引用格式：吴若玲,郭旦怀.大语言模型空间认知能力测试标准研究[J].地球信息科学学报,2025,27(5):1041-1052. [Wu R L, Guo D H. Research on evaluation standards for spatial cognitive abilities in large language models[J]. Journal of Geo-information Science, 2025,27(5):1041-1052.] **DOI:**10.12082/dqxxkx.2025.240694; **CSTR:**32074.14.dqxxkx.2025.240694

大语言模型空间认知能力测试标准研究

吴若玲,郭旦怀*

北京化工大学 信息科学与技术学院,北京 100029

Research on Evaluation Standards for Spatial Cognitive Abilities in Large Language Models

WU Ruoling, GUO Danhuai*

College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China

Abstract: [Objectives] Understanding whether Large Language Models (LLMs) possess spatial cognitive abilities and how to quantify them are critical research questions in the fields of large language models and geographic information science. However, there is currently a lack of systematic evaluation methods and standards for assessing the spatial cognitive abilities of LLMs. Based on an analysis of existing LLM characteristics, this study develops a comprehensive evaluation standard for spatial cognition in large language models. Ultimately, it establishes a testing standard framework, SRT4LLM, along with standardized testing processes to evaluate and quantify spatial cognition in LLMs. **[Methods]** The testing standard is constructed along three dimensions: spatial object types, spatial relations, and prompt engineering strategies in spatial scenarios. It includes three types of spatial objects, three categories of spatial relations, and three prompt engineering strategies, all integrated into a standardized testing process. The effectiveness of the SRT4LLM standard and the stability of the results are verified through multiple rounds of testing on eight large language models with different parameter scales. Using this standard, the performance scores of different LLMs are evaluated under progressively improved prompt engineering strategies. **[Results]** The geometric complexity of input spatial objects influences the spatial cognition of LLMs. While different LLMs exhibit significant performance variations, the scores of the same model remain stable. As the geometric complexity of spatial objects and the complexity of spatial relations increase, LLMs' accuracy in judging three spatial relations decreases by only 7.2%, demonstrating the robustness of the test standard across different scenarios. Improved prompt engineering strategies can partially enhance LLM's spatial cognitive Question-Answering (Q&A) performance, with varying degrees of improvement across different models. This verifies the effectiveness of the standard in analyzing LLMs' spatial cognitive abilities. Additionally, Multiple rounds of testing on the same LLM indicate that the results are convergent, and score differences between different LLMs exhibit a stable distribution. **[Conclusions]** SRT4LLM effectively measures the spatial cognitive abilities of LLMs and serves as a standardized evaluation tool. It can be used to assess

收稿日期：2024-12-17；**修回日期：**2025-02-25.

基金项目：国家自然科学基金项目（42371476）；中央高校基本科研业务费资助项目（buctrc202132）。 [**Foundation items:** National Natural Science Foundation of China, No.42371476; Fundamental Research Funds for the Central Universities, No. buctrc202132.]

作者简介：吴若玲（2001— ），女，北京人，硕士生，主要从事地理大模型研究。E-mail: ling010511@163.com

***通讯作者：**郭旦怀（1973— ），男，江西南康人，博士，教授，主要从事地理人工智能理论与应用研究。E-mail: gdh@buct.edu.cn

LLMs' spatial cognition and support the development of native geographic large models in future research.

Key words: large language model; spatial cognition; spatial relation; evaluation standard; prompt engineering; chain of thought; spatial scene

***Corresponding author:** GUO Danhuai, E-mail: gdh@buct.edu.cn

摘要:【目的】大语言模型是否具有空间认知能力及其水平的量化方法,是大模型和地理信息科学研究的重要问题。目前还缺乏大语言模型空间认知能力测评的系统方法和标准。本文在研究现有大语言模型特征的基础上开展大语言模型在空间认知问题上的测试标准研究,最终形成一套具备大语言模型空间认知能力的测试标准框架SRT4LLM和测试流程,用以评估大语言模型的空间认知能力。**【方法】**SRT4LLM分别从空间场景中的空间对象类型、空间关系和Prompt策略3个维度研究构建测试体系,包括3种空间对象类型、3种空间关系和3种提示工程策略,并制定标准化测试流程。通过对8个不同参数的大语言模型进行多轮测试,验证SRT4LLM标准的有效性和结果的稳定性,据此标准测试不同大模型在逐次改进的提示工程的得分。**【结果】**输入空间对象的几何复杂度影响大语言模型空间认知,不同大模型表现差异较大,但同一大模型得分较为稳定,随着空间对象几何复杂性和空间关系复杂度的增加,大语言模型对3种空间关系的判断准确率最大仅有7.2%的小幅下降,显示本测试标准在不同场景间的较强的适应性。改进的提示工程能够部分改善大语言模型的空间认知问答得分,且不同大模型的改善程度差异较大,本标准具备挖掘大模型空间认知能力的能力。同一大语言模型多轮测试表明测试结果收敛,大模型之间的分值差也较稳定。**【结论】**SRT4LLM具备度量大语言模型空间认知能力的能力,可作为大语言模型空间认知能力的评估的标准化工具,可为下一步构建原生的地理大模型提供支持。

关键词:大语言模型;空间认知;空间关系;测试标准;提示工程;思维链;空间场景

1 引言

以ChatGPT、DeepSeek为代表的大语言模型(简称“大模型”)因其基于大规模文本数据预训练,在自然语言理解、文本补全、多语种翻译、自动摘要和文本生成等自然语言处理任务中表现出接近甚至优于人类的能力[1-2]。此外,大模型在诸如图像理解、声音复刻、视频生成等[3-4]非语言理解领域也展现出广阔的应用前景。研究发现,大模型具备处理地理描述文本的能力,可以回答地理相关的非推理性问题,也能够根据语言描述生成地图[5]。进一步研究发现,大模型输出结果会因为空间对象的输入方式和提示方法的不同产生波动,在部分任务中波动高达40%[6]。现有大模型在自然语言理解能力逐渐趋同,但对空间问题的理解能力呈现较大差异。大模型是否具有空间认知能力?如何量化不同大模型在空间认知能力上的差异?成为迫切需要解决的研究问题。

空间认知是人类认知能力的重要组成部分,也是地理学研究的基本问题之一。空间认知一般是以空间场景作为载体开展研究,经典的空间场景认知能力一般以空间关系作为空间认知的媒介,拓扑关系、方向关系和距离关系是空间关系的重要组成部分,定性空间关系的描述与相似性度量是常用量表[7-10]。这些经典方法的有效性在众多理论研究和认知测试中得到验证,人工智能模型的空间认知能力研究也基本延续这一量表方法。如Ji等[11]评估了大语言模型在几何形状和空间关系表示方面的能力,实验结果显示空间关系表示准确率可达73%。Yamada等[12]提出了一种基于空间网格方法,用于评估大模型对序列输入中隐式空间结构的表示能力,尽管该方法在空间接近性问题存在明显错误,但对其他空间关系问题能提供较合理的答案。作为RCC模型的提出者,Cohn采用对话的方式对大语言模型的空间常识边界进行了辩证评估[13],通过聚焦定性空间推理任务,对ChatGPT-4的RCC-8关系的推理能力进行了3项实验评估,揭示其在此方面的表现特征[14]。这些研究表明,大语言模型具备一定的空间认知能力,不同大模型空间认知能力有显著差异,但目前的研究没有提出对大模型空间认知能力的统一测试标准。

