CCL24-Eval任务3系统报告: 基于大型语言模型的中文空间语义评测

霍世图[1], 王钰君[1], 吴童杰[1]
[1]北京师范大学　国际中文教育学院　北京　100875
{20222109017, 202221090022, 202221090014}@mail.bnu.edu.cn

摘要

本研究的任务旨在让大模型进行实体识别、角色识别、异常识别、信息推理、同义识别任务，综合评估大模型的空间语义理解能力。其中，我们使用普通提示词、工作流提示词和思维链三种提示词策略来探讨大模型的空间语义理解能力，最后发现ERNIE-4在1-shot的普通提示词上表现最佳。最终，我们的方法排名第六，总体准确率得分为56.20%。

关键词： 空间语义；大语言模型；提示词工程

System Report for CCL24-Eval Task 3: Evaluation of Chinese Spatial Semantics Based on Generative Language Models

Shitu Huo[1], Yujun Wang[1], Tongjie Wu[1]
[1]Beijing Normal University, School of International Chinese Education, Beijing 100895
{20222109017, 202221090022, 202221090014}@mail.bnu.edu.cn

Abstract

The task of this paper aims to comprehensively assess large models' spatial semantic understanding capabilities through entity recognition, role recognition, anomaly detection, information inference, and synonym recognition tasks. We explored the spatial semantic understanding capabilities of large models using three prompt strategies: general prompts, workflow prompts, and chain-of-thought prompts. In the end, we found that ERNIE-4 performed best with 1-shot general prompts. Our system ranked sixth overall, with an accuracy score of 56.20%.

关键词： Spatial Semantics；Large Language Model；Prompt Engineering

1 引言

在自然语言处理领域，大型语言模型取得了显著进展。现有的大模型主要基于注意力机制的Transformer (Vaswani et al., 2017)，利用缩放定律（Scaling Laws）大幅提升模型性能 (Kaplan et al., 2020)，从而使得BERT (Devlin et al., 2018)和GPT-3 (Floridi and Chiriatti, 2020)等模型能够捕捉复杂的语言结构和语境关系，甚至可以批量改写和生成逼真的文本，从而端对端地完成机器翻译、语义分析等自然语言处理任务 (Brown et al., 2020; Zhao et al., 2023)。

大语言模型具有较好的语义理解能力，与人类的语义判断具有极高的相关性，并展现出作为理论语言学新兴研究工具的潜力 (Tjuatja et al., 2023)。在早期，大语言模型常被视为一只“随机鹦鹉”，其学习机制是单纯统计某个词语出现的频率 (Bender et al., 2021)。然而后续的研究表明，大模型在没有遵循语言学理论的情况下，能够有效整合语义和语法的多重信息 (Piantadosi, 2023)，区分同一个词在不同语境中的多种用法 (Petersen and Potts, 2023)，甚至可以有效表征时间和空间 (Gurnee and Tegmark, 2023)。

空间语义评测是评估自然语言处理系统对空间表达理解能力的重要手段。传统的空间语义评测方法主要依赖人工标注，但这种方法标注成本高且可扩展性差。如今，基于生成式语言模型开展评测任务逐渐成为趋势。基于上述背景，本研究拟探究以下问题：1）大模型对空间语义的理解程度如何？2）在理解空间语义的具体任务上，大模型各有哪些优劣？

本研究基于第四届中文空间语义理解评测任务（SpaCE2024），首先介绍空间语义评测的背景和相关研究，然后通过实验分析不同模型的空间语义理解能力，最后对实验结果进行讨论和分析，以更好地了解大模型在空间语义理解方面的能力边界。

