基于知识库的空间关系推理数据集自动生成研究 *

詹卫东[1,2,3] 胡 楠[2] 肖力铭[2] 孙春晖[2]

（1. 北京大学中国语言学研究中心 北京 100871；2. 北京大学中文系 北京 100871；
3. 多媒体信息处理全国重点实验室 北京 100871）

［**摘要**］空间关系推理任务旨在考察大语言模型对空间关系的理解以及空间常识推理能力。本文提出了一种空间关系推理数据集自动生成方法，并基于此方法生成了 2040 道中文空间推理测试题，用于面向大语言模型的空间语言理解能力评测大赛 SpaCE2024 中。相比以往的空间推理数据集，本数据集涵盖了更多的空间关系类型、空间图式和评价维度。参赛系统在本文评测数据集上的平均准确率为 0.34，显示基于本文方法生成的空间关系推理测试数据集对大语言模型而言具有高挑战性。

［**关键词**］数据合成；空间常识；推理；大语言模型；语言理解能力测试

［**中图分类号**］TP391.1［**文献标识码**］A［**文章编号**］1003–5397（2025）04–0016–11

DOI:10.16499/j.cnki.1003-5397.2025.04.009

Knowledge-Driven Automatic Generation of Spatial Reasoning Datasets

ZHAN Weidong, HU Nan, XIAO Liming, SUN Chunhui

Abstract: Spatial relation reasoning tasks are designed to evaluate Large Language Models' comprehension of spatial relations and their capacity for spatial commonsense reasoning. In this paper, we propose an automatic generation method for spatial relation reasoning datasets. Based on this method, we generated 2,040 Chinese spatial reasoning test items for SpaCE2024, an evaluation competition on spatial language understanding for LLMs. Compared with previous spatial reasoning datasets, the dataset covers a wider range of spatial relation types, spatial schemas and evaluation dimensions. The average accuracy of participating systems on this dataset was 0.34, indicating that the spatial relation reasoning dataset generated by our method poses a significant challenge to Large Language Models.

［收稿日期］2025-05-30

［作者简介］詹卫东，北京大学中文系教授，博士生导师，主要研究计算语言学、现代汉语语法和语言知识工程；胡楠，北京大学中文系博士生，主要研究计算语言学；肖力铭，北京大学中文系博士生，主要研究计算语言学；孙春晖，北京大学中文系博士生，主要研究计算语言学。

* 本研究得到教育部人文社会科学重点研究基地重大项目“面向机器语言能力评测的综合型语言知识库研究”（22JJD740004）的资助。北京大学中文系王佳骏、邢丹、王希豪、李楠、张子涵、崔香、蔡奇栩、邓思锐、秦宇航等多位同学参与了本文研究工作，在此一并致谢。

Abstract: data synthesis; spatial common sense; reasoning; large language model; language understanding evaluation

一　引言

空间表达描述了物体之间的空间方位关系。要准确理解文本中的空间语义，不仅需要语言知识，还需要调用空间认知能力，构建空间场景，并进行空间方位信息相关的推理。聚焦于训练和评估机器的空间场景建模能力、空间语义理解能力和空间方位推理能力，自然语言处理（NLP）领域的研究者构建了一系列的空间推理数据集，表 1 中列举了空间推理数据集的代表性工作。

表 1　空间推理数据集的代表性工作

数据集名称	测试任务	数据集涉及的空间关系
bAbI（2015）	实体空间方位推理 空间路径发现	上、下、左、右、东、西、南、北（8 类）
StepGame（2022）	基于网格的实体间空间方位关系判断	上、下、左、右、左上、左下、右上、右下（8 类）
SpartQA（2021）& SpaRTUN（2022）	基于图片的实体间空间方位关系判断	上、下、左、右、东、西、南、北、远、近、内、外（12 类）
RoomSpace（2024）	基于 3D 建模的实体间方位关系判断	上、下、左、右、东、西、南、北、远、近、前、后、左前、右前、左后、右后等（20 类）

