交互式任务中的应用扩展，空间推理逐渐成为检验结构化世界知识和状态建模能力的重要方向。现有文本空间推理基准覆盖空间关系判断、组合式空间描述理解以及类导航推理等任务 [24], [6], [25]。StepGame[26]、SpatialEval[27] 等数据集被广泛用于考察多步空间推断，但核心评价仍主要是输出准确率。已有结果表明，LLMs 可在部分文本空间任务上取得非平凡表现，且链式思维提示、显式参照结构或可视化式推理可进一步提升性能 [13], [28], [29]。然而，模型也对表层改写、组合复杂度和分布偏移较为敏感 [30], [31]。这说明准确率虽能反映外显行为，却难以区分正确答案来自结构化空间计算、语言启发式、提示策略还是选项偏置。相关综述也指出，行为成功与结构化、组合性空间表征证据之间仍存在鸿沟 [32], [33]。因此，仅扩大基准覆盖面不足以回答模型“如何”完成空间推理。本文将任务变量设计为可被表征分析和因果干预直接访问的对象，进一步考察隐藏状态中的空间变量、可解释特征及其行为作用。

1.3 LLMs 的探测与机制分析

另一类研究关注神经语言模型内部编码了何种信息 [34], [35], [6], [36], [37]。线性探测用于检验特定变量是否可从隐藏状态解码；近期研究进一步分析表征几何结构，或借助稀疏自编码器识别可解释特征 [14], [38], [39], [40], [41]，也可通过因果干预测量行为效应 [42]。尽管机制分析已被用于多种语言和语义现象，其在空间推理中的应用仍然有限 [43], [44], [45], [46], [47]。更重要的是，单独的探测结果只能说明某个变量“可被读出”，不能证明模型自身“正在使用”该变量。本文因此将线性探测、SAE 特征分析和因果干预放在同一框架中，旨在从相关性证据推进到功能性与因果性理解。

综上，现有 LLM 空间推理研究多依赖行为基准，尚难揭示内部机制；机制可解释性研究虽已用于语言与语义现象分析，但对空间认知相关计算的讨论仍有限。本文以空间认知中的核心计算成分设计任务，并结合探测、特征分析与因果干预，在跨模型、跨语言条件下检验 LLMs 是否形成、保留并使用内部空间表征。

2 空间推理任务的计算分解

为区分大语言模型中的空间计算与表层语言模式匹配，本文提出基于计算分解的任务分类体系。不同于按领域、数据来源或表面题型组织任务，本研究依据正确推理所需的最小计算基元划分任务，使每类任务都对应明确的中间空间变量，并可被后续探测和干预直接访问。

本文识别出三类空间计算成分：（1）关系空间推理，检验局部关系能否整合为全局结构；（2）视角旋转变换，检验模型能否对既有空间状态执行参考系或朝向变换；（3）空间序列执行，检验模型能否在多步操作中维持并更新空间状态。三类任务分别对应空间结构建立、空间状态变换和空间状态更新，构成便于机制分析的测试套件。图 2 展示了三类任务族及其代表性示例。