2 相关研究

这一部分首先介绍与空间语义有关的语言学研究，然后总结自然语言处理领域关于空间语义评测的研究概况。

2.1 语言学视角下的空间语义研究

认知语言学基于人们对世界的经验和对世界进行感知和概念化的方法 (张敏, 1998)，借助隐喻等方式将抽象概念具象化。空间关系在认知语言学中占据了重要地位，是人类最早习得的能力之一 (Akhundov, 1986; Clark, 1973; 张敏, 1998; 赵艳芳, 2001)。Lakoff and Turner (1989)指出，作为一种意象图式隐喻（image schema metaphor），空间隐喻将具体的空间概念投射到抽象的语言结构中，这样的投射可以传递空间关系及其内在逻辑。这是空间语义形成的认知基础。[0]

具体来说，人在感知外部世界时总是对空间敏感，包括运动、方向、地点等信息。Langacker (1982)高度重视空间语义对语言形成的作用，提出空间语法（Space Grammar）。Jackendoff (1983)的主题关系假设（Thematic Relations Hypothesis, TRH）认为人类语言概念结构中的事件（event）和状态（state）都是通过空间概念化组织起来的，所有语义场内的关系都类似空间组织关系。Lakoff (1987)的形式空间化假设（SFH）与之类似，并且更进一步从空间语义的视角讨论了一些基本句型的形成。Johnson (1987)概括出了和空间语义相关的27个最重要的意象图式，认为这是人类空间范畴推理的基础。Pütz and Dirven (1996)也指出地点位置对人类概念的形成有基础性作用。很多和空间范畴相关的概念也已经成为认知语言学中的重要工具和分析手段，如Talmy (1983)的图形-背景理论、Langacker (1987)的“基体-侧面”“射体-界标”关系等。

在这些基本概念和技术手段的指导下，国内外都出现了一批有关空间语义的微观研究。国外研究中，Hawkins (1984)基于空间认知对英语介词进行了全面研究。Vandeloise (1994)系统考察了法语空间介词的句法结构和语义表征。Herskovits (1986)对英语的空间表达式进行了跨学科的调查。Svorou (1993)则从认知普遍性的角度对空间介词进行了跨语言比较研究。国内研究中，廖秋忠 (1986)引进参照点的概念研究方位词，突破了传统语法对方位词静态研究的局限。刘宁生 (1994)讨论了汉语如何选择空间方位的参照物、目的物和方位词，从而表达物体的空间关系。齐沪扬 (1998)建立了现代汉语空间系统的理论框架，极大地开拓了汉语空间语义研究的视野。

汉语方位词“上”“下”始终是微观研究的重点。崔希亮 (2000)对“在X上”进行解析，显示了其空间语义及其心理延伸。蓝纯 (2003)比较了汉语的“上”“下”和英语的up和down，指出两种语言中相似方位词存在不同空间语义。白丽芳 (2006)更加深入地考察了汉语的“上”和“下”，指出两者在同一语言系统中也有不对称，“上”比“下”更基本、语义更丰富。徐丹 (2008)考察了汉语时空表达的语言特点，认为和其他语言相比，汉语特有的采用“上”“下”这样的纵向结构表示时间，体现了汉人独特的认知观念。

[0] “空间语义”和“空间范畴”是一对内涵和外延都类似的概念，本体研究中一般不作区分。如无特殊必要，下文也不区分两者。

除了关于方位词的静态研究，有关位移动词和位移事件的动态研究也是国内空间语义研究的热门话题。陆俭明 (2002)最先明确界定位移动词，认为这类动词“含有向着说话者或离开说话者位移的语义特征”，这个概念蕴含了很强的空间语义观。此后的研究中，动词的位移性逐渐作为动词的一种语义特征使用，和语义特征分析法或构式研究结合在一起，如张国宪 (2006)对表状态“在+处所+V状”和“V状+在+处所”两类构式的研究，雍茜 (2013)对三类“在+L”构式的研究，曾传禄 (2014)对“V起来”“V得（不）过来/过去”等结构的语义分化研究等。