然而，现有的空间推理数据集仍然存在着较多的不足之处。比如，bAbI（Weston et al.，2015）数据集中的子任务 Task17“实体空间方位推理”（Positional Reasoning）、Task19“实体路径发现”（Path Finding）与空间推理任务相关，但其测试形式和内容都较为简单，推理步数较少，并仅涉及 8 类基本空间关系。StepGame（Shi et al.，2022）在一定程度上扩展了空间关系的类型，并且将推理步数的范围提高到了 4 到 10 步，大大提高了推理难度。不过，StepGame 基于网格的空间关系推理和现实世界的空间推理仍有较大不同，为了保证答案的确定性，StepGame 将实体置于网格中，对各个空间关系做了严格的单位和角度的限制，例如，“左边”在 StepGame 中被定义为“一个实体的横坐标比另一个实体的横坐标小 1 个单位”，与现实世界差异较大。

SpartQA（Mirzaee et al.，2021）和 SpaRTUN（Mirzaee et al.，2022）在出题形式上不局限于网格中抽象的空间方位关系，而是基于 NLVR 数据集（Suhr et al.，2017）中的简单几何图形来编写物体间方位关系的脚本描述，在提高推理难度的同时，贴近现实世界的空间推理，涉及的空间关系类型拓展至 12 类。RoomSpace（Li et al.，2024）数据集采用了 3D 建模方法来模拟日常生活场景中的空间实体图式，并将实体间复杂的空间关系视为约束满足问题（Constraint Satisfaction Problem，CSP）来构建空间推理题自动生成算法，所涉及的空间关系进一步提升至 20 类。

尽管上述数据集在空间关系类型、推理复杂度和现实贴合度方面不断取得进展，但在模型评估方面仍显不足。在评估过程中，目前的空间推理评测往往只给出模型的总体正确率，无法细粒度地评估空间推理能力，难以为模型优化提供针对性指导。评测数据的价值不仅在于筛选模型优劣，更在于帮助开发者诊断问题、定位短板、指导改进。

为实现更具解释性的空间推理能力评估，评测用数据集的自动生成方式，需要在已有的依靠模板生成的基础上优化。本文提出以结构化知识库为支撑的数据集自动生成框架，具体包括三个核心模块：（1）建立基于常识的空间方位关系推理知识库，结构化地描述空间图式及其约束规则；（2）设计基于知识库的自然语言试题自动生成算法，实现从知识库到题库的自动转换；（3）根据题目的知识标签，构建多维度的评估指标，实现对大语言模型空间推理能力的动态细粒度评估。

二　空间方位关系推理知识库的构建

空间方位关系推理知识库是实现空间推理测试题自动生成的基础。知识库中记录了特定空间图式中各个实体的空间关系信息。空间图式、空间方位模板、模板间推理关系及初始事实共同构成了空间关系推理知识库。

（一）空间图式

一个空间实体并不是孤立的，一般总要与其他实体产生一定的空间关系，它们共同构成一个空间图式。空间图式是对现实世界中无数具体的空间场景的抽象，如：桌上放着一本书、树上停着一只鸟、一艘渡轮正在渡河、一个人正横穿马路，等等，这些场景经抽象后形成的构型称为“空间图式”（Talmy，2005）。

将具体空间场景抽象为空间图式有利于固定空间实体的分布和实体的数量，从而集中考察特定类型的空间关系。同时，不同的空间图式涉及的空间关系有同有异，设计不同类型的空间图式时，主要考虑因素是提高各种空间图式对空间关系的整体覆盖率。

本研究尝试构建了四种空间图式，均由生活中的空间场景抽象而来，如图 1 所示，分别为六人向心面对围坐的“向心六角”（图 1- 甲）、六人离心背对围坐的“离心六角”图式（图 1- 乙）、六个实体在 3×2 置物架上放置的“三层两列”图式（图 1- 丙）和四人两两面对面的“四人卡座”图式（图 1- 丁）。