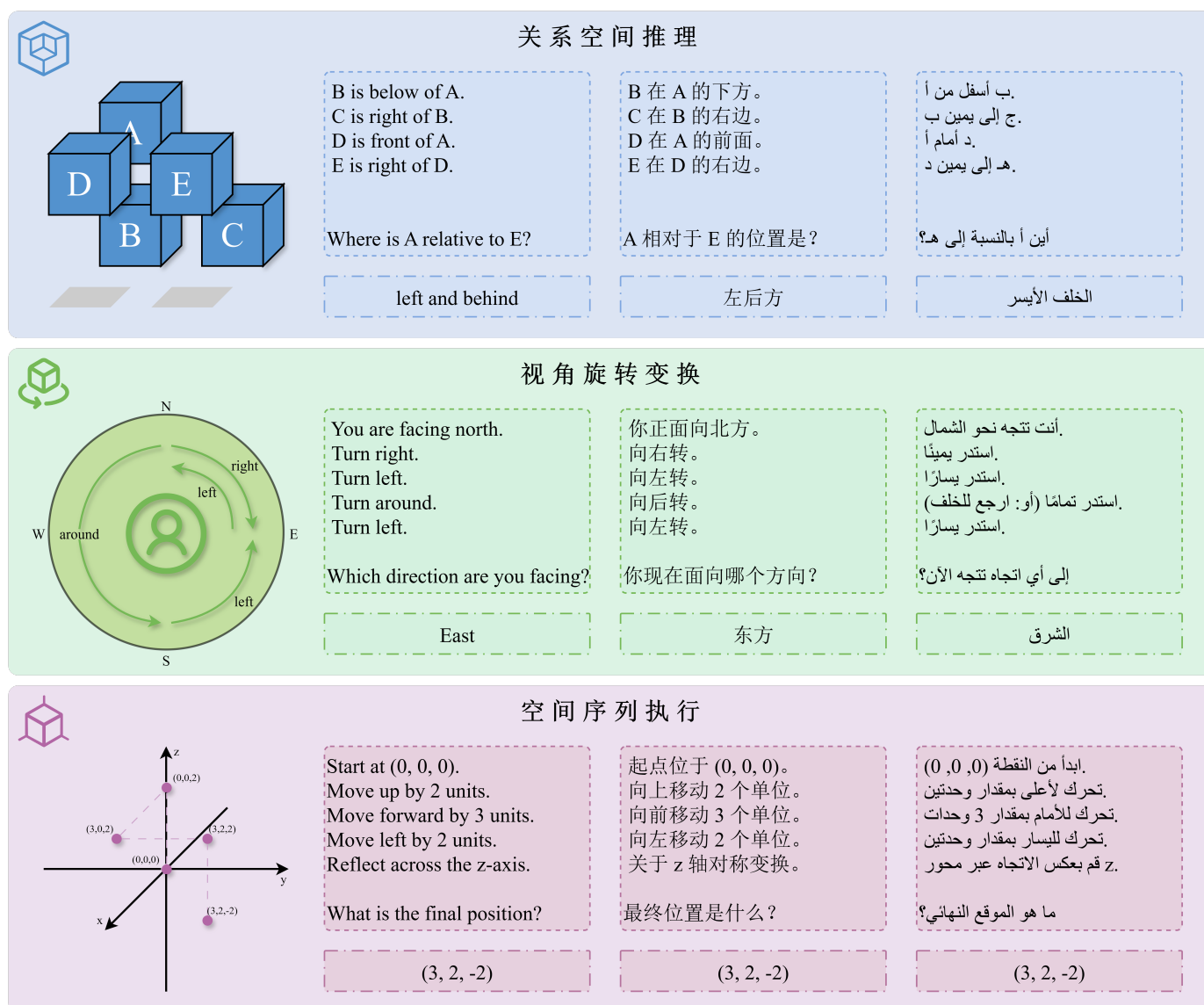


图 2: 本文提出的三类空间任务族示意图

2.1 任务设计原则

所有任务族遵循四项共同原则。**抽象性**：使用抽象实体与关系，减少世界知识依赖。**组合性**：要求整合多个空间约束或操作，测试模型是否能组合使用空间信息。**参数化**：通过实体数量、步骤数或维度控制难度，以分析性能随复杂度变化的规律。**机制可及性**：明确定义中间空间变量，使线性探测、SAE 特征分析和反事实干预指向同一组可解释对象。所有任务均在标准单次前向推理条件下评估，不诱导显式推理轨迹；任务由受控规则程序生成，以保证跨任务和跨语言的计算等价性。

2.2 三类空间任务族

本文构建三类空间任务族，分别对应空间结构建立、空间状态变换和空间状态更新。

关系空间推理任务。关系空间推理构成空间理解能力的基础层。其核心在于，模型能否根据多个局部关系（如 A 在 B 左侧、 B 在 C 上方）建立全局一致的空间结构，并推导未被直接给出的隐含关系。该任务要求模型把成对关系组织为整体空间网络，而非只识别单个空间词。

视角旋转变换任务。视角旋转变换体现空间理解能力的操作性维度。该任务要求模型在保持内部一致性的前提下，对已有空间状态实施旋转、镜像、视角切换或参考系转换，并报告变换后的朝向或关系。与关系推理关注“结构如何建立”不同，视角变换关注“已有结构能否被规则化操作”，因而可检验内部表征是否具有几何一致性。

空间序列执行任务。空间序列执行体现空间理解能力的过程性要求。每个样本给定初始位置和一系列移动或变换命令，模型需在执行全部步骤后计算最终位置。该任务突出时间维度上的状态维护：模型不仅要理解单步操作，还要将多个操作累积为一致的最终空间状态。若中间状态无法保存，误差将在后续步骤中放大。因此，空间序列执行抽象了导航与路径积分的计算核心。

2.3 与人类空间认知机制的功能对应

本文分类体系以计算为导向，与人类空间认知中的经典区分存在功能类比：关系推理类似认知地图中的空间整合与推断，视角变换类似心理旋转中的几何操作，空间序列执行类似导航中的路径积分与动态更新。该类比为任务设计提供理论锚点，并不预设神经实现等同。

2.4 跨语言任务构建

为区分空间计算与语言特异性线索，本文在英语、汉语和阿拉伯语中分别构建所有任务族。任务并非逐句直译，而是在保持核心空间逻辑、约束数量、推理步数和答案唯一性一致的前提下，允许各语言采用自然表达。若模型形成稳定的抽象空间机制，相似空间变量应在不同语言中呈现相近的层级分布与功能特征；若表征依赖具体语言路径，则可能出现外部行为相近但内部机制不同的现象。

3 实验设置与机制分析方法

在上一节完成任务分解后，本节说明如何将其转化为可复现实验，并追踪模型内部的空间信息。本文建立从行为表现到内部表征、再到因果作用的分析链条：首先在跨语言任务集上评估外部行为并提取各层隐藏状态；随后通过线性探测定位空间变量的层级分布；进一步借助稀疏自编码器分离可解释特征；最后通过激活修补与特征消融检验这些表征是否参与输出生成，从而区分“可读出”、“被使用”和“具有因果影响”三种证据。



图 3: 本文分析框架概览

3.1 实验设置

模型：本文评估两个多语言 LLM 家族，即 Qwen2.5-7B[48] 和 Llama-3-8B[49]，并同时纳入基础模型与指令微调模型。所有实验均使用公开检查点，不进行任务特定微调，以考察预训练和指令微调后形成的一般空间表征及其行为可用性。

推理策略：所有任务均采用标准单次前向推理，不使用链式思维提示或显式推理轨迹。模型通过输出单个 token 在多项选择格式中给出答案，以减少外部草稿、解释文本和提示工程的影响，使后续探测和干预更集中于隐式空间表征。

数据与评估：对于每个任务族和每种语言，本文构建 2,000 个训练样例和 200 个留出测试样例，且各难度水平均匀分布。测试数据从不用于探测器训练或 SAE 训练。除非另有说明，结果均报告于同分布测试集。表 I 汇总了数据集统计信息。