近来对空间语义的研究，则主要集中于汉（语）方（言）比较、汉（语）外（语）比较、空间语义的认知和心理现实性等问题，如贾红霞 (2009)、尹蔚彬 (2014)、李云兵 (2016; 2020)、祝克懿 (2018)等。可以说，对空间语义的研究在国内已经步入了一个新的阶段，跨学科和实证研究不断涌现，对包括汉语在内的人类语言空间语义的认识也在不断提高。

2.2 自然语言处理领域的空间语义评测研究

自然语言处理领域的空间语义评测研究主要关注如何从自然语言中提取和理解与物理空间相关的信息。深度学习方法出现之前，自然语言处理中的空间语义任务大致经历了以下阶段：阶段一主要关注空间语义网络的层级和关系定义 (Tappan, 2004)，通过明确空间实体之间的层级关系和语义连接，初步奠定了空间信息处理的基础；阶段二则侧重特定空间语义任务，如空间实体识别 (Kordjamshidi et al., 2011)、空间关系判定等，使用机器学习方法在特定数据集上进行的半监督或无监督的训练。

然而，此前的研究大多采用非语言的形式化方法，没有充分考虑人类在自然语言中表达空间关系的方式，因此并不能有效理解自然语言中的抽象空间概念 (Stock, 1998; Renz and Nebel, 2007; Bateman et al., 2007)。例如，人类在交流空间信息时，自然语言片段中存在着不确定性和模糊性，通过构建与空间相关的知识分类，如拓扑关系和度量关系 (Tappan, 2004)等类别的方式发挥的效果有限。

为了解决自然语言中空间表述的模糊性问题，尤其是空间介词的语义识别，Kordjamshidi 等 (2011)提出了空间角色标注（Space Role Labelling，SpRL）任务，Roberts (2012)使用联合方法来识别和分类空间角色，首先从训练和测试数据中使用CRF模型提取特征，捕捉词语之间的依赖关系，并使用最大熵和朴素贝叶斯分类器来消除介词含义的歧义。同时，利用SemEval-2007 (Litkowski and Hargraves, 2007)的介词项目（TPP）的注释数据来学习介词的空间意义，然后在识别轨迹和地标角色的空间角色标签器中使用介词消歧的结果，从而实现介词的空间或非空间意义的二元分类。该方法能够同时考虑空间关系中的所有元素（如轨迹物、地标和指示物），允许使用基于整个关系的特征集 (Roberts and Harabagiu, 2012)，实现了空间关系介词消歧。

SpaceEval 2013 (Kolomiyets et al., 2013)扩展了SpRL 任务，在静态空间关系识别之外引入了运动关系（Movelink）和运动标签，用于注释运动动词或名词性运动事件及其类别并从空间语义的角度来分类事件。SpaceEval 2015 (Pustejovsky et al., 2015)则通过设定空间元素识别和分类任务、运动信号识别、运动关系识别等子任务，全面评估系统在识别和分类各种空间概念及其关系中的表现。

在中文方面，SpaCE空间语义测评借鉴了上述成果，构建了一系列高质量评测数据集，为机器的空间语义理解提出了更高要求 (詹卫东et al., 2022; 岳朋雪et al., 2023)。任务设置上，要求模型不仅能在富含空间信息的语料中执行识别和分类任务，还要进行方位推理和异形同义识别等多层次的空间语义理解。此外，在大语言模型掌握世界知识并“涌现”出空间语义识别和规划能力后，针对空间语义理解任务的设定必然更加复杂且全面。

3 数据集

本研究的数据集涵盖五大任务类别和两种选择题形式，共有九个小类题目，包括报刊、文学作品、中小学课本等一般领域与交通事故、体育动作、地理百科等专业领域，旨在考察模型在实体识别、角色识别、异常判断、方位推理和语义识别五个维度的空间语义理解能力。表1分别展示了训练集、验证集和测试集的数据情况，每一条数据都包含了题目编号、文本、选项和答案，评测采用选择题形式，题目选项设置为4个。在数据分布上，空间方位信息推理题目最多，空间异形同义识别题目最少，题型以单选题为主。总的来看，数据集的分布差异呈现了题目的多样性和复杂性，给本次评测也带来了一定的挑战。