在大模型的空间认知研究中,大模型的“幻觉问题”是一个无法回避的问题。“幻觉问题”指的是大模型输出的回答与客观事实或用户预期显著不一致,但形式规范的现象[15],提示工程(Prompt Engineering)是抑制大模型幻觉问题的重要方法之一,即通过为特定的推理任务设计策略性提示(Prompt),在不改变大模型参数的情况下,引导大模型产生更准确且符合用户预期的输出,降低生成内容的不确定性,从而提升大模型的实际效能。随着大模型技

术的发展，提示工程的研究日益深入，如Few-Shot Prompting通过为大模型提供少量但高质量的输入输出示例[16]，得以提高大模型在复杂任务上的理解能力；思维链（Chain-of-Thought, CoT）[17]则通过引入一系列中间推理步骤，约束大模型推理方向，从而提高大语言模型执行复杂推理的能力。目前缺乏专门针对空间认知能力的大模型提示工程方法的研究。此外，由于空间对象的描述方式与文本存在较大的差异，研究人员观察到在大模型问答任务中，空间对象的几何特征和空间布局会对大模型的空间认知产生较大影响[18]，但缺乏量化研究。

在此背景下，本文针对大模型在空间认知的应用，构建评测标准，以揭示大模型认知能力的量化机制。具体而言，本文从空间对象的几何形态、空间对象的布局方式及提示策略3个维度，对大模型的性能进行系统性评估，旨在测试其在不同空间场景下的适应性与输出稳定性。同时，将输入对象的几何特征及空间分布等关键因素纳入测试框架，以确保评测的全面性与科学性。为此，本文提出了一种系统化测试框架SRT4LLM，并通过多轮实验验证当前大模型在该框架下的表现，从而评估测试标准的有效性及测试结果的稳定性。

2　大模型空间认知测试标准

依据ISO的标准，软件测评标准应涵盖标准化要素，包括术语、流程、文档、技术规范及适用于各软件开发生命周期阶段的测试评估模型。本文聚焦于大模型的空间认知能力测试，重点研究测评标准、测试流程设计及测试样例的构建。首先，构建多层次测试标准框架与测试流程，涵盖测试空间场景、空间关系、提示工程和测试步骤，以全面覆盖空间对象形态与布局的多样性和难度梯度；为降低提示对测试结果潜在的干扰，结合Few-shot Prompting和CoT方法，优化提示策略，以评估不同大模型在多种提示方案下的适应性；最后，通过对不同参数规模的大语言模型进行多轮实验，验证测试标准的有效性，为进一步优化大模型的空间认知能力提供理论支撑。

2.1　大模型空间认知测试框架SRT4LLM

大模型的空间认知能力测试与传统软件测试存在显著差异，为规范测试内容并减少不同大模型在不同测试数据中的得分偏差，需构建标准化空间认知测试框架。本文提出了面向大模型的空间认知测试标准框架SRT4LLM（图1），采用递进式的测试方法，逐层挖掘大模型对空间关系的理解与推理能力，确保测试的全面性和科学性。SRT4LLM从以下4个层面对大模型测试进行约束：

（1）空间关系：涵盖拓扑关系、方位关系和距离关系。这3类关系是空间认知的核心要素，构成了所有复杂空间场景的基础。SRT4LLM以此为认知基准，构建系统化测试方案。

（2）空间场景：包括空间对象的输入方式、空间对象类型的组合方式和空间场景复杂度3个维度，均遵循从简单到复杂的设计原则。空间对象几何类型涵盖圆形、矩形和多边形；空间关系主要涉及面–面关系和点–线–面关系；简单空间场景由基础几何对象圆形和矩形构成，复杂空间场景则包含多种不规则多边形、点和折线的组合。

（3）提示工程策略：采用3种复杂度渐进的方法——简单提示、引导提示和示例提示，以充分挖掘不同大模型的空间认知能力，并评估其在不同提示策略下的表现。

（4）测试脚本模板：大模型具备较强的上下文理解能力，为减少测试顺序对结果的干扰，制定标准化的测试脚本，涵盖任务指令、空间关系定义、测试对象坐标序列和提示工程策略，以确保测试过程的可重复性与稳定性。

2.1.1　空间关系定义

定性空间关系描述更符合人类的空间感知与表达习惯[19]。本文基于定性空间关系对大模型进行测试，涵盖3类基本关系：拓扑关系、方向关系和距离关系。

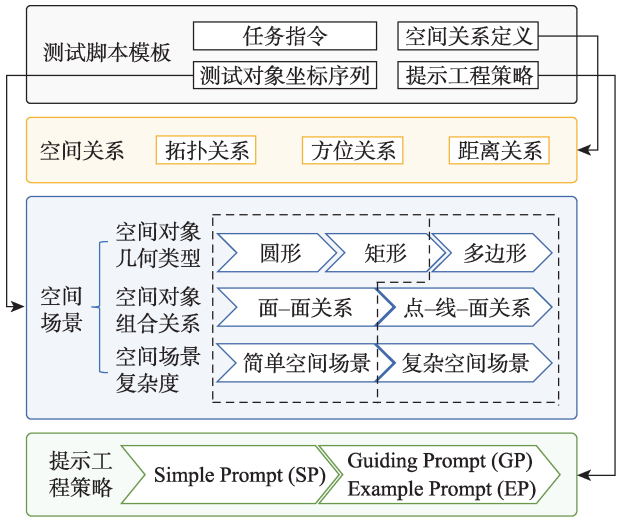


图1　SRT4LLM测试架构

Fig. 1　The test architecture of SRT4LLM

(1)拓扑关系

拓扑关系描述空间实体之间的位置逻辑结构关系,独立于具体形状和大小。SRT4LLM采用RCC-8模型[7-8]对拓扑关系进行刻画,包括DC(离散)、EC(外接)、PO(部分重叠)、TPP(内接)、NTPP(被包含)、TPPi(包含)、NTPPi(外部包含)和EQ(相等)8种关系(图2)。