任务族	训练集	测试集	难度范围（任务跳数）
关系空间推理	2 000	200	3–10
视角旋转变换	2 000	200	2–10
空间序列执行	2 000	200	2–10
每种语言合计	6 000	600	–
三种语言合计	18 000	1 800	–

表 I: 各任务族的数据集统计

评价指标：本文报告行为准确率，并分析性能随任务复杂度的变化。对于表征分析，本文使用线性探测 R^2 、回归误差（MAE/RMSE）和层级趋势，刻画空间信息在模型内部出现的位置。行为指标和表征指标分开报告，以避免把“可读出变量”直接等同于“能可靠完成任务”。

多语言设计：所有任务分别在英语、汉语和阿拉伯语中独立构建。提示并非从某一种语言直接翻译，而是在允许自然语言差异的同时保持计算等价。所有模型均在三种语言上评估，且不进行语言特定调优。

模型	语言	关系空间推理任务准确率	视角旋转变换任务准确率	空间序列执行任务准确率
Llama3-8B-Instruct	EN	49.0 (98/200)	28.0 (56/200)	62.0 (124.0/200)
	ZH	47.5 (95/200)	28.0 (56/200)	52.5 (105/200)
	AR	31.0 (62/200)	26.0 (52/200)	57.5 (115/200)
Qwen2.5-7B-Instruct	EN	50.0 (100/200)	23.0 (46/200)	55.5 (111/200)
	ZH	39.5 (79/200)	23.5 (47/200)	58.0 (116/200)
	AR	39.5 (79/200)	24.5 (49/200)	47.0 (94/200)
Qwen2.5-7B-Base	EN	4.0 (8/200)	2.0 (4/200)	2.0 (4/200)
	ZH	0.0 (0/200)	2.5 (5/200)	0.0 (0/200)
	AR	3.5 (7/200)	3.0 (6/200)	0.5 (1/200)

表 II: 不同任务族和语言上的准确率（%）

如表 II所示，模型在空间任务上的行为表现明显不均衡：关系推理和空间序列执行达到中等准确率，而视角旋转变换普遍接近随机水平，说明空间推理不能被视为单一能力。性能也未随复杂度单调下降：关系推理和视角变换中，准确率随步骤长度非单调波动，提示内部处理可能存在策略切换；空间序列执行则表现出更稳定的退化趋势，说明基于坐标的更新更容易被模型以数值或模板化路径处理。

跨语言评估显示，模型表现具有中等语言依赖性。总体而言，英语和汉语表现相近，阿拉伯语在关系推理任务上明显落后。空间序列执行的跨语言差异最小，提示坐标格式提供了更接近语言无关的计算接口；关系词和参考系表达则更易受具体语言路径影响。

按难度（跳数/步数）细分后，关系推理和视角变换准确率呈非单调波动，在中等难度区间骤降后局部回升，提示模型可能切换至粗粒度启发式策略；空间序列执行准确率则近似单调下降，与逐步累积误差一致。非单调区段的统计显著性经分段线性回归的 Chow 检验确认 ($p < 0.05$)。

3.2 空间变量的线性探测

本文训练线性探测器，从模型隐藏状态中解码任务相关空间变量，包括相对位置向量（任务族 1）、朝向向量（任务族 2）和绝对空间坐标（任务族 3）。探测器在训练集上训练，并在留出测试集上评估；文中报告 R^2 、MAE 和 RMSE 作为解码质量指标。线性可解码性并不直接等同于模型真正“使用”该信息，而是定位候选空间表征的第一步。为评估解码指标是否高于随机水平，本文将训练标签随机打乱后训练探测器，重复 100 次获得零分布。所有报告的非负 R^2 值均显著高于零分布的 95% 分位数 ($p < 0.01$ ，经 Bonferroni 校正)。视角变换任务中的负 R^2 表明探测器预测误差超过均值预测基线，与其接近随机的行为表现一致。

跨模型和跨任务的层级曲线显示出明显模型特异性（图 4）。Qwen 系列的空间信息通常在中间层达到峰值，并在最终层附近衰减；Llama3-8B-Instruct 的空间信息虽留存至最终层附近（第 31 层），但强度偏弱。视角旋转变换任务在所有模型和语言中均表现出较低甚至为负的 R^2 ，说明朝向变量难以被稳定解码，也解释了该任务的行为低迷。

指令微调并非空间表征强度的决定性因素：Qwen2.5-7B-Base 在关系推理上的探测 R^2 (0.382) 高于其指令微调版本 (0.366)，空间序列执行的 R^2 (0.394) 也与指令微调模型 (0.402) 几乎持平。然而，基础模型在行为层面几乎无法完成这些任务，表明空间能力的瓶颈不只在于“是否编码”，更在于能否将空间信息接入输出决策。