序号	任务类别	任务要求	题型	训练集	验证集	测试集
1	空间信息实体识别	选出文本空间信息的参照物	单选题	937	226	489
			多选题	161	24	81
2	空间信息角色识别	选出文本空间信息的语义角色，或者选出与语义角色相对应的空间表达形式	单选题	1074	186	746
			多选题	19	4	24
3	空间信息异常识别	从四个选项中选出文本空间信息异常的语言表达	单选题	1077	40	500
4	空间方位信息推理	基于文本给出的推理条件进行空间方位推理，从四个选项中选出推理结果	单选题	909	468	1509
			多选题	301	207	531
5	空间异形同义识别	从四个选项中选出能使两个文本异形同义或异义的空间义词语	单选题	4	44	517
			多选题	1	11	133
总计				**4483**	**1210**	**4530**

Table 1: SPaCE 2024数据集概况

4 实验过程

4.1 模型一览

如表2所示，本研究选取了来自OpenAI、智谱华章、阿里巴巴、百度和深度求索的六个具有代表性的模型，涵盖了不同模型架构和规模。这些模型的参数规模从720亿到2360亿不等，支持的上下文长度从3.2万到12.8万不等，均是在2024年本研究开展期间的最新或较新版本[1]。

模型	版本日期	开发者	模型大小	上下文	词表大小	是否开源	调用方式
GPT-4 Turbo	04-09	OpenAI	未披露	12.8万	10万	否	API
GPT-4o	05-13	OpenAI	未披露	12.8万	20万	否	API
GLM-4	未披露	智谱华章	未披露	12.8万	未披露	否	API
ERNIE-4	03-29	百度	未披露	8千	未披露	否	API
Qwen1.5-72B-chat	未披露	阿里巴巴	720亿	3.2万	15万	是	API
Deepseek-V2-chat	未披露	深度求索	2360亿	3.2万	10万	是	API

Table 2: 模型一览

4.2 提示词工程

本研究的提示词均采用Markdown格式的结构化格式，主要包含提示词策略、提示样本构建两个部分。

在提示词策略上，分别采用普通提示（Vanilla Prompt）、工作流（Workflow）、思维链（Chain of Thought, CoT）三种方式构建提示词。在提示样本构建上，本研究的普通提示词和工作流提示词都采用0-shot、1-shot、3-shot，思维链采用1-shot。对于思维链提示词，我们参考了Wei (2022)的提示词，将其改为“想法”和“答案”两部分，让输出更为结构化，从而方便思维链和答案的提取。

关于样本的选取，训练集每条数据都有一个文本（C）、一个问题（Q）、四个选项（O）和一个答案（A）。对于每条数据，我们将其组织为一个样本$S_i = \{C_i, Q_i, O_i\}$。随后，使用Sentence-BERT (Reimers and Gurevych, 2019)将这些样本转换为向量。接下来，针对每个任务类别，我们计算了所有样本向量的平均值，作为该类别的簇心。最后，通过计算每个样本向

[1]本研究开展日期为2024年5月1日至5月17日。

量与簇心的语义相似度，分别找出距离簇心最近的1个和3个样本，作为1-shot和3-shot的训练数据。

在思维链中，样本示例需要有思考过程，因此我们用GPT-4撰写了样本的思维链过程。此外，由于异形同义识别任务的训练集只有1道多选题，我们人工将异形同义识别任务的其中2道单选题改编为多选题，以确保能够构建3-shot。

普通提示词示例

#Goal: 从四个选项中选出文本中的空间信息参照物。注意，只需回答option的一个key，不需要回答value，不需要解释。
*Text:** <text>
*Question:** <question>
*Option:** <option>
*Answer:**

