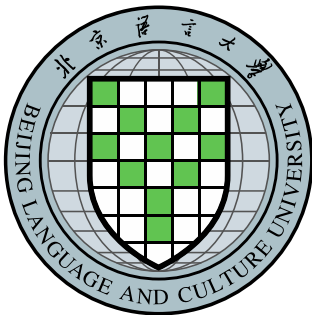




大语言模型空间推理能力的内部表征机制研究

安纪元¹, 杨麟儿¹, 王梦焰¹, 鲁鹿鸣², 安维华¹, 杨尔弘¹

¹北京语言大学, ²北京师范大学

2026年10月17日

目录

1

研究背景

4

结果分析

2

我们的方法

5

总结

3

实验设置

空间智能

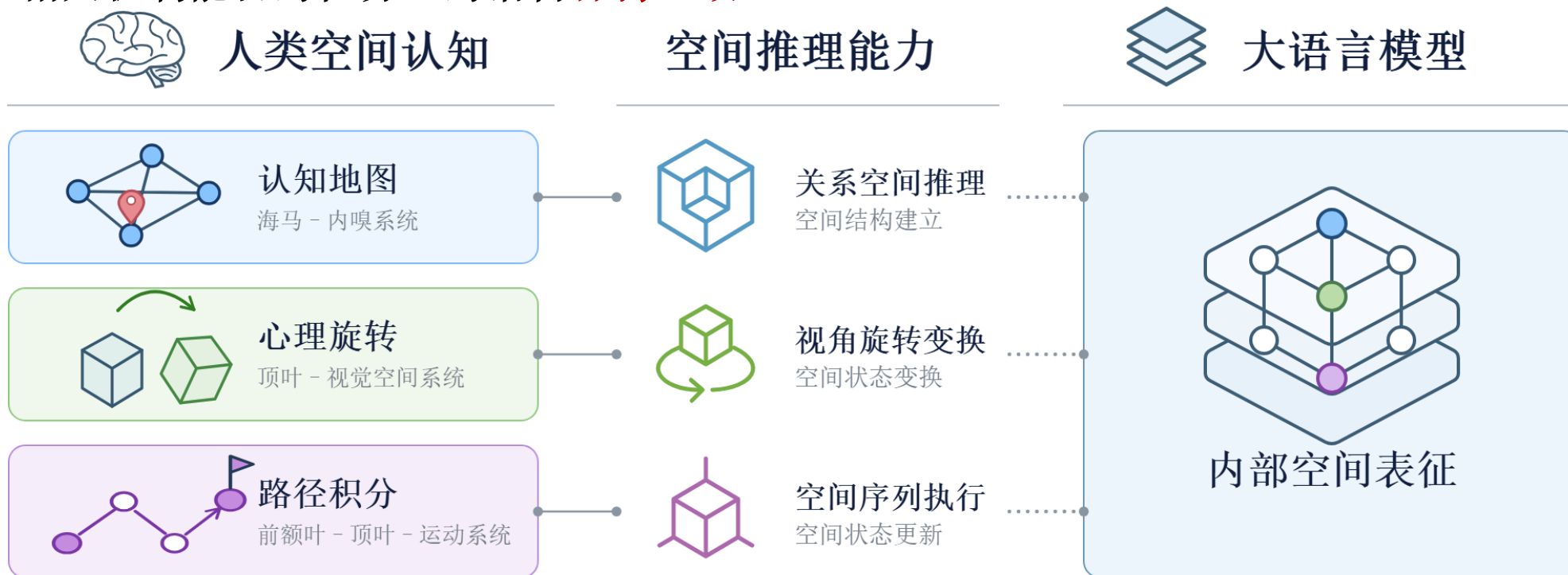
- **空间智能**是理解空间关系、构建环境表征，并据此进行推理与行动的能力。
- 具身智能与世界模型要求模型理解环境结构并据此规划行动^{[1][2]}。
- **研究焦点**：模型是否真正形成并调用空间表征，而不仅是在任务中给出正确答案？

可信空间推理

- **可信空间推理**不仅要求结果正确，还要求推理依据可验证、过程可解释。
- 现有评估主要关注模型能答对多少题，但是难以解释答案产生的**内部依据**。
- 正确答案可能来自空间计算，也可能受到语言启发、选项偏置或任务格式的影响。