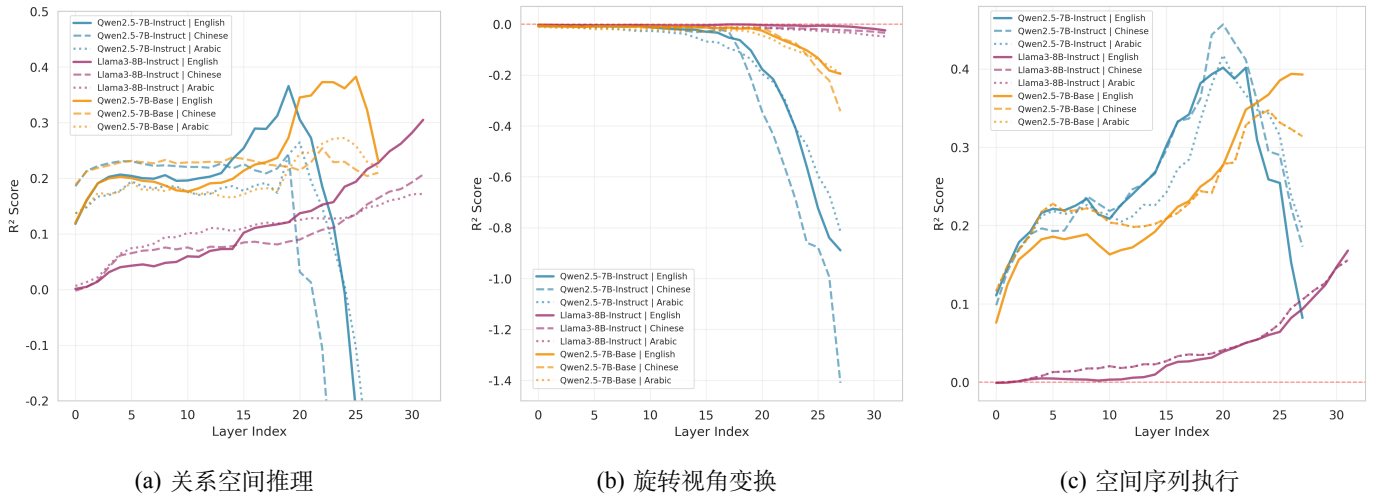


图 4: 三类空间任务族空间变量预测的逐层 R^2 分数

3.3 SAE 空间选择性特征分析

为从分布式隐藏状态中分离可解释的空间特征，本文在各任务族探测性能的峰值层上训练稀疏自编码器 (sparse autoencoders, SAEs)。SAE 采用标准架构，扩展因子为 32，L1 惩罚系数经网格搜索在 $\{10^{-4}, 5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}\}$ 范围内选取（以验证集重构误差与稀疏性的 Pareto 最优为准），使用 Adam 训练 100 个 epoch，批大小为 512，学习率为 10^{-3} 。空间选择性特征定义为：在空间任务输入上的平均激活值超过非空间控制输入平均激活值 2 个标准差以上。SAE 揭示出少量空间选择性特征，通常仅占有所有发现特征的 3%–5%，如图 5 所示，说明空间信息集中于少量可识别但高度任务化的特征子集。

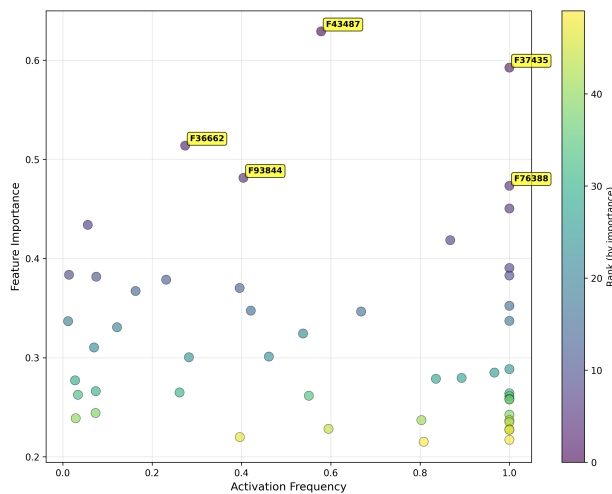


图 5: 空间推理任务中特征重要性与激活频率的关系（以视角旋转变换任务、Qwen2.5-7B-Instruct、汉语为例）

跨任务分析显示，模型内部存在清晰的空间选择性特征，包括响应基本方位的单元（任务族 2）、坐标范围特征（任务族 3）和关系轴特征（任务族 1）。基于梯度的归因分析进一步表明，空间选择性更强的特征对空间预测贡献更大，说明它们并非探测器偶然利用的统计痕迹，而是功能性候选单元。

不同任务族之间的特征重叠有限（12%–18%），说明不同空间计算大体依赖不同特征集合。跨语言分析显示，若依据激活频率衡量，特征存在部分共享；但若依据因果归因排序，重叠显著降低。这提示模型可能在不同语言中到达相似答案，却调用不同内部路径。

3.4 空间表征的因果干预

仅凭探测和特征选择不足以证明空间表征参与了行为生成。为此，本文采用激活修补（activation patching）和 SAE 特征消融，评估所发现的空间表征是否因果性地影响模型行为。

在激活修补实验中，本文为每个测试样本选取一个仅在目标空间变量上不同的配对样本，并将其在对应层的残差流激活作为反事实激活。将原始前向传播中的目标层激活替换为反事实激活后，测量输出概率分布朝反事实方向偏移的程度。在空间序列执行任务中，替换中间层激活会系统性地输出推向反事实轨迹，其中第 14–18 层效果最强；最终层虽最接近输出生成，干预影响却最弱（图 6）。这与探测结果相互印证：空间信息主要在中间层被构造和使用，但未稳定保留为最终层的显式决策变量。

SAE 特征消融进一步确认了功能特异性。移除排名最高的空间特征使视角变换准确率下降 29%（配对 t 检验， $p < 0.01$ ），使空间序列执行准确率下降 14.5%（ $p < 0.05$ ）；消融非空间控制特征几乎不产生影响（ $\Delta < 2\%$ ， $p > 0.1$ ）。显著性检验基于 200 个测试样本的 bootstrap 重采样（10,000 次），报告 95% 置信区间。这表明本文识别的空间表征不仅可解码，而且在特定任务和层级中具有因果作用；干预效果集中于中间层也说明空间信息未稳定贯通至最终输出层。