工作流提示词示例

#Role: 你是一位擅长空间信息实体识别的专家。
#Goal: 从四个选项中选出文本中的空间信息参照物。注意，只需回答option的一个key，不需要回答value，不需要解释。
#Workflow: 1.阅读text：细致阅读提供的text，特别关注其中的空间信息描述。2.分析option：查看所有option，识别哪些可能是text中的空间参照物。3.选择正确option：对比text与option，选择最匹配的空间信息参照物。
*Text:** <text>
*Question:** <question>
*Option:** <option>
*Answer:**

思维链提示词示例

#Goal: 从四个选项中选出文本中的空间信息参照物。注意，只需回答option的一个key，不需要回答value，写出Thought和Answer。
*Text:** <text>
*Question:** <question>
*Option:** <option>
*Thought:** <thought>
*Answer:**

4.3 实验设置

针对不同的**提示词输出内容**，我们采用了不同的答案提取方法以优化提取过程。在普通提示和工作流提示中，提示词要求模型直接输出选项，选项之间用英文逗号“,”隔开，以便后续转换为列表格式进行评估。然而，模型有时输出答案后可能继续输出其他内容。对此，我们首先将其转换为列表，接着遍历每个元素，提取每个元素中的首字符。在思维链方法中，提示词要求模型先输出思路，再输出答案。我们使用正则表达式来提取答案，所使用的正则表达式为`<**Answer:**\n(.+?)(\n\n|$)>`。由于不同模型的指令遵循能力存在差异，我们还会自动检查每个答案是否都为A、B、C、D 四个选项之一，如不符合，还需人工检查。

关于**模型输出结果及其评测**，本研究将`temperature` 设为0.1，以确保模型输出结果的稳定性。评测指标采用准确率（Accuracy），即模型答对的题目数量占所有题目的百分比。模型答对为1分，其他情况为0分。其他情况包括：模型认为选项都不符合要求，模型拒绝回答问题，或在多选题中未能全部答对。

5 结果

5.1 模型总体表现

表 3和表 4分别是模型验证集、测试集的总体表现，满分为100。ERNIE-4借助1个样本的普通提示得到53.88%，为本研究评测的最高分，而GLM-4在1个样本的工作流提示词得到了第二高分53.14%。最终测试集使用ERNIE-4和GLM-4进行预测，ERNIE-4达到了最高的准确率，为56.20%。

根据所有模型的表现情况，我们总结归纳以下结论：（1）大模型基座能力具有举足轻重的作用。大模型的表现并不是一成不变，但比如ERNIE-4、GLM-4等模型拥有较强的中文语义理解基座能力，能够很好地适应多种有挑战性的任务。（2）提示词的数量对模型结果有重要影响。单样本可以显著提升模型的空间语义理解能力，但相较于0-shot，1/3-shot可以显著提升模型的空间语义理解能力，但从1到3-shot，准确率升降不定（7例上升，5例下降）。（3）提示词策略不一定越复杂越好，简单的提示词策略可能也有出色的效果。思维链可以帮助模型更好地理解语义空间，但在此次空间语义测评中表现并不突出。

模型	普通			工作流			思维链
	0	1	3	0	1	3	1
ERNIE-4	50.25	**53.88**	52.73	**52.23**	52.73	**52.81**	**51.06**
GLM-4	**51.24**	52.01	52.23	50.49	**53.14**	50.41	50.82
GPT-4o	48.92	51.16	**52.89**	48.35	50.99	51.73	50.91
GPT-4 Turbo	48.18	50.99	51.54	47.43	51.49	47.77	50.74
Deepseek-V2-chat	48.84	49.83	49.98	46.69	49.42	49.83	46.78
Qwen1.5-72B-chat	44.71	46.61	46.45	42.81	45.70	45.04	45.45