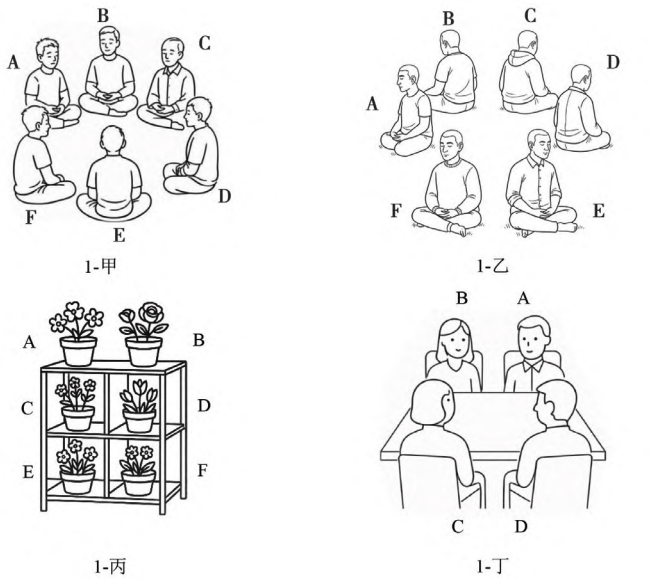
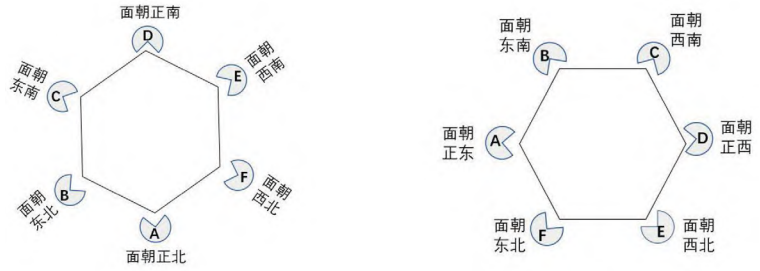


图 1　四种空间图式

在四种空间图式中引入更多空间关系，可以进一步扩展出更具体的空间场景，例如：引入了东、南、西、北等空间关系，为场景中的实体规定了特定的朝向或所处方位，就得到了“向心六角正南正北”（图 2- 甲）、“向心六角正东正西”（图 2- 乙）等容纳更多空间关系的更为具体的空间场景。空间场景的名称及空间图式对应的场景数量见表 2。

（二）空间方位模板

对于一个空间图式，可以采用模板加推理规则来描述其中实体的方位信息。



2-甲　　2-乙

图 2　向心六角空间图式的两种空间布局（点位图式）

表 2　空间图式及其具体的空间场景

空间图式	空间场景	场景数量
四人卡座	四人卡座（无朝向）；四人卡座正东正西；四人卡座正南正北	3
向心六角	向心六角（无朝向）；向心六角正东正西；向心六角正南正北	3
离心六角	离心六角（无朝向）；离心六角正东正西；离心六角正南正北	3
三层两列	三层两列正东正西	1

空间方位模板描述了一个空间图式中实体的位置信息或实体间的方位关系信息。表 3 列举了四个空间图式中空间方位模板的数量，并给出了若干示例。模板“X 在东侧”描述了一个空间实体 X 的位置信息，“X 在 Y 的正上方”描述了两个实体 X 和 Y 的相对方位关系。

表 3　四个空间图式中空间方位模板的相关信息及实例

空间图式	模板数量	模板示例
向心六角	156	X 和 Y 面对面、X 在 Y 的顺时针方向第一位、X 在 Y 的斜对面
离心六角	136	X 和 Y 背对背、X 在 Y 的逆时针方向第一位、X 在 Y 的斜后方
三层两列	172	X 在三层东侧、X 在 Y 的正上方、X 和 Y 在同一层
四人卡座	56	X 在 Y 的左手边、X 在 Y 的斜对面、X 在 Y 的正对面

对于一个模板而言，X 和 Y 可被具体的空间实体替代，以形成描述一个空间方位事实的命题，例如“茉莉在月季的正上方”就是由“X 在 Y 的正上方”派生得到的命题。

（三）模板间推理关系

空间模板之间的推理规则可以分为两种形式：（1）逻辑关系；（2）“条件—结论”推导规则。逻辑关系是两个空间方位模板之间的简单推导规则。表 4 列出了空间方位模板的五种逻辑关系类型，并以三层两列空间图式为例，展示了该图式中部分模板之间的逻辑关系。

除简单逻辑关系外，知识库中还引入了推导规则来描述空间关系之间更复杂的推理。规则由条件和结论两部分组成。条件部分是一个或多个模板，结论部分为一个模板，即单个或多个条件可推出一个结论。表 5 列举了空间方位模板的部分推理规则实例。