(2)方向关系

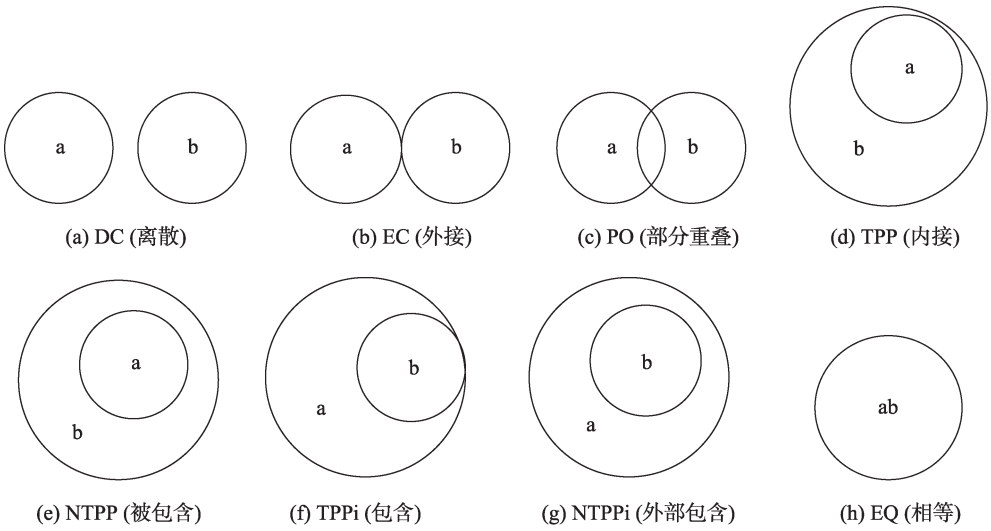
方向关系刻画空间实体的空间排列顺序。SRT4LLM采用基于投影的9方位系统,包含Up(上)、Down(下)、Left(左)、Right(右)、Upper Left(左上)、Lower Left(左下)、Upper Right(右上)和Lower Right(右下)8种关系(图3)。

(3)距离关系

距离关系描述空间实体之间的定性距离约束。SRT4LLM采用单调递增的距离范围划分方式[10](图4)。考虑不同空间场景的坐标尺度差异,在简单的空间场景中,δ_0=2,δ_1=4,在复杂的空间场景中,δ_0=10,δ_1=20,根据此定义,将参考对象的周围空间分为Close、Medium和Far 3种层次关系,以确保距离关系的稳定性与适用性。

2.1.2 空间场景设置

在大模型测试设计中,空间对象的表达形式对测试目标具有重要影响。输入真实地名可用于评估大模型地理常识认知能力,而输入空间几何对象更适合测试其空间场景认知表现。为消除不同大模型中地名语料的语种差异导致的偏差,SRT4LLM采用几何对象组合作为空间场景输入。所有测试场景统一采用直角坐标系,空间对象的位置信息以(x, y)坐标形式表示。具体而言,点对象用二维坐标表示,线对象用顶点坐标序列,面状对象采用首尾相连的坐标序列,而圆对象则通过中心



注:a和b表示2个空间对象。

图2 RCC-8模型[8]

Fig. 2 RCC-8 model[8]

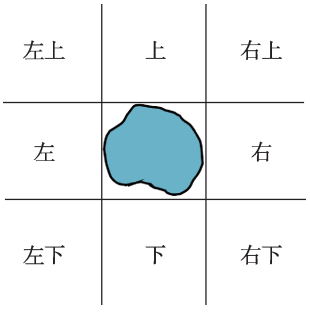
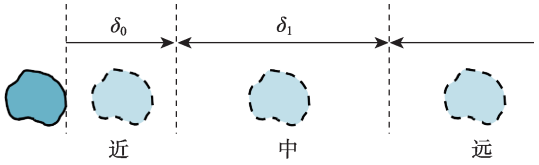


图3 基于投影的9方位系统[9]

Fig. 3 The 9-directional system based on projection[9]



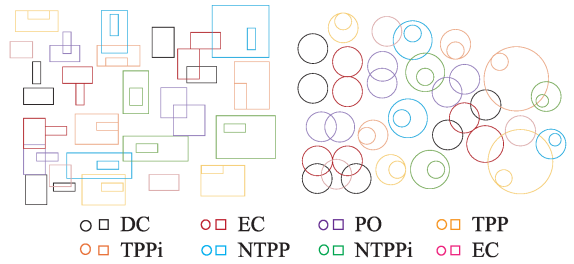
注:δ_0和δ_1表示递增的区间数值。

图4 单调递增的距离范围划分的定性距离表示[10]

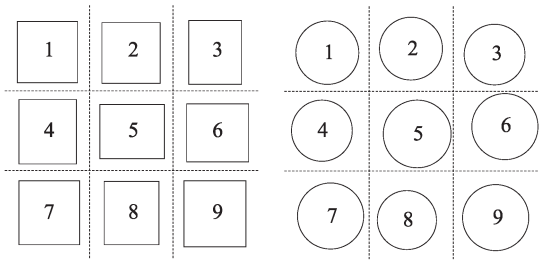
Fig. 4 Qualitative distance representation of monotonically increasing distance range division[10]

点坐标和半径进行描述。

空间场景中空间对象布局的方式可能影响大模型输出结果，为降低布局差异带来的测试结果波动，本研究设计了垂直、水平和对角线3种布局，以缓解单一布局可能导致的偏差。场景的复杂性方面，测试设计遵循2项原则：① 逐步增加空间对象的形态复杂度，从简单的圆形与矩形扩展到不规则多边形；② 增强空间对象组合关系的复杂度，将单一的面-面关系扩展到点、线、面之间的空间关系。基于此，在简单空间场景基础上，构建复杂测试场景，以评估大模型对不同复杂度空间认知任务的适应性。简单的空间场景（图5）基于3种布局（垂直、水平和对角线）设计测试任务，并针对3种空间关系（拓扑、方向、距离）构建测试集。其中，拓扑关系测试场景覆盖RCC-8模型中的8种关系，每种关系在3个布局共包含24组测试数据。方向关系测试场景从9个对象中随机选择3组，共有24组测试数据。距离关系测试场景则关注x和y_i之间的关系，每种距离关系包含8组测试场景，综合考虑不同布局，总计24组测试数据。

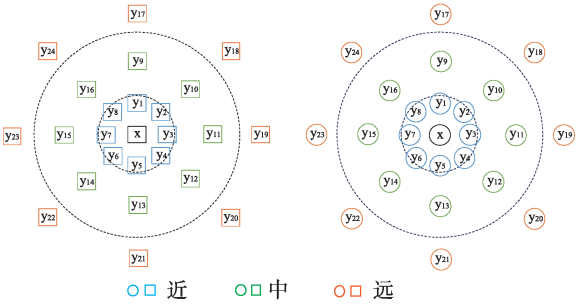


(a) 拓扑关系测试场景



注：数字1—9为空间对象编号。

(b) 方位关系测试场景



注：x表示距离关系测试中其中一个空间对象，y_i表示第i组测试中的另一个空间对象，i取值1~24，即24组测试。

(c) 距离关系测试场景

图5　简单空间测试场景

Fig. 5　Simple spatial test scenes

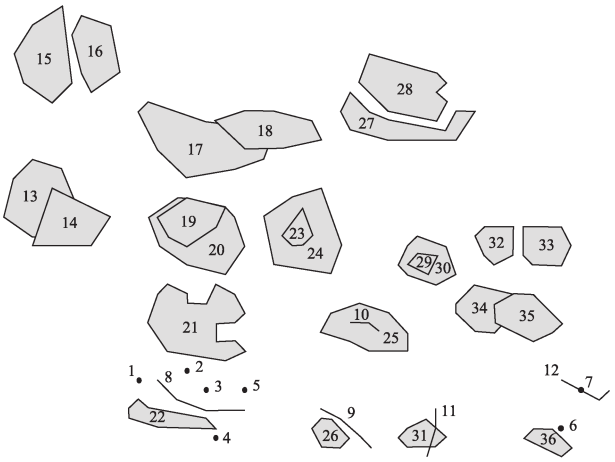
此外，文献[18]设计提出了一套用于定性空间关系的计算和空间认知测试的场景，涵盖更高复杂度的空间对象的形状与布局。本研究将该复杂场景（图6）纳入测试标准，作为评估大模型空间认知能力评估的基准测试集，以增强测试体系的全面性与有效性。