Table 3: 模型在验证集的总体表现

模型	样本数量	提示词	测试集准确率
ERNIE-4	1	普通	**56.20**
GLM-4	1	工作流	54.52

Table 4: 模型在测试集的最终表现

5.2 模型具体表现

大模型使用某个提示词策略达到最佳结果，但不等于在该提示词策略下各个方面都表现优秀。基于此，本研究进一步探究五个模型的具体表现，涵盖实体识别、角色识别、异常识别、空间推理、同义词识别、单选题和多选题7个维度。为了展示模型的实际表现和最大潜力，表 5列出了每个模型在验证集中的实际最佳性能和潜在最佳性能。其中，实际最佳性能是指某一维度在最优提示词下的得分，潜在最佳性能是该维度在不同提示词中的最高得分。

整体上看，所有模型在角色识别任务的表现最优，在空间推理任务的表现相对最差，单选题得分普遍高于多选题。在实际最佳性能方面，ERNIE-4表现最好，在4个任务和单选题表现优异，GLM-4次之，在角色识别任务和多选题的得分最高。在潜在最佳性能方面，ERNIE-4和GLM-4依然保持稳定，模型可以在实体识别和角色识别这两个任务发挥更大潜力。此外，Deepseek-V2-chat在同义识别任务的潜在表现尤其值得关注，可以达到与ERNIE-4、GLM-4比肩的水平。

下列以ERNIE-4的实际最佳性能的结果为例，探究其在验证集不同任务类别的表现，更为细致地了解模型空间语义理解的特点。

5.2.1 在实体识别题目的表现

实体识别类题目考察模型能否识别空间方位词和语境中已经出现过实体的同指关系。由于这类空间方位词和实体的关系在语境中是固定的，模型比较容易从语境中学到这个知识，其对实体的空间语义关系理解比较准确（如1题和4题）。

		实体识别	角色识别	异常识别	空间推理	同义识别	单选题	多选题
ERNIE-4	实际最佳	**79.20**	95.26	**87.50**	29.92	**65.45**	**61.20**	25.20
	潜在最佳	**80.40**	**96.84**	**87.50**	29.92	**65.45**	**61.20**	25.20
GLM-4	实际最佳	78.40	**95.79**	85.00	29.33	60.00	58.30	**32.93**
	潜在最佳	78.40	**96.84**	85.00	29.33	63.63	59.64	**32.93**
GPT-4o	实际最佳	76.40	93.68	80.00	**30.52**	60.00	58.09	32.52
	潜在最佳	76.40	95.79	80.00	**30.52**	65.45	59.34	33.74
GPT-4 Turbo	实际最佳	76.80	95.26	72.50	28.59	54.54	59.54	20.73
	潜在最佳	76.80	95.26	80.00	29.48	61.82	59.92	23.17
Deepseek-V2-chat	实际最佳	74.40	95.26	77.50	26.22	52.73	56.33	24.80
	潜在最佳	74.40	96.84	82.50	29.04	**65.45**	56.74	29.67
Qwen1.5-72B-chat	实际最佳	71.60	91.05	67.50	23.11	52.73	55.50	11.79
	潜在最佳	72.40	93.68	67.50	24.74	54.54	55.50	16.67

Table 5: 模型在验证集的实际最佳性能和潜在最佳性能

1题：周游口袋里只有五元钱……所以蹬三轮车的上来拉生意时，他理都不理他们，而是从西装口袋里掏出个玩具手机，这个玩具手机像真的一样，里面装上一节五号电池，悄悄按上一个键，手机的铃声就会响起来。（题目：___里面装上一节五号电池）

4题：回家以后，她给丈夫算了一笔账：我每天上下班路程要花3个小时，工作8小时，中午吃饭1小时，总共在外边花12小时……（题目：总共在___外边花12小时）

值得一提的是，当需要判断的实体在语境中和其他实体有领属或广义领属关系，尤其是其他实体没有在语境中出现时，ERNIE-4对实体选择的错误率较高。58题没有出现实体“屋子”，67题存在干扰项“大厅”（实际应该选“楼上”，两者有广义领属关系），模型对两句话的判断都出现了问题。