（四）初始事实

初始事实是一组空间方位命题，对一个特定空间图式进行全局描述。例如，在四人卡

表 4　模板间逻辑关系及其示例

逻辑关系	逻辑表达式	示例	
		p（模板 1）	q（模板 2）
等价	$p \longleftrightarrow q$	X 的左邻在 Y 的左下方	Y 的左邻在 X 的左上方
		X 在 Y 的右边	Y 在 X 的左边
蕴含	$p \rightarrow q$	X 的左邻在 Y 的左下方	X 在 Y 正下方
		X 在 Y 的右边	X 和 Y 左右相邻
包括	$q \rightarrow p$	X 与 Y 同层	X 在 Y 左边
		X 在 Y 的下方	X 在 Y 的右下方
冲突	$\neg(p \cap q)$	X 和 Y 隔了一层	Y 所在层和 X 所在层相邻
		X 在西侧	X 在东侧
互反	$(\neg p \longleftrightarrow q) \cup (p \longleftrightarrow \neg q)$	X 与 Y 同侧	X 与 Y 不同侧
		X 与 Y 在同一层	X 与 Y 不同层

表 5　空间方位模板的推理规则示例

空间图式	推理规则示例	
	条件	结论
向心六角	① J 在 K 的顺时针方向第一位 ② L 在 J 的顺时针方向第一位	L 在 K 的顺时针方向第二位
四人卡座	① J 在 L 的左手边 ② K 在 L 的正对面	J 在 K 的斜对面

座空间图式中，分别以 A、B、C、D 来指代四个空间实体，有初始事实如下：

命题 1：B 在 A 的右手边；命题 2：C 在 B 的正对面；命题 3：D 在 C 的右手边

以上三条命题定义了四人卡座空间图式中所有实体之间的空间关系（对应上文图 1-丁）。基于初始事实，程序可以自动运用知识库中的逻辑关系和推理规则，从而生成关于该空间图式中所有实体（A、B、C、D）之间的空间关系事实，并依据这些事实生成推理题（参见下文中的“出题模块”）。

三　空间推理题的自动生成

本研究设计的空间推理题自动出题程序（AutoQuest）工作流程如图 3 所示。限于篇幅，本文仅重点介绍程序中的推理模块和出题模块。为构建数据集，程序还需要对试题的数量、分布等进行控制，程序开始运行时，在读入“知识库”的同时，还要读入“试题配置”文件，该文件中包含了这些控制信息。为对大模型的性能表现进行细粒度分析，对自动生成的每道题，还要详细记录题目的特征信息，程序输出“题库”结果的同时，还会输出“试题信息”文件，记录题目的各项详细特征。

推理模块的功能是根据用户配置读取知识库，包括初始事实、空间推理模板、模板间逻辑关系和推理规则，然后迭代应用全部逻辑规则，从初始事实推导出实体间的全部空间关系，生成完整事实库。

出题模块从事实库中筛选能描述空间图式全貌的命题组作为题干备用文本，从该组

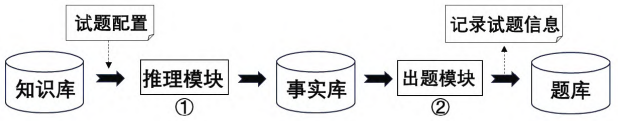


图 3　空间推理题自动出题程序的工作流程

命题中随机选取一个命题（如“A 在 C 的斜对面”），隐去其中的一个实体（如“C”），这样，该命题就变成了一个问题（如：A 在 ___ 的斜对面），隐去的实体（如“C”）就是问题的答案。因为题型是选择题而非填空题，程序还要再选 2~3 个实体作为选择题的干扰选项。最终生成的推理题还需要经过必要的自然语言润色，包括将空间图式转换为现实场景描述，将 A、B、C 等符号转为人名或物品名称等（如“小张、月季花”等）。

本节以“四人卡座”空间图式的出题流程为例，介绍这两个模块的具体工作流程。

（一）推理模块

推理模块的工作流程如图 4- 甲所示。

1. 用户配置解析：接收输入参数（如空间图式类型、题目数量、涉及实体数量等）。

2. 加载知识库：

①加载初始事实，以上文给出的四人卡座初始事实（对应图 1- 丁）为例。

②加载该空间图式对应的空间关系信息的逻辑关系和推理规则，例如，在四人卡座空间图式中，“X 在 Y 的右手边”等价于“Y 在 X 的左手边”。“X 在 Y 的左手边”和“Z 在 X 的正对面”可推导出“Z 在 Y 的斜对面”。