2.1.3　提示工程策略

在SRT4LLM框架中，为全面评估大语言模型的空间认知能力，采用3种提示工程策略：简单提示（Simple Prompt, SP）、引导提示（Guiding Prompt, GP）和示例提示（Example Prompt, EP）。不同提示策略通过不同上下文引导方式与推理需求，以系统的方式测试大模型的空间认知表现。

（1）简单提示(Simple Prompt, SP)

SP作为基础的测试，在无额外引导的情况下，主要衡量其预训练语料库中空间关系的存量及对空间关系的默认推理能力。SP仅包含任务指令和空间关系的定义（以```标签标识），并输入空间对象的坐标信息。在测试过程中这些作为上下文信息输入给大模型引导其回答。以拓扑关系为例（下同），SP算法如下：



注：数字1—36为空间对象编号。

图6　复杂空间测试场景参考[18]

Fig. 6　Complex spatial test scenes reference [18]

算法1 Simple Prompt (SP)

输入：空间对象形状*shape*; 空间对象坐标*coordinate*; 空间关系定义*relation*

输出：Simple Prompt模板

1. 初始化Simple Prompt模板：

Your task is to determine the topological relation between two closed geometrical shapes in the same coordinate system. The eight kinds of topological relations will be delimited with ``` tag. ```{*relation*}```.

The two geometrical shapes are {*shape*} that will be given their *positions* by coordinates: {*coordinate*}.

2. 将具体测试数据的*shape*、*coordinate*和*relation*分别替换模板中的占位符{*shape*}; {*coordinate*}; {*relation*}

3. 返回生成的SP提示语

（2）引导提示(Guiding Prompt, GP)

SP仅能测试出大模型知识库中的空间关系知识，难以衡量其推理能力。研究表明，优化提示词可以有效激发大模型的推理能力，其中，CoT是一种提升大模型复杂推理任务表现的技术。对于复杂问题，如算术推理(arithmetic reasoning)、常识推理(commonsense reasoning)、符号推理(symbolic reasoning)等，大模型很难直接给出正确答案。CoT通过要求大模型在输出最终答案之前，显式输出逐步的推理步骤这一方法来增强大模型的推理。由于拓扑关系(RCC-8)的层次递进性，GP能够引导大模型遵循这样的层次关系一步步推理（图7）。

Follow these steps in the reasoning process, The steps will be separated by {delimiter}:
Step 1: {delimiter} Determine the position in the coordinate system.
Step 2: {delimiter} Are x and y connected or discrete?
Step 3: {delimiter} If connected, are x and y externally connected or overlapping?
Step 4: {delimiter} If overlapping, are x and y partially overlapping or is one part of the other?
Step 5: {delimiter} Which is part of which?
Step 6: {delimiter} Are x and y tangent?

图7 标准思维链prompt模板

Fig. 7 Standard CoT template

SRT4LLM在设计引导提示时，简化了思维链的复杂推理过程，针对空间问题加入关键推理要点。测试表明，其效果显著优于具有复杂推理过程的CoT。GP算法如下：

算法2 Guiding Prompt (GP)

输入：空间对象形状*shape*; 空间对象坐标*coordinate*; 空间关系定义*relation*

输出：Guiding Prompt模板

1. 初始化Guiding Prompt模板：

Your task is to determine the topological relation between two closed geometrical shapes in the same coordinate system. The eight kinds of topological relations will be delimited with ``` tag. ```{*relation*}```.

The two geometrical shapes are {*shape*} that will be given their *positions* by coordinates: {*coordinate*}.

Pay attention to the following points in responding:

(1) By specifying the range of *x*-coordinate and *y*-coordinate, clearly define the positions of two geometrical shapes in the coordinate system.

(2) Two {*shape*} can only be overlapping if their *x*-coordinate and *y*-coordinate overlap at the same time.

(3) TPP, NTPP, TPPi, NTPPi and EQ are special cases of PO and should be categorized separately if their definitions are met. If not, categorized as PO.

2. 将具体测试数据的*shape*、*coordinate*和*relation*分别替换模板中的占位符{*shape*}; {*coordinate*}; {*relation*}

3. 返回生成的GP提示语

（3）示例提示(Example Prompt, EP)

EP在SP的基础上引入Few-shot策略，每次测试前提供2个推理示例，以指导大模型生成更优输出。该策略能够强化大模型对空间关系的理解，并减少错误分类。EP算法如下：

算法3 Example Prompt (EP)

输入：空间对象形状*shape*; 空间对象坐标*coordinate*; 空间关系定义*relation*

输出：Example Prompt模板

1. 初始化Example Prompt模板：

Your task is to determine the topological relation between two closed geometrical shapes in the same coordinate system. The eight kinds of topological relations will be delimited with ``` tag. ```{*relation*}```.

You will be given two cases to learn how to reason the question out.

Case 1 - The two geometrical shapes are {*shape*} that will be given their position by coordinates: rectangle *x*: (5, 6), (7, 6), (7, 7), (5, 7), (5, 6); rectangle *y*: (4, 5), (8, 5), (8, 8), (4, 8), (4,5).

Answer 1 - Based on the given coordinate information, the position of the two circles in the coordinate system can be determined:

rectangle *x*: *x*-coordinate ranges from 5 to 7, *y*-coordinate ranges from 6 to 7.

rectangle *y*: *x*-coordinate ranges from 4 to 8, *y*-coordinate ranges from 5 to 8.