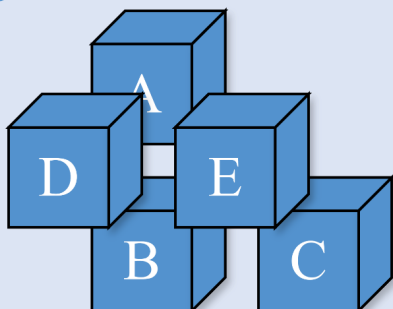
研究背景

- 构建空间推理能力基准，以此研究大语言模型空间推理能力的内部表征机制：
 - 空间信息在模型内部**何处**、以**何种形式**出现？（与人类是否相似？）
 - 这些信息是否对输出决策具有**因果作用**？
 - 相关机制能否跨任务、跨语言**保持一致**？



我们的方法——数据集构建

- 空间推理能力任务定义：



关系空间推理

B is below of A.
C is right of B.
D is front of A.
E is right of D.

Where is A relative to E?

left and behind

B 在 A 的下方。
C 在 B 的右边。
D 在 A 的前面。
E 在 D 的右边。

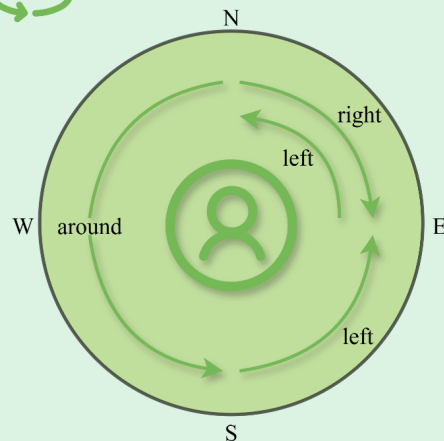
A 相对于 E 的位置是？

左后方

ب أسفل من أ.
ج إلى يمين ب.
د أمام أ.
هـ إلى يمين د.

أين أ بالنسبة إلى هـ؟

الخلف الأيسر



视角旋转变换

You are facing north.
Turn right.
Turn left.
Turn around.
Turn left.

Which direction are you facing?

East

你正面向北方。
向右转。
向左转。
向后转。
向左转。

你现在面向哪个方向？

东方

أنت تتجه نحو الشمال.
استدر يمينًا.
استدر يسارًا.
استدر تمامًا (أو: ارجع للخلف).
استدر يسارًا.

إلى أي اتجاه تتجه الآن؟

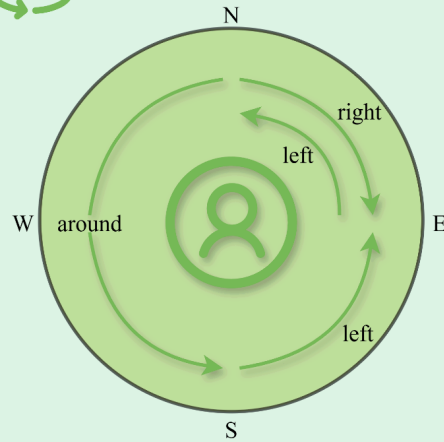
الشرق

我们的方法 — 数据集构建

• 空间推理能力任务定义:



视角旋转变换



You are facing north.
Turn right.
Turn left.
Turn around.
Turn left.

Which direction are you facing?

East

你正面向北方。
向右转。
向左转。
向后转。
向左转。

你现在面向哪个方向?

东方

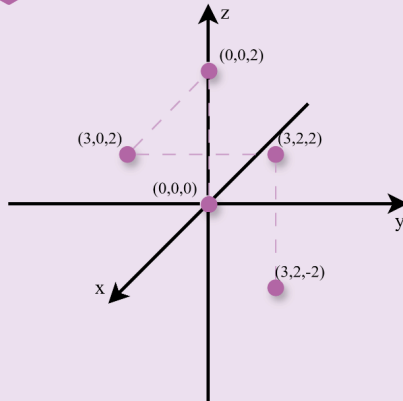
أنت تتجه نحو الشمال.
استدر يميناً.
استدر يساراً.
استدر تمامًا (أو: ارجع للخلف).
استدر يساراً.