4 分析与讨论

本节基于行为评估、线性探测、SAE 特征分析和因果干预结果，讨论 LLMs 空间推理机制。下文依次分析空间表征的条件性与碎片化、跨语言机制差异，以及对空间推理评测的启示。

4.1 空间表征的条件性与碎片化

本文的核心问题是，大语言模型是否形成了结构化空间表征，还是其空间行为仅由浅层语言启发式产生。实验结果并不支持简单二分：LLMs 的确编码了部分空间信息，其中一部分还会因果性地影响行为；但这些表征是条件性的、碎片化的，且未被稳定整合到最终决策之中。

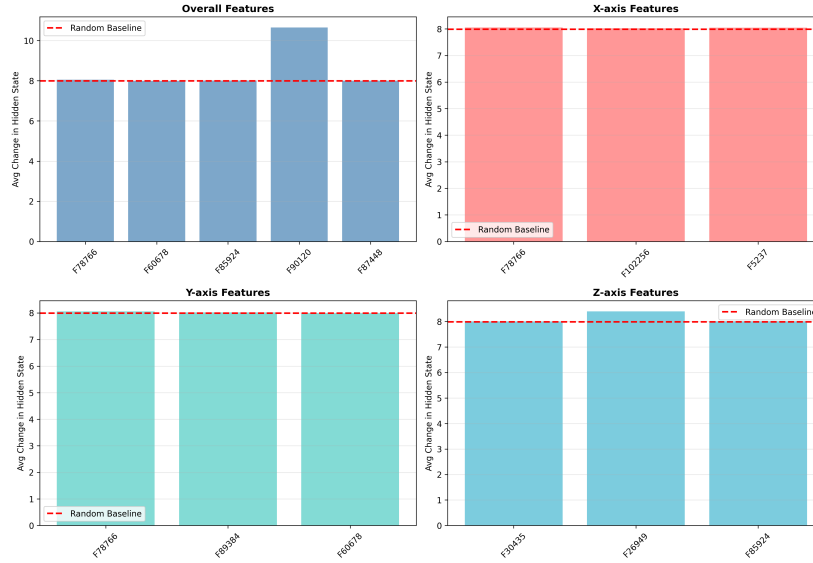


图 6: 不同维度空间特征的干预效应 (以空间序列执行任务, Qwen2.5-7B-Instruct, 英语为例)

线性探测显示, 任务相关空间变量能够从隐藏状态中解码。Qwen 系列在关系推理和空间序列执行上的最高 R^2 可达 0.37–0.40, 表征具备一定几何一致性; Llama3-8B-Instruct 的空间编码整体较弱, 尤其空间序列执行的 R^2 仅为 0.168 (约为 Qwen2.5-7B-Instruct 的 40%)。因果干预进一步证实, 对中间层进行激活修补能够可靠地将模型输出推向反事实空间轨迹。

与此同时, 空间表征是脆弱且瞬时的。Qwen 系列的探测性能在中间层达到峰值后于最终层急剧下降, 说明中间阶段构建的空间结构未持续保留至输出阶段; Llama3-8B-Instruct 的空间信息虽留存至最终层附近, 但强度较低, 同样未能有效支撑空间推理。此外, 空间特征具有明显任务特异性, 在关系推理、朝向推理和坐标推理之间重叠有限。相比通常能以共享表征支持多种空间计算的认知地图, 当前 LLMs 尚未表现出这种统一性。

综合来看, LLMs 展现出条件性且部分的空间编码: 模型能够构建局部空间表征, 但这种表征既不持久, 也未在不同任务族之间形成统一。这解释了其为何能在部分空间基准上取得非平凡表现, 却仍在参考系变换和组合泛化中表现脆弱。

基础模型与指令微调模型的对比进一步揭示了编码与可用性的分离。Qwen2.5-7B-Base 在行为层面几乎无法完成空间任务 (三族任务准确率均低于 4%), 但内部表征中包含较强空间信息: 关系推理 R^2 为 0.382, 超过 Qwen2.5-7B-Instruct 的 0.366 和 Llama3-8B-Instruct 的 0.305; 空间序列执行 R^2 为 0.394, 与表现最好的 Qwen2.5-7B-Instruct (0.402) 几乎持平 (表 III)。这一现象可能说明空间编码与任务导向可用性可分离, 也可能反映基础模型缺乏指令遵循和多选题答题能力。为区分二者, 本文在补充实验中选取基础模型可可靠完成的非空间事实检索任务, 发现其探测 R^2 与行为准确率呈正相关 ($r = 0.64$), 而空间任务上无此相关性 ($r = 0.03$)。这表明空间任务中“编码强而行为弱”的分离并非基础模型的普适特征, 而是空间推理的特有问题。与此同时, Qwen2.5-7B-Base 的最佳探测层位于第 22–26 层, 而非最后若干层, 进一步说明空间信息不会自然流向最终输出层。

不同任务族中的失败模式进一步揭示了 LLM 空间推理的能力边界。**关系空间推理: 组合坍塌。**在关系任务中, 准确率随链长非单调下降, 并在中等复杂度处急剧下滑。错误分析显示, 模型常保留一个空间轴而丢失另一空间轴, 表明多约束组合部分失效; 更高复杂度下的性能恢复可能反映模型转向粗粒度启发式。**视角旋转变换: 表征缺失。**朝向跟踪接近随机水平, 且在所有层上呈现近零的探测 R^2 。错误模式显示模型偏向高频方向 (如北、东), 说明其选择主要依赖语言先验, 而非几何状态跟踪。**空间序列执行: 算术通道主导。**空间序列执行表现最强 (Llama3-8B-Instruct 在英语中达 62%, Qwen2.5-7B-Instruct 在汉语中达 58%), 但错误分析显示不少失败源于算术错误。排除算术错误后, 其空间性能更接近关系推理水平, 说明坐标任务的成功部分受益于数值计算通道。