Next, reason about the topological relation between two rectangles:
(1). The two rectangles overlap in both the *x* and *y* coordinate ranges, so it is not DC(*x, y*) or EC(*x, y*).
(2). Rectangle *x*'s x and *y* coordinates are both completely contained in rectangle *y*, so they are not partially overlapping or identical. It is a TPP(*x, y*) or NTPP(*x, y*) relation.
(3). Rectangle *x* and rectangle *y* are not tangent, so it is a NTPP (*x, y*) relationship.
Therefore, it is concluded that the two rectangles are NTPP(*x, y*).
Case 2 - The two geometrical shapes are {*shape*} that will be given their position by coordinates: rectangle x: (1, 2), (3, 2), (3, 5), (1, 5), (1, 2); rectangle y: (3, 3), (5, 3), (5, 4), (3, 4), (3, 3).
Answer 2 - Based on the given coordinate information, the position of the two rectangles in the coordinate system can be determined:
rectangle *x*: *x*-coordinate ranges from 1 to 3, *y*-coordinate ranges from 2 to 5.
rectangle *y*: *x* coordinate ranges from 3 to 5, *y* coordinate ranges from 3 to 4.
Next, reason about the topological relation between two rectangles:
(1). The *y*-coordinate ranges of the two rectangles overlap, but the *x*-coordinate ranges do not, so it is DC(*x, y*) or EC(*x, y*), not the others.
(2). The x-coordinate ranges of the two rectangles do not overlap, but are connected, indicating that the two rectangles are externally connected, so it is EC(*x, y*).
Therefore, it is concluded that the two rectangles are EC(*x, y*).
Question - The two geometrical shapes are {*shape*} that will be given their position by coordinates: {*coordinate*}.

2. 将具体测试数据的*shape*、*coordinate*和*relation*分别替换模板中的占位符{*shape*}; {*coordinate*}; {*relation*}
3. 返回生成的EP提示语

2.1.4 测试脚本模板

当前大模型的测试数据集[20-22]通常采用"问题–答案"对(Q&A)形式，并以客观选择题形式构建测试数据。然而，当测试空间关系认知能力时，传统选择题设计难以有效评估大模型的推理能力。主要原因在于，通用预训练语料库中的空间关系专业语料(如“拓扑分离”、“拓扑交叉”等)出现频率较低，导致大模型对相关概念的理解能力不足，从而影响测试有效性。此外，为消除上下文记忆对测试结果的干扰，亟需建立标准化的测试脚本模板。

本研究提出的测试模板在每次测试前向大模型提供空间关系的标准定义(表1)，以减少因概念模糊导致的理解偏差。同时，考虑到选择题可能对大模型的输出产生暗示，测试改为简答题形式，使大模型需独立完成推理与判断，从而更准确地评估其空间关系认知能力。

为降低因任务描述的上下文环境引起的差异导致的结果偏差，本研究提出一套标准化提问模板，包含4个核心要素：任务指令、空间关系定义、测试对象坐标序列及提示工程策略。通过结构化设计，该模板确保输入一致性，从而提供稳定的测试基准，支持后续性能评估。此外，鉴于测试涉及多语种大模型对比分析，为确保实验的一致性和可比性，所有测试均采用英文作为标准输入语言，尽可能减少因语言差异可能引入的变量，提高测试的客观性与严谨性。

表1 SRT4LM 3种基础空间关系的定义

Tab. 1 Definitions of three spatial relations in SRT4LLM

空间关系	SRT4LLM对空间关系的定义
拓扑关系	(1) DC(*x, y*): *x* is disconnected from y. (2) EC(*x, y*): *x* is externally connected to y without any overlap. (3) PO(*x, y*): *x* partially overlaps *y*, with neither being a part of the other. (4) TPP(*x, y*): *x* is a tangential proper part of *y*. (5) NTPP(*x, y*): *x* is a nontangential proper part of *y*. (6) TPPi(*x, y*): *y* is a tangential proper part of *x*. (7) NTPPi(*x, y*): *y* is a nontangential proper part of *x*. (8) EQ(*x, y*): *x* is identical with *y*.
方位关系	(1) Up(*x, y*): *y* is roughly above *x*. (2) Down(*x, y*): *y* is roughly below *x*. (3) Left(*x, y*): *y* is roughly to the left of *x*. (4) Right(*x, y*): *y* is roughly to the right of *x*. (5) Upper Left(*x, y*): *y* is roughly to the upper left of *x*. (6) Lower Left(*x, y*): *y* is roughly to the lower left of *x*. (7) Upper Right(*x, y*): *y* is roughly to the upper right of *x*. (8) Lower Right(*x, y*): *y* is roughly to the lower right of *x*.
距离关系	Qualitatively describe the relation by delimiting the distance range. (1) Close(*x, y*): The length of the distance from x to y is $[0, \delta_0]$. (2) Medium(*x, y*): The length of the distance from x to *y* is $(\delta_0, \delta_0+\delta_1]$. (3) Far(*x, y*): The length of the distance from x to *y* is $(\delta_0+\delta_1, +\infty)$.

2.2 SRT4LLM测试流程

SRT4LLM采用3类Prompt工程策略(SP、GP、EP),针对3类空间对象(圆形、矩形、多边形)进行测试。每类空间对象包含24组场景,对一种空间关系进行24×3×3=216次测试,3种空间关系共进行648次提问。测试在8个大模型上分别调用API自动化执行(图8),若测试过程中出现异常结果或因连接问题导致中断,则重启测试流程,确保数据完整性。最终,测试结果由人工统计正确分类的场景数,并计算正确率。随后进入下一轮测试迭代。

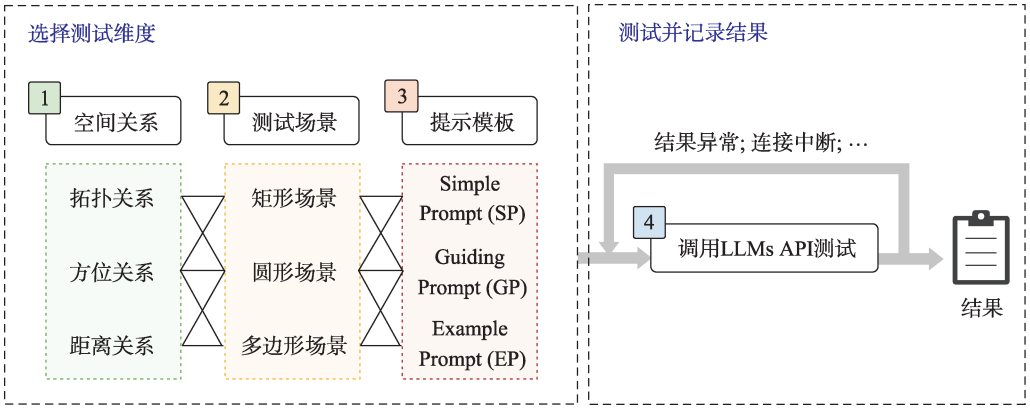


图8 SRT4LLM测试流程

Fig. 8 The evaluation process of SRT4LLM

3 实验与结果分析

3.1 大模型选择

为验证测评标准的有效性,本研究依据大模型的发布时间、用户规模及讨论热度等因素,选取了8种主流大模型作为测试对象,包括ChatGLM3,ERNIE Bot,Gemini,GPT-3.5,GPT-4,LLaMa2,QWEN和SparkDesk(表2)。测试实验在2024年1—2月开展,其中,ERNIE Bot、Gemini、GPT-3.5、GPT-4、QWEN和SparkDesk是通过调用官网API接口实现自动化测试,而ChatGLM3和LLaMa2基于阿里云的DashScope通过API接口进行调用。