إلى أي اتجاه تتجه الآن؟

الشرق



空间序列执行



Start at $(0, 0, 0)$.
Move up by 2 units.
Move forward by 3 units.
Move left by 2 units.
Reflect across the z-axis.

What is the final position?

$(3, 2, -2)$

起点位于 $(0, 0, 0)$.
向上移动 2 个单位。
向前移动 3 个单位。
向左移动 2 个单位。
关于 z 轴对称变换。

最终位置是什么?

$(3, 2, -2)$

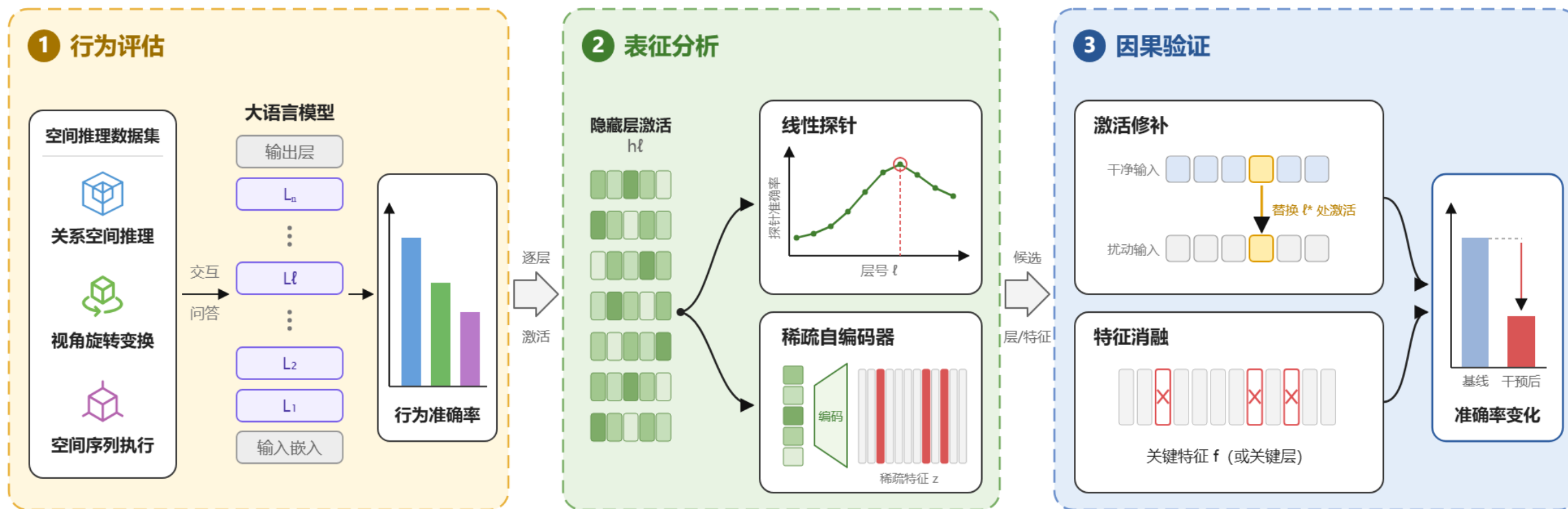
ابدأ من النقطة $(0, 0, 0)$.
تحرك لأعلى بمقدار وحدتين.
تحرك للأمام بمقدار 3 وحدات.
تحرك لليساار بمقدار وحدتين.
قم بعكس الاتجاه عبر محور z.