3. 将初始事实输入到事实库中，推理器接受事实库作为输入，开始推理。

4. 循环推理：推理器遍历当前事实库中的命题，将它们与空间关系信息的逻辑关系和推理规则进行匹配，生成新事实，并将新事实加入到事实库中。例如，命题 1“B 在 A 的右手边”应用逻辑关系推导出新命题“A 在 B 的左手边”。更新后的事实库再次输入到推理器中，对逻辑关系和推理规则进行匹配，例如新事实“A 在 B 的左手边”与“C 在 B 的正对面”应用表 5 所示的四人卡座推理规则，即可得到新事实“A 在 C 的斜对面”。

5. 若无新事实生成，则循环推理终止，输出事实库。同时记录全部推理路径。

从上述四人卡座空间图式的 3 条初始命题出发，经过推理模块后，输出的事实库

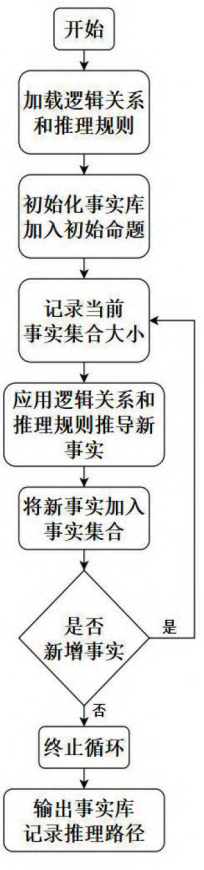


图 4- 甲　推理器

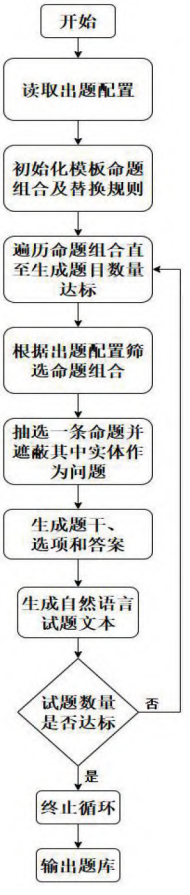


图 4- 乙　出题器

图 4　AutoQuest 程序流程图

中共有 158 条事实。这些事实（命题）接下来作为输入，进入出题模块，用于构建题干和问题。

（二）出题模块

出题模块的主要工作流程如图 4- 乙所示：

1. 筛选题干命题：从事实库中筛选出等价于初始事实的一组事实。
2. 根据出题设置，从事实库中抽取一条命题，隐去其中的实体，作为问题，例如：
   抽选命题 5：C 在 A 的斜对面。
   隐去实体 C，得到问题：___ 在 A 的斜对面？
3. 构建选项：(1) B (2) D (3) C (4) 以上选项都不是
4. 记录问题涉及到的空间关系、答案的数量、推理的过程等信息。
5. 为题目添加前导语。
6. 建立抽象变量与具体的自然语言实体的映射，例如：A 映射为铁拐李，B 映射为汉钟离，C 映射为姜子牙，D 映射为何仙姑。

最终输出题目如下：

> 四人来到饭店吃火锅，选了四人卡座坐下。卡座分列一张长方形桌子长边两侧，每排卡座上坐两人。面对面而坐。已知：汉钟离在铁拐李的右手边，姜子牙在汉钟离的正对面，何仙姑在姜子牙的右手边。
>
> **问题：** ____ 在铁拐李的斜对面？
>
> **选项：** A. 汉钟离 B. 何仙姑 C. 姜子牙 D. 以上选项都不是
>
> **答案：** C. 姜子牙

值得一提的是，出题模块可以控制生成题目的难度。例如，在上述示例中，抽选到的问题命题是“___ 在铁拐李的斜对面”，根据已知条件，至少需要两步推理才能得到答案（由命题 1 和命题 2 推出）；如果抽选到的命题是“___ 在汉钟离的正对面”，因为题干中已经出现了“姜子牙在汉钟离的正对面”，因此只需要 1 步推理便能推知答案（由命题 2 推出）。

四　大模型在空间推理任务上的测试结果初步分析

基于人工编写的空间方位关系推理知识库和 AutoQuest 自动出题程序，我们构建了一个空间关系推理数据集，并作为 SpaCE2024（https://2030nlp.github.io/SpaCE2024/）中的一个子任务（空间方位关系推理，Spatial Position Reasoning, SPR），评估大语言模型的空间推理能力。六个参赛系统全部采用大语言模型完成任务，有关参赛系统选择的模型以及提分策略可参见 SpaCE2024 的技术报告（Xiao et al., 2024；王士权等，2024）。下面重点分析参赛系统的评测结果。