3.2 测评结果分析

3.2.1 空间场景对象几何形状影响分析

实验设计涉及不同复杂度的几何场景,包括圆形、矩形和多边形,以逐步增加推理难度。测评结果(表3)显示,8个大模型在圆形、矩形和多边形空间场景的3种空间关系判断准确率分别为35.8%、38.8%、53.0%,28.6%、39.7%、51.4%,以及32.6%、37.2%、46.5%。

圆形对象由于其位置仅依赖于圆心与半径,判断相对简单,大模型整体表现较优。矩形对象需考虑边界条件,增加了几何复杂性,对大模型的推理能力提出更高要求。多边形对象则涉及多个顶点及边界的复杂空间关系,进一步提高了任务难度,导致准

表2 测试大模型

Tab. 2 Tested large language models

大模型名称及版本	发布机构	发布时间	测试版本号
ChatGLM3	智谱AI	2023年10月27日	ChatGLM3-6B
ERNIE Bot	百度	2023年3月16日	ERNIE-Bot-turbo-0922
Gemini	Google	2023年12月6日	gemini-pro
GPT-3.5	OpenAI	2022年11月30日	gpt-3.5-turbo
GPT-4	OpenAI	2023年3月15日	gpt-4-0125-preview
LLaMa2	Meta AI	2023年7月19日	LLaMa2-13B-chat
QWEN	阿里云	2023年3月16日	qwen-max
SparkDesk	科大讯飞	2023年5月6日	sparkv3.5

表3　3种空间场景上的测试准确率

Tab. 3　Accuracy on three spatial scenes　　(%)

大模型	拓扑关系			方位关系			距离关系		
	圆形	矩形	多边形	圆形	矩形	多边形	圆形	矩形	多边形
ChatGLM3-6B	22.2	13.9	4.2	11.8	11.8	13.9	25.0	18.0	11.1
ERNIE-Bot	9.7	16.0	8.3	10.4	10.4	18.1	34.8	30.6	29.2
Gemini-pro	38.2	25.0	45.8	38.2	37.5	40.3	51.4	58.3	55.6
GPT-3.5	40.3	27.8	29.2	54.2	52.1	41.7	81.9	85.4	45.8
GPT-4	78.5	72.9	63.9	87.5	95.1	72.2	98.6	92.4	79.2
LLaMa2-13B	22.9	14.6	11.1	14.6	18.1	15.3	25.0	32.0	16.7
Qwen-max	47.2	29.2	54.2	55.5	60.4	68.1	76.4	62.5	84.7
Sparkv3.5	27.1	29.9	44.4	38.2	31.9	27.8	31.3	32.0	50.0
平均值	35.8	28.6	32.6	38.8	39.7	37.2	53.0	51.4	46.5

确率下降。然而，整体而言，大模型在不同几何场景的表现未出现显著下降，体现了较强的适应性。

综上所述，SRT4LLM测试标准验证了大模型在不同场景的适应性，表明大模型在地理空间任务中的能力受场景复杂度影响。从8个大模型的测试结果（图9）看，除个别大模型在3类空间对象上的准确率有明显波动外，大部分大模型在不同空间对象类型中回答准确率相对稳定，表现出较高的一致性，进一步验证了SRT4LLM能够有效运用于各类空间场景。此外，各大模型在空间认知任务中的准确率存在较大差异，表明SRT4LLM具备测试大模型在空间认知能力上的区分能力。

3.2.2　提示工程影响分析

本研究采用3种提示工程策略：SP、GP和EP，为大模型提供不同程度的提示信息。测评结果（表4）表明，使用SP、GP和EP提示策略时，大模型在3类空间关系判断任务上的平均准确率分别为29.6%、32.3%、35.4%，31.0%、35.8%、47.4%以及44.3%、46.6%、57.3%。

SP策略适用于简单问题，但对复杂问题的推理指导不足。GP与EP策略能够显著提升大模型的

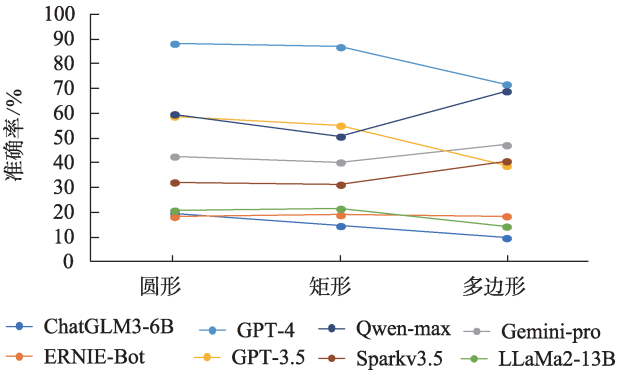


图9　大模型在3类空间场景上的准确率

Fig. 9　Accuracy of tested large language models on three types of spatial scenes

表4　Simple Prompt (SP)、Guiding Prompt (GP)和Example Prompt (EP)策略下的大模型空间认知准确率

Tab. 4　Spatial cognitive accuracy of large language models using Simple Prompt (SP), Guiding Prompt (GP), and Example Prompt (EP)　　(%)

大模型	拓扑关系			方位关系			距离关系		
	SP	GP	EP	SP	GP	EP	SP	GP	EP
ChatGLM3-6B	**12.0**	11.5	9.9	12.0	10.4	**16.1**	2.1	4.2	**42.7**
ERNIE-Bot	8.3	**13.5**	9.9	14.6	11.5	**16.7**	25.5	37.5	29.7
Gemini-pro	36.5	36.0	**43.7**	26.0	33.9	**57.3**	54.2	43.2	**68.2**
GPT-3.5	27.1	27.6	**40.1**	36.5	40.6	**62.0**	60.9	62.5	**70.9**
GPT-4	63.6	72.4	**73.4**	73.4	**87.5**	84.4	86.5	**89.6**	85.9
LLaMa2-13B	12.0	16.2	**16.7**	13.0	15.6	**18.7**	15.1	20.3	**32.3**
Qwen-max	40.1	45.3	**53.1**	50.5	63.6	**75.0**	67.7	71.9	**91.7**
Sparkv3.5	**37.0**	35.9	36.5	21.9	23.4	**49.0**	42.2	**43.3**	37.0
平均值	29.6	32.3	**35.4**	31.0	35.8	**47.4**	44.3	46.6	**57.3**

注：3类空间关系中3种Prompt的准确率最高得分用粗体标注，优化后准确率下降的加注下划线。

空间认知能力，主要原因可能有以下3点：

（1）上下文信息的补充：由于大模型缺乏人类的直觉认知，适当补充提示信息可以帮助其更好地理解任务。GP通过提供回答线索，EP通过提供回答示例，均为大模型的任务解析与推理提供了明确方向。