ما هو الموقع النهائي؟

$(3, 2, -2)$

我们的方法——机制分析框架

- 结合表层行为、内部表征与因果验证



模型回答是否**正确**？

空间变量在哪里**显著**？

干预表征是否**影响**输出结果？

哪些特征有选择性？

- 模型选择

- 探究不同训练阶段/模型架构，在空间推理表征机制是否存在**共性结论**（泛化性）。

比较维度	模型组合	考察目的
不同训练阶段	Qwen2.5-7B- Base Qwen2.5-7B- Instruct	是否因为指令微调变化？
不同模型系列	Qwen2.5 -7B-Instruct Llama3 -8B-Instruct	是否具有模型家族差异？

- 推理设置

- 减少外部推理过程与答案格式对机制分析的**影响**。

推理设置	设置目的
标准单次前向推理（不使用CoT）	减少显式思考过程的影响
避免生成中间解释（Thinking）	聚焦隐藏状态中的空间信息
多项选择，单 token 作答	减少答案措辞与长度差异

- 任务设计与难度控制

- 抽象** 实体与关系：减少世界知识依赖，主要依据题目给出的空间约束作答。

任务	考察目标	难度范围
关系空间推理	整合局部关系，建立整体结构	3–10 跳
视角旋转变换	跟踪朝向与参考系变化	2–10 步
空间序列执行	维护并更新位置状态	2–10 步

- 跨语言任务设置

- 验证空间推理能力与语言依赖是否解耦，是否存在**共享的内部空间表征机制**。
 - 英语（EN）——印欧语系（Indo-European）
 - 汉语（ZH）——汉藏语系（Sino-Tibetan）
 - 阿拉伯语（AR）——亚非语系（Afro-Asiatic）

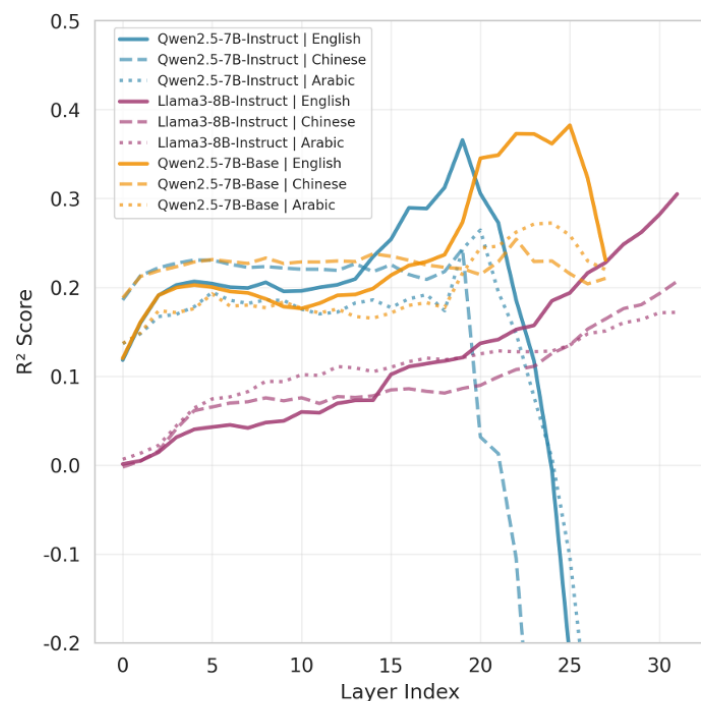
结果分析 — 行为准确率

- 空间推理表现具有明显任务差异

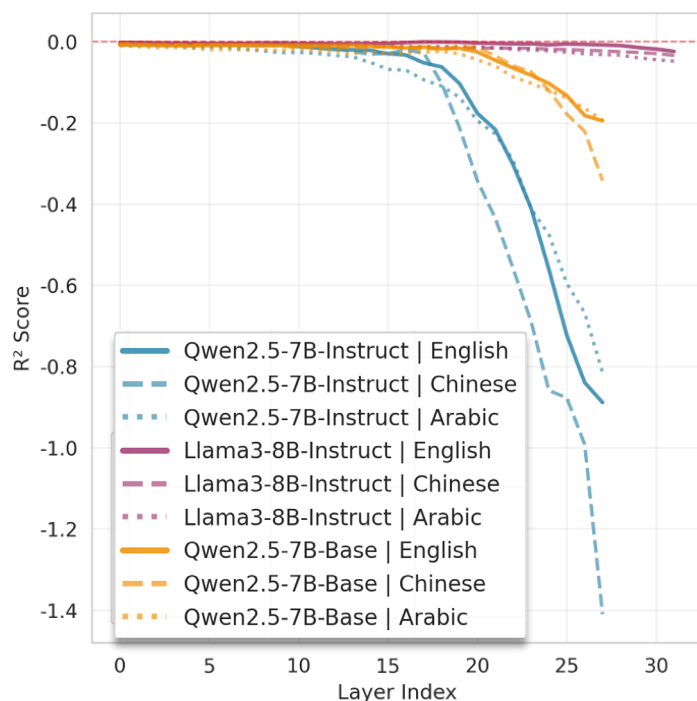
模型	语言	关系空间推理	视角旋转变换任务准确率	空间序列执行任务准确率
Llama3-8B-Instruct	EN ✓	49.0 (98/200)	28.0 (56/200)	62.0 (124.0/200)
	ZH ✓	47.5 (95/200)	28.0 (56/200)	52.5 (105/200)
	AR	31.0 (62/200)	26.0 (52/200)	57.5 (115/200)
Qwen2.5-7B-Instruct	EN	50.0 (100/200)	23.0 (46/200)	55.5 (111/200)
	ZH	39.5 (79/200)	23.5 (47/200)	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)