（一）总体表现：准确率和稳定性

SpaCE2024 使用准确率和稳定性指标，衡量大语言模型在空间推理任务上的总体表现，在宏观层面进行系统间的横向对比。准确率得分反映模型推理能力的高低，稳定性得分呈现模型推理能力的稳健程度。我们对 10 道测试题目进行了不影响推理过程和结果的形式干扰，包括：让题目重复出现、改变问题的提问方式、变换选项顺序，形成 3 道“稳定题”。如果大模型在测试题及稳定题上都给出一致的推理结果，那说明大模型能够稳健地进行空间推理。表 6 展现了六个参赛系统的最好成绩和平均成绩。

表 6　SpaCE2024 空间推理任务参赛系统的准确率和稳定性

考察指标	评测结果
准确率	最好系统：0.37；最差系统：0.31；均值：0.34
稳定性	最好系统：0.80；最差系统：0.40；均值：0.52

在准确率上，表现最佳的系统仅 37%，显示基于本文方法生成的空间关系推理测试数据集，对大语言模型具有高挑战性。在稳定性上，大模型的空间推理能力在面对形式干扰时表现出显著的不稳定性，特别是对选项的顺序较为敏感。此外，模型在部分稳定题上表现为“稳定地错”，模型可能建立了系统但错误的推理模式，发现并研究此类错误有助于为模型调优提供方向。

（二）细粒度指标分析

1. 从空间关系角度考察大语言模型的准确率

表 7　从空间关系不同类型角度看参赛系统的得分（部分关系）

空间关系	测试集题量	参赛系统均值	空间关系	测试集题量	参赛系统均值
横向不相邻	225	0.47	右	514	0.31
正对面	56	0.43	东	32	0.31
斜对面	113	0.40	南	22	0.30
逆时针	116	0.37	北	41	0.26
顺时针	93	0.35	西	59	0.25
横向相邻	263	0.35	上	37	0.24
左	491	0.34	下	35	0.24

表 7 展示了大模型在考察题量较大的空间关系上的准确率得分。模型在推理对面关系、相邻关系和顺逆时针关系时相对更好，但在推理左右、东西南北和上下关系时表现不佳。这可能是由于前者通常与情景中的结构逻辑排列有关，如六边形中的对面关系即“+3”，不强调外部参考系，模型可通过序列模式进行建模；而后者则依赖明确的空间参考框架，如观察者朝向、物体自身朝向或地图方位，这类空间关系的认知和建模建立在身体经验和感知框架之上，当前的语言模型在这一维度严重缺失，进而在推理能力上表现出不足。为了形成更多的规律性认识，后续我们将利用本文生成方法的优势，精准控制知识库中的空间关系标签，定量、大量生成特定空间关系的题目进行评测。

2. 从空间图式的角度考察大语言模型的准确率

表 8 展示了大模型在不同空间图式上的准确率得分。模型在四人卡座图式中的表现明显好于六角图式和三层两列图式，可能受到实体数量和空间关系类型数量的影响。图式实体越多，空间关系类型越多，需要处理的空间关系越复杂。值得注意的是，当实体的位置信息在题目中被隐去时，以上规律并不显现。模型在隐去一个实体位置信息的题目上的准确率为 0.25，而在所有实体位置都出现的题目上可达 0.28，说明大模型补全空间场景以及推理隐含空间信息的能力还有待提高。

在六角图式中，模型在没有朝向限制的图式上的表现显著优于有朝向限制的图式。模型可能难以对东南西北及其复合方向构建稳定、统一的绝对方向参考系，导致在此类图式中出现推理错误。此外，模型在离心六角图式的得分略低于向心六角图式，可能与六人背向中心的情景在生活中较为少见有关，模型缺乏在离心六角图式中推理的经验。

表 8　从空间图式不同类型角度看参赛系统的得分

空间图式	朝向 / 方位	测试集题量	参赛系统均值
四人卡座	无东南西北	59	0.58
	东西	60	0.57
	南北	60	0.63
离心六角	无东南西北	260	0.42
	东西 + 复合方向	260	0.26
	南北 + 复合方向	260	0.24
向心六角	无东南西北	261	0.43
	东西 + 复合方向	260	0.25
	南北 + 复合方向	260	0.29
三层两列	东西	260	0.28