模型	语言	关系空间推理	视角旋转变换	空间序列执行
Llama3-8B-Instruct	EN	$R^2=.305 / L=31 / \text{MAE}=.57 / \text{RMSE}=.75$	$R^2=-.001 / L=17$	$R^2=.168 / L=31 / \text{MAE}=2.14 / \text{RMSE}=3.27$
	ZH	$R^2=.207 / L=31 / \text{MAE}=.60 / \text{RMSE}=.80$	$R^2=-.003 / L=0$	$R^2=.156 / L=31 / \text{MAE}=2.19 / \text{RMSE}=3.27$
	AR	$R^2=.172 / L=31 / \text{MAE}=.63 / \text{RMSE}=.82$	$R^2=-.002 / L=4$	$R^2=.156 / L=31 / \text{MAE}=2.19 / \text{RMSE}=3.27$
Qwen2.5-7B-Instruct	EN	$R^2=.366 / L=19 / \text{MAE}=.55 / \text{RMSE}=.71$	$R^2=-.008 / L=9$	$R^2=.402 / L=20 / \text{MAE}=1.86 / \text{RMSE}=2.80$
	ZH	$R^2=.243 / L=19 / \text{MAE}=.60 / \text{RMSE}=.78$	$R^2=-.006 / L=0$	$R^2=.457 / L=20 / \text{MAE}=1.65 / \text{RMSE}=2.67$
	AR	$R^2=.265 / L=20 / \text{MAE}=.59 / \text{RMSE}=.76$	$R^2=-.005 / L=0$	$R^2=.418 / L=20 / \text{MAE}=1.82 / \text{RMSE}=2.74$
Qwen2.5-7B-Base	EN	$R^2=.382 / L=25 / \text{MAE}=.55 / \text{RMSE}=.70$	$R^2=-.007 / L=0$	$R^2=.394 / L=26 / \text{MAE}=1.83 / \text{RMSE}=2.79$
	ZH	$R^2=.254 / L=22 / \text{MAE}=.59 / \text{RMSE}=.77$	$R^2=-.003 / L=0$	$R^2=.347 / L=24 / \text{MAE}=1.93 / \text{RMSE}=2.90$
	AR	$R^2=.272 / L=24 / \text{MAE}=.59 / \text{RMSE}=.76$	$R^2=-.011 / L=0$	$R^2=.347 / L=24 / \text{MAE}=1.93 / \text{RMSE}=2.90$

表 III: 按模型、语言和任务族统计的最佳层探测结果

4.2 空间表征的跨语言差异

多语言评估揭示出分层的语言依赖模式。在行为层面，不同语言之间的性能差异最高可达 14 个百分点，尤其体现在关系推理任务中；坐标任务（空间序列执行）的跨语言差异最小，与数值格式更具语言不变性的直觉一致。在表征层面，空间信息在不同语言中出现在相似层级，说明模型具备某种共享计算架构。然而，SAE 分析显示，不同语言中具有因果重要性的特征重叠有限，即不同语言可能达到相近探测表现，却依赖不同内部特征。这挑战了“LLM 空间推理天然语言无关”的观点。

4.3 对空间推理评测的启示

空间表征是必要但非充分的。表征强度与行为准确率之间的差距表明，仅编码空间变量不足以保证可靠推理。主要瓶颈在于空间信息能否向最终决策层传播、整合并转化为答案。因此，未来评测不应只问模型内部是否含有某个变量，还应追问该变量是否被正确路由到行为生成环节。**隐式推理与显式推理。**禁用链式思维提示可揭示模型的隐式空间结构，但这种结构较为脆弱。本文推测，显式推理通过外化中间状态来稳定空间表征，其作用更类似补偿性支架，而非仅仅暴露隐藏能力。**任务形式至关重要。**不同任务族的表现差异凸显了任务形式的影响。基于坐标的任务可受益于算术计算路径，而关系任务和朝向任务要求模型维持抽象空间组合，这是 LLMs 难以持续完成的。某一任务形式上的高表现并不等同于一般空间理解能力；空间评测需要同时覆盖关系整合、参考系变换和状态更新等互补维度。

5 结论与未来工作

本文从机制分析视角考察了大语言模型的空间推理能力。通过将空间推理分解为关系空间推理、视角旋转变换和空间序列执行三类计算基元，并结合线性探测、稀疏自编码器分析与因果干预，本文发现：当前 7-8B 参数规模的 LLMs 并非完全依赖浅层语言启发式，其隐藏状态中确实包含可解码、且在一定条件下影响行为的空间信息。然而，这类空间表征仍具有明显局限。空间信息多集中于中间层，并在传向最终决策层时衰减；基础模型可表现出“编码强而行为弱”的脱耦现象；不同任务族依赖的内部特征重叠有限；跨语言结果进一步显示，相似行为可能由不同内部路径支撑。总体而言，当前 LLMs 具备有限的、条件性的空间编码能力，但尚未形成稳健、统一、可迁移的空间认知机制。

未来工作可在更大规模和更多架构的模型上验证本文发现，进一步分析空间表征的尺度规律与跨任务共享机制；同时，可将该框架扩展至多模态和具身场景，并探索通过辅助空间预测、对比学习、表征正则化或显式状态记忆等方法增强模型空间表征的稳定性、可组合性与跨语言一致性。