结果分析 — 线性探针

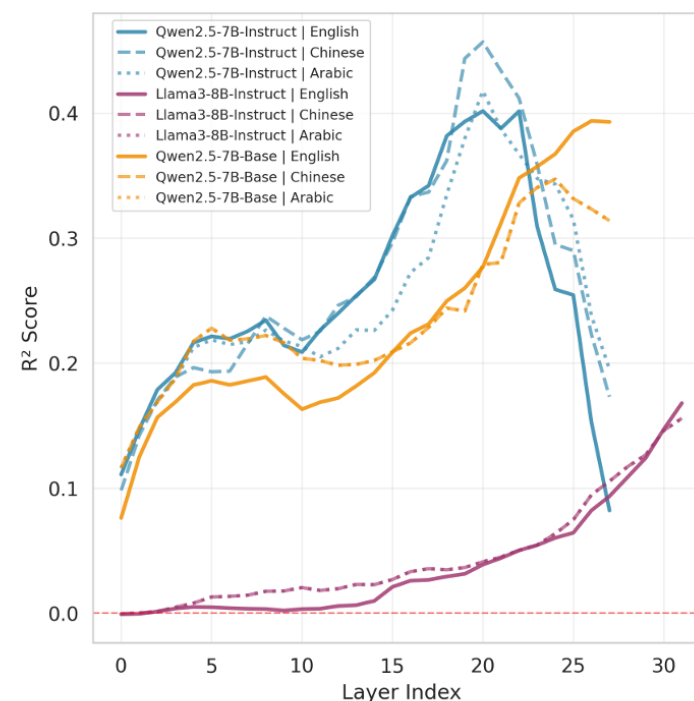
- 空间信息的分布具有**模型差异**
 - 峰值位置与末层变化不同，提示空间信息并非沿统一路径持续保留。