（三）对大语言模型空间方位关系推理能力的多角度考察

表 9　SpaCE2024 空间推理任务评测结果多角度汇总表

考察项	评测结果
准确率	最好系统：0.37；最差系统：0.31；均值：0.34
稳定性	最好系统：0.80；最差系统：0.40；均值：0.52
依赖信息条数	依赖信息越少，模型表现越好；依赖信息越多，表现越差
答案选项个数	答案选项个数越少，模型表现越好；答案选项个数越多，表现越差
空间图式类型	四人卡座 > 向心 / 离心六角 > 三层两列
空间关系类型	对面 > 相邻（横向、纵向、顺逆时针）> 左右 > 东南西北 > 上下
表达精确程度	精确 > 非精确
空间实体数量	4/4>6/6>5/6

除了空间关系和空间图式的角度，基于本文方法生成的评测数据集还支持从依赖信息条数、答案选项个数、表达精确程度等角度对准确率进行细粒度分析。表 9 汇总了 SpaCE2024 空间推理任务的多角度评测结果。

依赖信息条数和答案选项个数从推理能力维度考察模型的表现。“依赖信息条数”指求解题目必须使用的已知条件数量。参赛系统在依赖一条信息的题目上仅有 0.5 的准确率，且随着信息条数的增加而下降。“答案选项个数”指选择题的答案数量。大模型在多选题上的准确率得分比单选题低 15%，准确率随着答案数量的增加而下降。从推理的角度看，正确选项数的增加不仅意味着输出空间增大，更意味着模型需要独立地构建多个推理路径，并对多个中间结论进行组合判断。空间表达的精确程度同样影响推理路径的数量。精确的空间表达指示一个具体的位置，如“顺时针第一个”，而非精确的空间表达指示一片范围，如“不相邻”，意味着要推导多个实体的位置。这一过程对模型的推理能力提出了更高要求。

总体来看，空间推理任务的难度较高，参赛系统的平均准确率和稳定性均不理想。进一步的细粒度分析表明，大语言模型的空间推理能力在推理能力维度和空间知识维度的多个特征上存在短板，亟需通过专门的空间知识建模与推理能力训练进一步提升。

五　结语

本文提出了一种自动生成中文空间推理题的方法，通过人工构建小规模知识库，实现大规模、自动化生成高质量数据集，用于评估大语言模型的空间推理能力。本文方法的优点可以概括为：

（1）题目生成受到命题逻辑的约束，能最大程度确保题库的正确性。

（2）题目的推理难度可通过设定推理步数进行量化，从而实现题目难度可控。

（3）出题过程中可自动采集知识库中蕴含的空间关系、推理所需步数和空间图式等特征，形成评估大语言模型空间推理能力的多维度、细粒度框架。

相较于以往的空间推理任务数据集，本文研究工作设计了 4 种空间图式的具体 10 种空间场景，包含了 30 种空间关系，大大拓宽了空间知识的覆盖面。同时，本文首次对大语言模型的空间推理能力进行了诊断式分析，从总体表现、推理能力、空间知识三个维度展开细致评估，发现空间推理任务对于当前大语言模型而言仍是高挑战任务，大语言模型的空间推理能力尚不成熟：

（1）在空间知识维度上，大语言模型不擅长推理左右、东南西北、上下等需要参照系的基础空间关系，在多人且空间关系复杂的场景中表现不佳。

（2）在推理能力维度上，大语言模型在多步推理与多项输出方面存在短板，而且依赖显式、完整的信息表达，推理隐含信息的能力有待提升。

基于空间图式的合成数据技术可以为评测大语言模型的空间推理能力提供推理难度可调、空间知识分布可控的高质量数据集。除用于模型评测外，未来生成更大规模数据集还可以直接用于模型训练。在后续研究中，一方面将考虑优化空间方位知识库，包括补充目前缺少的前后、里外、远近关系，以及增加空间图式的类型；另一方面，尝试将本文的合成数据思路拓展到其他领域，如时间常识、自然常识和社会关系常识推理等。我们期待，通过探索，能够在大语言模型时代，寻找到一个可以更好地发挥语言学研究成果的价值的可行路径，在人类专家知识和机器学习所需的数据之间架起一座桥梁。

［参考文献］

［1］ 王士权，付薇薇，方瑞玉等 . 基于上下文学习与思维链策略的中文空间语义理解［A］. 第二十三届中国计算语言学大会论文集［C］. 太原：中国中文信息学会，2024.