致谢

本研究受国家语委科研项目 (ZDA145-17)、教育部人文社会科学研究一般项目 (23YJCZH264)、北京语言大学中央高校基本科研业务费专项资金 (26RGZN19)、北京语言大学研究生创新基金 (中央高校基本科研业务费专项资金) 项目 (25YCX134) 资助。感谢评审专家对本文的细致审阅和宝贵建议。

参考文献

- [1] J. Wang, Z. Wu, Y. Li, H. Jiang, P. Shu, E. Shi, H. Hu, C.-Y. Ma, Y.-H. Liu, X. Wang, Y. Yao, X. Liu, H. Zhao, Z. Liu, H. Dai, L. Zhao, B. Ge, X. Li, T. Liu, and S. Zhang, “Large language models for robotics: Opportunities, challenges, and perspectives,” *ArXiv*, vol. abs/2401.04334, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:266899905>
- [2] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, K. Choromanski, T. Ding, D. Driess, K. A. Dubey, C. Finn, P. R. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. J. Joshi, R. C. Julian, D. Kalashnikov, Y. Kuang, I. Leal, S. Levine, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. S. Ryoo, G. Salazar, P. R. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. H. Vuong, A. Wahid, S. Welker, P. Wohlhart, T. Xiao, T. Yu, and B. Zitkovich, “Rt-2: Vision-language-action models transfer web knowledge to robotic control,” *ArXiv*, vol. abs/2307.15818, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:260293142>
- [3] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. H. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. R. Florence, “Palm-e: An embodied multimodal language model,” in *International Conference on Machine Learning*, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:257364842>
- [4] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. Chawla, O. Wiest, and X. Zhang, “Large language model based multi-agents: A survey of progress and challenges,” in *International Joint Conference on Artificial Intelligence*, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:267412980>
- [5] F. Yi, W. Fu, and H. Liang, “Model-based reinforcement learning: A survey,” 2018. [Online]. Available: <https://api.semanticscholar.org/CorpusID:108339287>
- [6] Y. Yamada, Y. Bao, A. K. Lampinen, J. Kasai, and I. Yildirim, “Evaluating spatial understanding of large language models,” *ArXiv*, vol. abs/2310.14540, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:264426397>
- [7] M. Du, B. Wu, Z. Li, X. Huang, and Z. Wei, “Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models,” in *Annual Meeting of the Association for Computational Linguistics*, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:270371420>
- [8] F. Li, D. C. Hogg, and A. G. Cohn, “Reframing spatial reasoning evaluation in language models: A real-world simulation benchmark for qualitative reasoning,” *ArXiv*, vol. abs/2405.15064, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:270045529>
- [9] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, “Finetuned language models are zero-shot learners,” *ArXiv*, vol. abs/2109.01652, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:237416585>
- [10] T. W. Holliday, J. R. Gautney, and L. Friedl, “Right for the wrong reasons,” *Current Anthropology*, vol. 55, pp. 696 – 724, 2014. [Online]. Available: <https://api.semanticscholar.org/CorpusID:116546048>
- [11] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. S. Zemel, W. Brendel, M. Bethge, and F. Wichmann, “Shortcut learning in deep neural networks,” *Nature Machine Intelligence*, vol. 2, pp. 665 – 673, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:215786368>
- [12] C. Xie, Y. Huang, C. Zhang, D. Yu, X. Chen, B. Y. Lin, B. Li, B. Ghazi, and R. Kumar, “On memorization of large language models in logical reasoning,” *ArXiv*, vol. abs/2410.23123, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:273695832>
- [13] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, F. Xia, Q. Le, and D. Zhou, “Chain of thought prompting elicits reasoning in large language models,” *ArXiv*, vol. abs/2201.11903, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:246411621>
- [14] J. Hewitt and C. D. Manning, “A structural probe for finding syntax in word representations,” in *North American Chapter of the Association for Computational Linguistics*, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:106402715>
- [15] Y. Elazar, S. Ravfogel, A. Jacovi, and Y. Goldberg, “Amnesic probing: Behavioral explanation with amnesic counterfactuals,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 160–175, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:227408471>
- [16] D. Yamins and J. J. DiCarlo, “Using goal-driven deep learning models to understand sensory cortex,” *Nature Neuroscience*, vol. 19, pp. 356–365, 2016. [Online]. Available: <https://api.semanticscholar.org/CorpusID:16970545>
- [17] C. Firestone, “Performance vs. competence in human-machine comparisons,” *Proceedings of the National Academy of Sciences*, vol. 117, pp. 26 562 – 26 571, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:222352613>
- [18] R. B. Ekstrom, J. W. French, and H. H. Harman, “Manual for kit of factor-referenced cognitive tests,” 1976. [Online]. Available: <https://api.semanticscholar.org/CorpusID:141329865>
- [19] R. N. Shepard and J. Metzler, “Mental rotation of three-dimensional objects,” *Science*, vol. 171, pp. 701 – 703, 1971. [Online]. Available: <https://api.semanticscholar.org/CorpusID:16357397>
- [20] M. J. Eckardt, “The hippocampus as a cognitive map,” *Journal of Nervous and Mental Disease*, vol. 168, pp. 191–192, 1980. [Online]. Available: <https://api.semanticscholar.org/CorpusID:144502478>
- [21] N. Burgess, “Spatial cognition and the brain,” *Annals of the New York Academy of Sciences*, vol. 1124, no. 1, pp. 77–97, 2008.
- [22] E. C. Tolman, “Cognitive maps in rats and men,” *Psychological review*, vol. 55 4, pp. 189–208, 1948. [Online]. Available: <https://api.semanticscholar.org/CorpusID:42496633>