(a) 关系空间推理



(b) 旋转视角变换

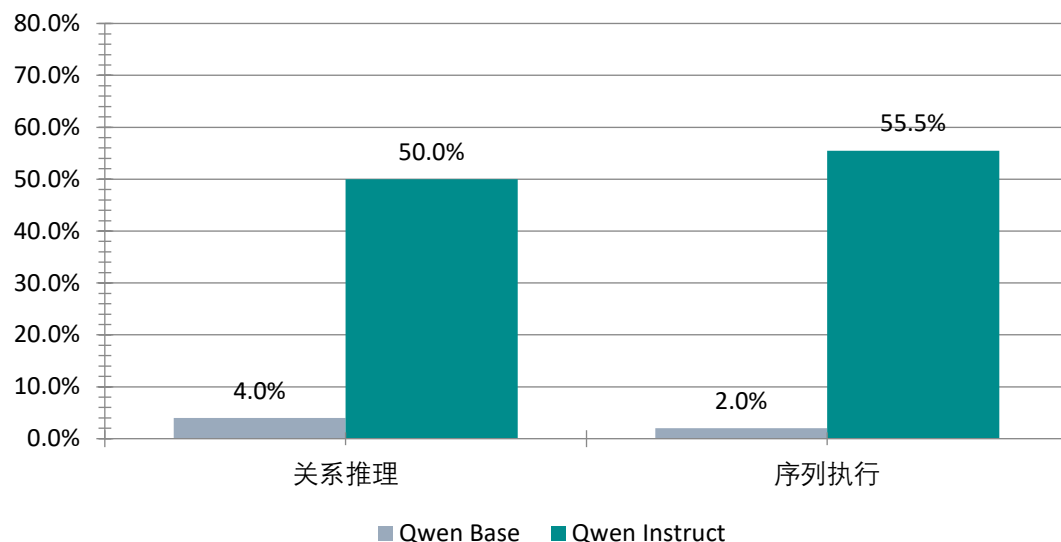


(c) 空间序列执行

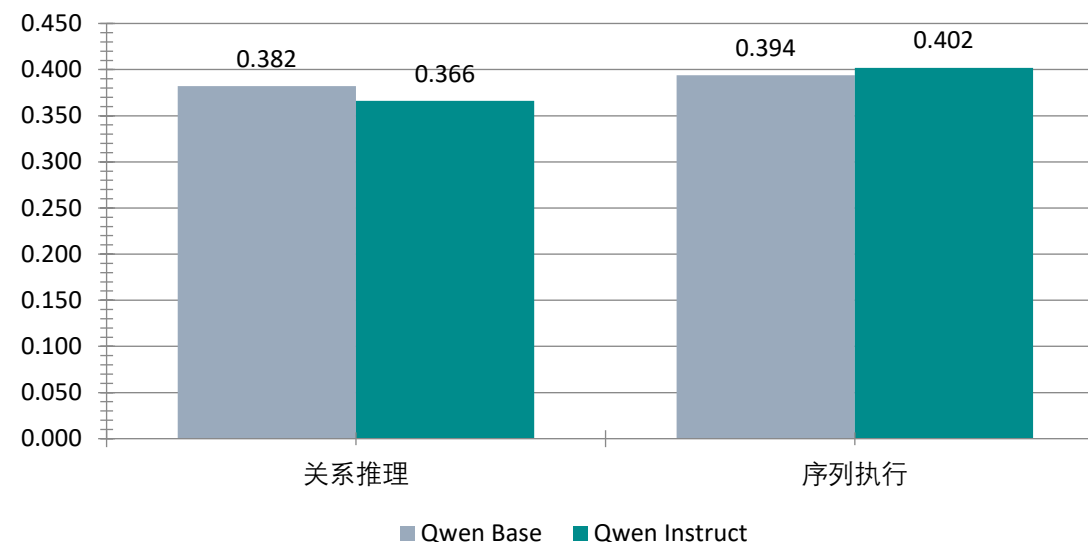
结果分析 — 线性探针

- 空间编码与答题行为存在**脱耦**现象
 - Base 模型已保留较强的可解码信息，但在答题设置下难以有效使用。

行为：推理准确率