（2）注意力的有效引导：GP能够聚焦于关键信息，减少大模型对无关内容的注意分散；同时，使用准确而简洁的语言表述，使大模型接收到的提示信息更具有效性，从而提高回答准确性。

（3）复杂任务的分解：EP提供了问题解答的参考框架，帮助大模型分解复杂任务，将其转化为易于处理的逻辑步骤，从而增强大模型在处理复杂空间关系时的能力。

综上，提示工程策略有助于提升大模型空间认知性能，尤其是在复杂任务场景中，优化提示能够充分挖掘大模型的潜在能力，显著提高其在空间认知问题中的准确性和鲁棒性。

从8个大模型的评估结果（图10）看，提示工程对大模型空间认知能力的表现有显著影响。随着提示的复杂性逐步增加，大部分大模型的准确率呈现上升趋势。这一结果表明，恰当的提示工程能够提升大模型的空间认知表现。同时，SRT4LLM能够有效挖掘大模型的空间认知潜力。此外，各大模型在不同提示策略下的表现存在差异，也表明SRT4LLM能够区分大模型在空间认知能力上的差异，为后续大模型优化与提示设计优化提供重要依据。

3.2.3 稳定性分析

为了验证大语言模型在本研究评测标准下的结果稳定性，在每种空间关系的3类测试场景（圆形、矩形、多边形）中随机抽取10个测试样例，对8种大语言模型在SP条件下进行3轮测试，记录各轮测试结果，并计算均值和标准差（表5）。同时，对每个大模型在不同空间关系的准确率均值和标准差进行量化（图11），结果具有稳定性。

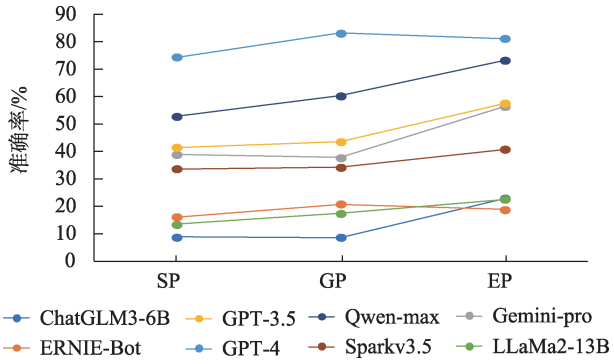


图10 大模型使用3种Prompt策略的准确率对比

Fig. 10 Accuracy comparison of tested large language models using three Prompt strategies

实验结果表明，大多数大模型在多轮测试中的准确率波动较小，显示出较高的稳定性。Sparkv3.5在所有测试任务中展现出最佳的稳定性，Qwen-max和GPT-4的准确率波动范围亦较小，标准差较低，表明其在不同测试场景下具备较高的一致性和稳定性。相较之下，ERNIE-Bot、Gemini-pro和GPT-3.5表现出较大的波动，稳定性相对较弱。此外，各大模型在拓扑关系、方位关系及距离关系3类任务上的表现较为均衡，Qwen-max大模型尤其突出。

整体而言，SRT4LLM能够有效评估大模型在不同空间任务和场景下的稳定性，并区分大模型在各任务类别中的性能。多数大模型的准确率具备稳定性，但部分大模型在特定任务的适应性存在差异，这进一步说明本评测标准的区分能力。

表5 多轮测试结果得分

Tab. 5 The result scores of multiple rounds of testing

大模型	拓扑关系					方位关系					距离关系				
	第一轮	第二轮	第三轮	平均值	标准差	第一轮	第二轮	第三轮	平均值	标准差	第一轮	第二轮	第三轮	平均值	标准差
ChatGLM3-6B	0.0	0.0	0.0	0.0	0.0	20.0	10.0	20.0	16.7	5.8	10.0	10.0	10.0	10.0	0.0
ERNIE-Bot	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.0	6.7	11.5	30.0	0.0	20.0	16.7	15.3
Gemini-pro	30.0	30.0	50.0	36.7	11.5	50.0	30.0	50.0	43.3	11.5	50.0	60.0	60.0	56.7	5.8
GPT-3.5	20.0	40.0	10.0	23.3	15.3	30.0	30.0	20.0	26.7	5.8	30.0	50.0	30.0	36.7	11.5
GPT-4	80.0	70.0	80.0	76.7	5.8	80.0	70.0	80.0	76.7	5.8	60.0	60.0	60.0	60.0	0.0
LLaMa2-13B	20.0	20.0	20.0	20.0	0.0	10.0	0.0	10.0	6.7	5.8	10.0	0.0	10.0	6.7	5.8
Qwen-max	60.0	60.0	70.0	63.3	5.8	70.0	60.0	60.0	63.3	5.8	60.0	60.0	60.0	60.0	0.0
Sparkv3.5	20.0	20.0	20.0	20.0	0.0	40.0	40.0	40.0	40.0	0.0	50.0	50.0	50.0	50.0	0.0

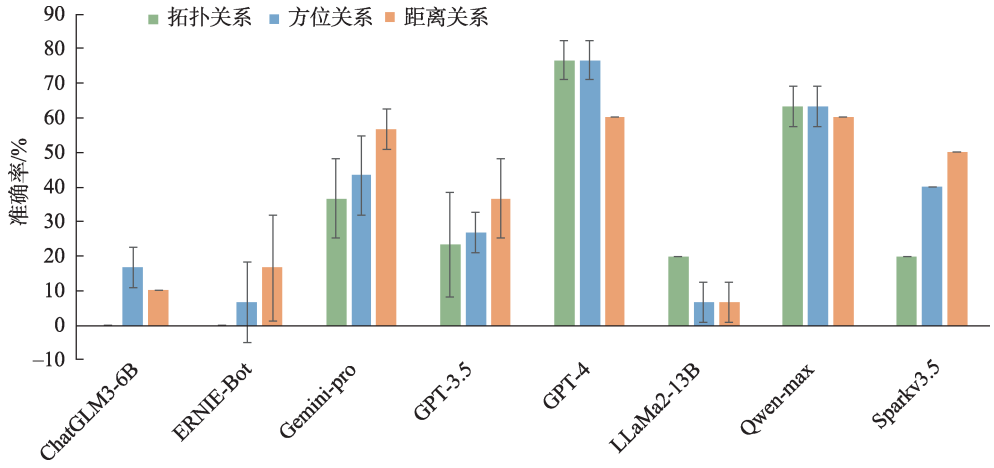


图 11　大模型在 3 类空间任务上的准确率平均值和标准差

Fig. 11　The average accuracy and standard deviation of tested large language models on three types of spatial tasks