- [23] A. S. Etienne and K. J. Jeffery, "Path integration in mammals," *Hippocampus*, vol. 14, 2004. [Online]. Available: <https://api.semanticscholar.org/CorpusID:1646974>
- [24] M. Suzgun, N. Scales, N. Scharli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, and J. Wei, "Challenging big-bench tasks and whether chain-of-thought can solve them," in *Annual Meeting of the Association for Computational Linguistics*, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:252917648>
- [25] M. I. H. Rizvi, X. Zhu, and I. Gurevych, "Sparc and sparp: Spatial reasoning characterization and path generation for understanding spatial reasoning capability of large language models," in *Annual Meeting of the Association for Computational Linguistics*, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:270357373>
- [26] Z. Shi, Q. Zhang, and A. Lipani, "Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts," in *AAAI Conference on Artificial Intelligence*, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:248228026>
- [27] J. Wang, Y. Ming, Z. Shi, V. Vineet, X. Wang, and N. Joshi, "Is a picture worth a thousand words? delving into spatial reasoning for vision language models," *ArXiv*, vol. abs/2406.14852, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:270688598>
- [28] W. Wu, S. Mao, Y. Zhang, Y. Xia, L. Dong, L. Cui, and F. Wei, "Mind's eye of llms: Visualization-of-thought elicits spatial reasoning in large language models," *Advances in Neural Information Processing Systems* 37, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:268889526>
- [29] Y.-H. Liao, R. Mahmood, S. Fidler, and D. Acuna, "Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models," in *Conference on Empirical Methods in Natural Language Processing*, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:272689771>
- [30] M. Sharma, "Exploring and improving the spatial reasoning abilities of large language models," *ArXiv*, vol. abs/2312.01054, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:265609823>
- [31] Z. Li, Y. Cao, X. Xu, J. Jiang, X. Liu, Y. S. Teo, S.-W. Lin, and Y. Liu, "Llms for relational reasoning: How far are we?" in *Proceedings of the 1st International Workshop on Large Language Models for Code*, 2024, pp. 119–126.
- [32] W. Liu, Q. Xue, H. Wang, X. Yin, B. Yang, and W. Gao, "Spatial reasoning in multimodal large language models: A survey of tasks, benchmarks and methods," 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:283109806>
- [33] X. Zheng, Z. Dongfang, L. Jiang, B. Zheng, Y. Guo, Z. Zhang, G. Albanese, R. Yang, M. Ma, Z. Zhang, C. Liao, D. Zhen, Y. Lyu, Y. Fu, B. Ren, L. Zhang, D. P. Paudel, N. Sebe, L. van Gool, and X. Hu, "Multimodal spatial reasoning in the large model era: A survey and benchmarks," *ArXiv*, vol. abs/2510.25760, 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:282575141>
- [34] H. Maennel, I. M. Alabdulmohsin, I. O. Tolstikhin, R. J. N. Baldock, O. Bousquet, S. Gelly, and D. Keysers, "What do neural networks learn when trained with random labels?" *ArXiv*, vol. abs/2006.10455, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:219792883>
- [35] R. Krause and S. Reimann, "Items or relations - what do artificial neural networks learn?" *ArXiv*, vol. abs/2404.12401, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:269282516>
- [36] N. Nanda and J. Bloom, "Transformerlens," <https://github.com/TransformerLensOrg/TransformerLens>, 2022.
- [37] J. Bloom, C. Tigges, A. Duong, and D. Chanin, "Saelens," <https://github.com/decoderresearch/SAELens>, 2024.
- [38] J. Hewitt and P. Liang, "Designing and interpreting probes with control tasks," *ArXiv*, vol. abs/1909.03368, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:202538609>
- [39] S. Gupta, J. Robinson, D. Lim, S. Villar, and S. Jegelka, "Structuring representation geometry with rotationally equivariant contrastive learning," *ArXiv*, vol. abs/2306.13924, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:259252140>
- [40] M. Raghu, J. Gilmer, J. Yosinski, and J. N. Sohl-Dickstein, "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability," in *Neural Information Processing Systems*, 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:23890457>
- [41] H. Yan, Y. Xiang, G. Chen, Y. Wang, L. Gui, and Y. He, "Encourage or inhibit monosemanticity? revisit monosemanticity from a feature decorrelation perspective," in *Conference on Empirical Methods in Natural Language Processing*, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:270737676>
- [42] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and editing factual associations in gpt," in *Neural Information Processing Systems*, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:255825985>
- [43] C. Olsson, N. Elhage, N. Nanda, N. Joseph, N. Dassarma, T. J. Henighan, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, S. Johnston, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. B. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah, "In-context learning and induction heads," *ArXiv*, vol. abs/2209.11895, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:252532078>
- [44] N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen *et al.*, "Toy models of superposition," *arXiv preprint arXiv:2209.10652*, 2022.
- [45] W. Gurnee, N. Nanda, M. Pauly, K. Harvey, D. Troitskii, and D. Bertsimas, "Finding neurons in a haystack: Case studies with sparse probing," *Transactions on Machine Learning Research*, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:256416079>
- [46] M. Hanna, O. Belkina, S. Pezzelle, P. Lertvittayakumjorn, and L. Passaro, "How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model," *Advances in Neural Information Processing Systems* 36, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:264487910>
- [47] S. Marks, C. Rager, E. J. Michaud, Y. Belinkov, D. Bau, and A. Mueller, "Sparse autoencoders find highly interpretable directions in language models," *ArXiv*, vol. abs/2410.02739, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:274822666>
- [48] Q. A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, and K. D. et al., "Qwen2.5 technical report," *ArXiv*, vol. abs/2412.15115, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:274859421>
- [49] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, and et al., "The llama 3 herd of models," 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:271571434>