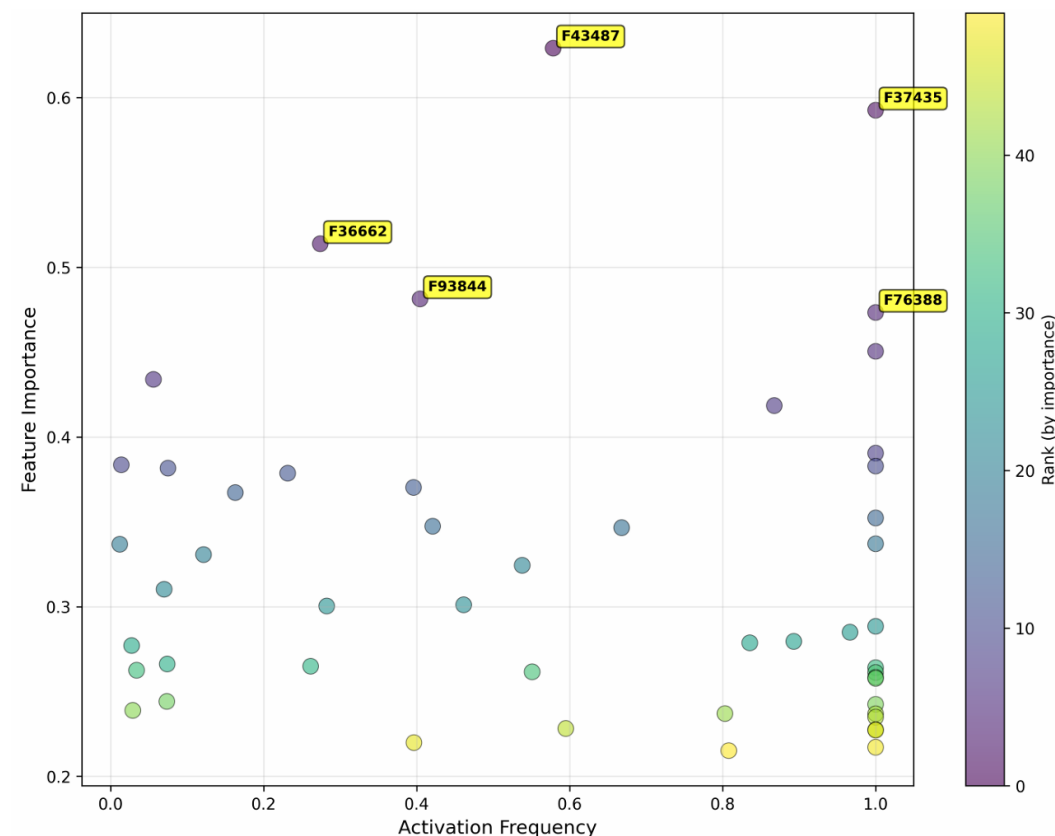
编码：最佳层探针 R^2



结果分析 — 自稀疏激活

- 空间特征稀疏，跨任务共享有限
 - 空间选择性特征占发现特征的比例： **3%–5%**
 - 不同空间任务之间的特征重叠： **12%–18%**

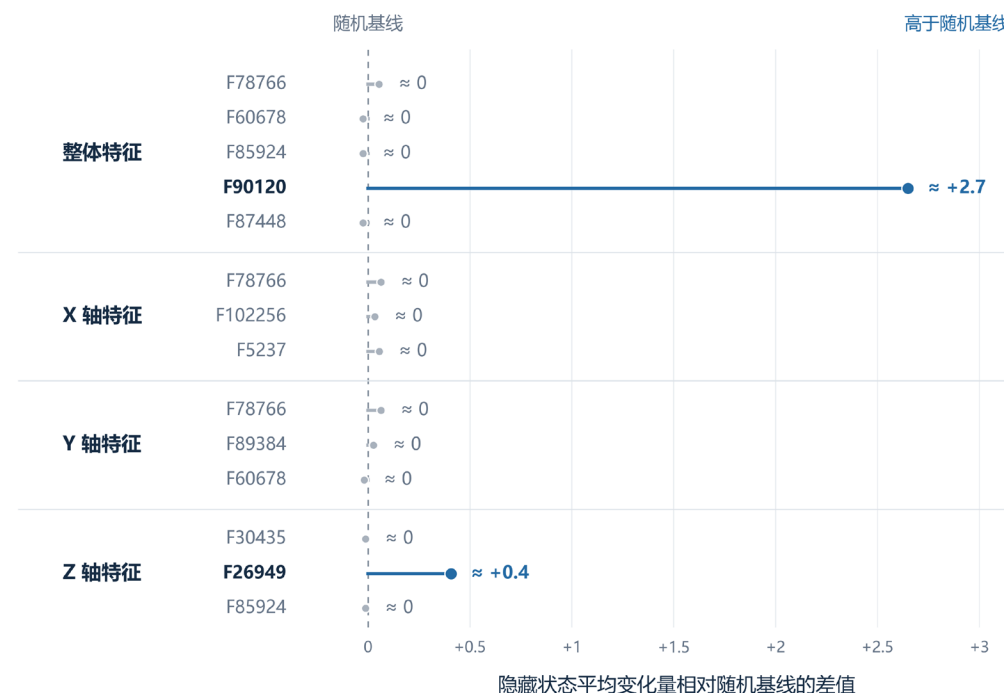
任务	特征
关系空间推理	关系轴特征
视角旋转变换	基本方位响应
空间序列执行	坐标范围特征