4　结语

本研究围绕大语言模型空间认知能力，初步探讨了系统性测试方法，并提出了一套标准测试框架 SRT4LLM，从空间对象类型、空间关系和提示工程 3 个维度构建完整的评测体系，并设计了标准测试流程。实验结果表明，SRT4LLM 不仅能够评估大语言模型在几何复杂场景中的表现，还可量化其在处理复杂空间关系时的适应能力和局限性，验证了该标准在衡量大模型空间认知能力方面的有效性。在提示工程的优化下，大语言模型的测试结果整体呈现稳步提升，表明提示策略对大模型输出具有显著影响。然而，不同大模型对 Prompt 策略的敏感度存在较大差异，进一步佐证 SRT4LLM 具备区分大模型能力的科学性和实用性。此外，多轮测试结果的稳定性验证了 SRT4LLM 的可靠性，可为研发原生大语言地理模型提供重要参考。

SRT4LLM 仍存在一定的局限性，如：统一采用英文提问模式可能抑制了 ChatGLM3、ERNIE-Bot、QWEN 和 Sparkv3.5 等中文语料大模型的表现。未来研究可考虑引入多模态输入场景，以更全面地评估大模型的空间认知能力，并挖掘其在实际地理认知任务中的应用潜力。在本论文修改期间，DeepSeek 发布，本文虽未将其纳入测试，但测试结果将在测试代码中共享。

利益冲突：Conflicts of Interest

所有作者声明不存在利益冲突。

All authors disclose no relevant conflicts of interest.

作者贡献：Author Contributions

吴若玲完成实验设计与实验，吴若玲和郭旦怀负责论文的写作和修改。所有作者均阅读并同意最终稿件的提交。

The experimental design and operation were completed by WU Ruoling. The manuscript was drafted and revised by WU Ruoling and GUO Danhuai. All authors read and approved the final manuscript for submission.

SRT4LLM 代码及测试用例见：https://github.com/LLING000/SRT4LLM

参考文献(References):

[1] Qin C W, Zhang A, Zhang Z S, et al. Is ChatGPT a general-purpose natural language processing task solver? [C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2023:1339-1384. DOI:10.18653/v1/2023.emnlp-main.85

[2] 陈炫婷,叶俊杰,祖璨,等. GPT 系列大语言模型在自然语言处理任务中的鲁棒性[J]. 计算机研究与发展,2024,61(5):1128-1142. [Chen X T, Ye J J, Zu C, et al. Robustness of GPT large language models on natural language processing tasks[J]. Journal of Computer Research and Development, 2024,61(5):1128-1142.] DOI:10.7544/issn1000-1239.202330801

[3] Jin P, Takanobu R, Zhang W C, et al. Chat-UniVi: Unified visual representation empowers large language models with image and video understanding[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2024:13700-13710. DOI:10.1109/CVPR52733.2024.01300

[4] 陈露,张思拓,俞凯. 跨模态语言大模型:进展及展望[J]. 中国科学基金,2023,37(5):776-785. [Chen L, Zhang S T,

Yu K. Cross-modal large language models: Progress and prospects[J]. Bulletin of National Natural Science Foundation of China, 2023, 37(5): 776-785.] DOI: 10.16262/j.cnki.1000-8217.20231026.006

[5] Zhang Y F, Wei C, He Z T, et al. GeoGPT: An assistant for understanding and processing geospatial tasks[J]. International Journal of Applied Earth Observation and Geoinformation, 2024,131:103976. DOI:10.1016/j.jag.2024.103976

[6] He J, Rungta M, Koleczek D, et al. Does prompt formatting have any impact on LLM performance?[J]. arXiv preprint arXiv:2411.10541, 2024. https://arxiv.org/abs/2411.10541v1

[7] Randell D A, Cohn A G. Modelling topological and metrical properties in physical processes[C]//Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning. ACM, 1989: 357-368. DOI:10.5555/112922.112958

[8] Randell D A, Cui Z, Cohn A G. A spatial logic based on regions and connection[C]//Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning. ACM, 1992: 165-176. DOI: 10.5555/3087223.3087240

[9] Frank A U, Egenhofer M J. Computer cartography for GIS: an object-oriented view on the display transformation[J]. Computers & Geosciences, 1992, 18(8): 975-987. DOI:10.1016/0098-3004(92)90015-J

[10] Hernández D, Clementini E, Felice P D. Qualitative distances[C]//Spatial Information Theory A Theoretical Basis for GIS. Berlin, Germany: Springer, 1995:45-57. DOI: 10.1007/3-540-60392-1_4

[11] Ji Y H, Gao S. Evaluating the effectiveness of large language models in representing textual descriptions of geometry and spatial relations[C]//12th International Conference on Geographic Information Science (GIScience 2023) (Leibniz International Proceedings in Informatics (LIPIcs)). Wadern, Germany: Schloss Dagstuhl Leibniz Center for Informatics, 2023, 277(43): 1-6. DOI: 10.4230/LIPIcs.GIScience.2023.43

[12] Yamada Y, Bao Y H, Lampinen A K, et al. Evaluating spatial understanding of large language models[J/OL]. Transactions on Machine Learning Research, 2024, 1-22[2025-02-10]. https://openreview.net/forum?id=xkiflfKCw3

[13] Cohn A G, Hernandez-Orallo J. Dialectical language model evaluation: an initial appraisal of the commonsense spatial reasoning abilities of LLMs[J]. arXiv preprint arXiv: 2304.11164, 2023. https://arxiv.org/abs/2304.11164v1

[14] Cohn A G. An evaluation of ChatGPT-4's qualitative spatial reasoning capabilities in RCC-8[C/OL]//36th International Workshop on Qualitative Reasoning (QR-23). 2023, 1-8[2025-02-10]. https://staff. fnwi. uva. nl/b. bredeweg/QR2023/pdf/07Cohn.pdf

[15] Huang L, Yu W J, Ma W T, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-55. DOI: 10.1145/3703155

[16] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. ACM, 2020:1877-1901. DOI:10.5555/3495724.3495883

[17] Wei J, Wang X Z, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. ACM, 2022: 24824-24837. DOI:10.5555/3600270.3602070

[18] 郭旦怀. 基于空间场景相似性的地理空间分析[M]. 北京:科学出版社,2016. [Guo D H. Geospatial analysis based on spatial scene similarity[M]. Beijing: Science Press, 2016.]

[19] Chen J, Cohn A G, Liu D Y, et al. A survey of qualitative spatial representations[J]. The Knowledge Engineering Review, 2015,30(1):106-136. DOI:10.1017/s0269888913000350

[20] Hendrycks D, Burns C, Basart S, et al. Aligning AI with shared human values[C/OL]//Proceedings of the International Conference on Learning Representations. 2021, 1-29[2025-02-10]. https://iclr.cc/virtual/2021/poster/2960

[21] Hendrycks D, Burns C, Basart S, et al. Measuring massive multitask language understanding[C/OL]//Proceedings of the International Conference on Learning Representations. 2021, 1-27[2025-02-10]. https://iclr. cc/virtual/2021/poster/2962

[22] Huang Y Z, Bai Y Z, Zhu Z H, et al. C-Eval: A multi-level multi-discipline Chinese evaluation suite for foundation models[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. ACM, 2023:62991-63010. DOI:10.5555/3666122.3668871

本文图文责任编辑：蒋树芳　黄光玉