［2］ 肖力铭，孙春晖，詹卫东等 . SpaCE2022 中文空间语义理解评测任务数据集分析报告［A］. 第二十二届中国计算语言学大会论文集［C］. 哈尔滨：中国中文信息学会，2023.

［3］ 肖力铭，詹卫东，穗志方等 .CCL23-Eval 任务 4 总结报告：第三届中文空间语义理解评测［A］. 第二十二届中国计算语言学大会论文集［C］. 哈尔滨：中国中文信息学会，2023.

［4］ 詹卫东，孙春晖，岳朋雪等 . 空间语义理解能力评测任务设计的新思路——SpaCE2021 数据集的研制［J］. 语言文字应用，2022，（2）.

［5］ 詹卫东 . 如何评估机器的语言能力？［C］. ChatGPT 来了——语言科学如何看待 ChatGPT［M］. 上海：上海教育出版社，2024.

［6］ Dai，H.，Liu，Z.，Liao，W.，et al. AugGPT：Leveraging ChatGPT for text data augmentation［J］. arXiv preprint arXiv：2302.13007，2023.

［7］ Eldan，R. & Li，Y. TinyStories：How small can language models be and still speak coherent english？［J］. arXiv preprint arXiv：2305.07759，2023.

[8] Josifoski,M.,Sakota,M.,Peyrard,M.,& West,R. Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction[J]. arXiv preprint arXiv: 2303.04132, 2023.

[9] Li,F.,Hogg,D. C. & Cohn,A. G. Reframing spatial reasoning evaluation in language models: A real-world simulation benchmark for qualitative reasoning[J]. arXiv preprint arXiv: 2405.15064, 2024.

[10] Li,X.,Yu,P.,Zhou,C.,et al. Self-Alignment with instruction backtranslation[J]. arXiv preprint arXiv: 2308.06259, 2023.

[11] Liu,H.,Ning,R.,Teng,Z.,et al. Evaluating the logical reasoning ability of ChatGPT and GPT-4[J]. arXiv preprint arXiv: 2304.03439, 2023.

[12] Mirzaee,R. & Kordjamshidi,P. Transfer learning with synthetic corpora for spatial role labeling and reasoning[J]. arXiv preprint arXiv: 2210.16952, 2022.

[13] Mirzaee,R.,Faghihi,H. R.,Ning,Q.,& Kordjmashidi,P. SpartQA: A textual question answering benchmark for spatial reasoning[J]. arXiv preprint arXiv: 2104.05832, 2021.

[14] Mukherjee,S.,Mitra,A.,Jawahar,G.,et al. Orca: Progressive learning from complex explanation traces of GPT-4[J]. arXiv preprint arXiv: 2306.02707, 2023.

[15] Shi,Z.,Zhang,Q. & Lipani,A. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts[A]. Proceedings of the AAAI Conference on Artificial Intelligence[C]. 2022,(10).

[16] Suhr,A.,Lewis,M.,Yeh,J.,& Artzi,Y. A corpus of natural language for visual reasoning[A]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) [C]. Vancouver: Association for Computational Linguistics, 2017.

[17] Talmy,L. The fundamental system of spatial schemas in language[A]. From Perception to Meaning: Image Schemas in Cognitive Linguistics[C]. Berlin: Mouton de Gruyter, 2005.

[18] Wu,D.,Zhang,J. & Huang,X. Chain of thought prompting elicits knowledge augmentation[J]. arXiv preprint arXiv: 2307.01640, 2023.

[19] Xiao,L.,Hu,N.,Zhan,W.,et al. Overview of CCL24-Eval task 3: The fourth evaluation on Chinese spatial cognition[A]. Proceedings of the 23rd China National Conference on Computational Linguistics (Volume 3: Evaluations)[C]. Taiyuan,China: Chinese Information Processing Society of China, 2024.

[20] Yehudai,A.,Carmeli,B.,Mass,Y.,et al. Genie: Achieving human parity in content-grounded datasets generation[J]. arXiv preprint arXiv: 2401.14367, 2024.

(责任编辑 常文斐)