结果分析 — 因果验证

- 部分空间特征会显著影响模型输出准确率
- 激活修补
 - 用配对反事实样本的目标层激活，替换原始样本对应层的激活。
 - 中间层（第 14–18 层）影响更强。
- 特征消融

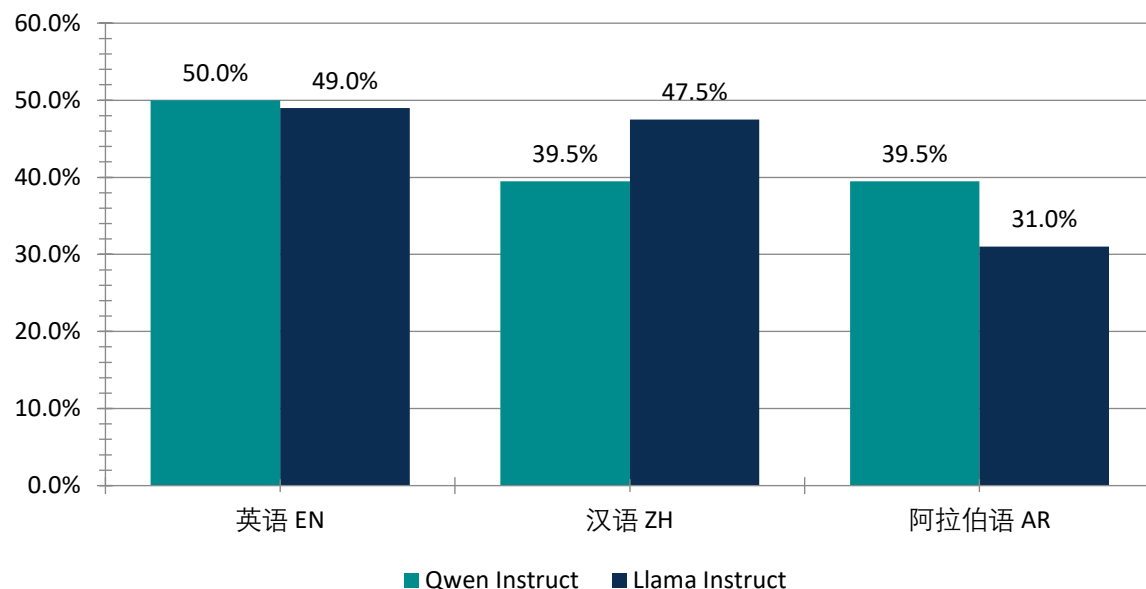
干预对象 / 任务	影响	显著性
空间特征 / 视角变换	准确率下降 29%	$p < 0.01$
空间特征 / 序列执行	准确率下降 14.5%	$p < 0.05$
非空间控制特征	变化 $< 2\%$	$p > 0.1$



结果分析 — 跨语言分析

- 任务表层准确率相近 + 线性探针层级相似 $\neq$ 内部功能特征共享

关系空间推理准确率 (Qwen-2.5-7B-Instruct)



最佳关系推理探针 (Qwen-2.5-7B-Instruct)

语言	最佳层	R ²
EN	19	0.366
ZH	19	0.243
AR	20	0.265

按因果重要性比较时，跨语言特征重叠有限。

总结

- 核心结论：空间信息存在，稳定表征与共享有限。
- 空间信息可被编码
 - 隐藏状态中存在可解码的空间变量，层级分布随模型与任务变化。
- 部分特征具有因果作用
 - 模型内部空间特征占比较小，干预部分特征可显著改变输出。
- 跨任务与跨语言共享有限
 - 模型相近的外部表现下，可能由不同内部推理路径支持。



论文



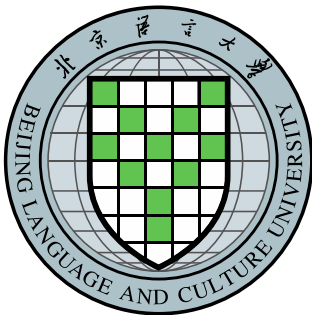
源码(Github)



扫二维码，添加我为朋友。



北京语言大学
语言监测与智能学习研究组



谢谢