58题：爸爸把我从床头打到床角，从床上打到床下，外面的雨声混合着我的哭声……（题目：___外面的雨声混合着我的哭声）

67题：楼上只有南面的大厅有灯亮。灯亮里有块白长布，写着点什么——林乃久知道写的是什么。其余的三面黑洞洞的……（题目：___三面黑洞洞的）

总的来说，ERNIE-4可以比较准确地判断语境中出现过的单独和方位词关联的实体，但对语境中没有出现的、以及语境中还存在和该实体有领属关系的干扰项时，模型的判断能力比较弱。

5.2.2 在角色识别题目的表现

角色识别类题目考察模型能否识别存在空间交互关系的两个实体。两个实体的空间交互关系有多种来源，包括领属关系带来的空间关系（如251题）、事件带来的空间关系（如第258题）、相对位置带来的空间关系（如259题）等。ERNIE-4对这些关系的认识都非常准确。

251题：时间过去近两个月，木沙江·努尔墩仍清楚地记得……在人工湖边的冰窟中，拉齐尼用一只手臂搂住孩子，另一只手努力托举着孩子……（题目：___的手努力托举着孩子）

258题：……正在站上值班的牛红生例行巡检，走到龙王路段时，发现一辆轿车从百米外的公路上猛然栽进路边的排水渠……牛红生只能在水中摸索，摸到车门后用力拽开，把人拉出水面……（题目：牛红生把___拉出了水面）

259题：文本同258题（题目：轿车栽进去时的初始位置是___）

当题目直接询问实体（包括抽象实体）的复杂空间关系（包括隐域空间关系）时，模型的识别能力下降，出现了较多错误。这类空间关系要么是隐涵的，要么是“元语言”意义上的，不容易从语境中直接得到。这里说的“元语言”，指题目选项中的抽象空间关系表达，如398题的四个选项为“路径”“方向”“起点”和“外部位置”，是抽象的描述方位的语言，和具体题干无关，这和“名词”“动词”“主语”“宾语”等“描述语言”的“元语言”不同。感谢审稿专家指出这一点。

398题：首先是水的气味，宽广的昌江流经鄱阳城奔向鄱阳湖，在城外留下韭菜湖、青山湖、土湖、东湖、球场湖五片湖……（题目：“鄱阳城”属于“昌江”流动时的___信息）

425题：几天以后李光头回来了，他在上海买了一辆红色的桑塔纳轿车，他有专车了……李光头从车里出来时，身穿一身黑色的意大利阿玛尼西装，那身破烂衣服扔在上海的垃圾桶里了……（题目：“破烂衣服扔在上海的垃圾桶里”发生在___）

总的来说，ERNIE-4判断两个具体实体的具体空间关系是最准确的，判断抽象实体的具体空间关系次之，判断抽象实体的抽象空间关系是最弱的。这和人类一般的认知能力相似，越是具体的对象和关系就越容易认知和识别。

5.2.3 在异常识别题目的表现

异常识别类题目考察模型能否识别存在异常或错误空间交互关系的若干实体。异常的空间关系要么是不合常理的（如441题），要么是自相矛盾的（如442题）。

441题：小红在下，我在上，走到四楼的东侧.....（题目：异常的空间方位信息是___，要求识别出“小红在下，我在上”）

442题：灵车缓缓地前进，牵动着千万人的心.....人们多么希望车子能停下来，希望时间能停下来！可是灵车渐渐地靠近了，最后消失在苍茫的夜色中了.....（题目：异常的空间方位信息是___，要求识别出“灵车渐渐地靠近并消失在夜色中”）

ERNIE-4的异常空间识别能力较好，但如果异常空间关系复杂，ERNIE-4就不容易将其识别出来。例如在478题中，由东向西行驶的车子左转弯是向南而不是向北。这种空间关系并不直观，模型在识别上也出现了问题。对这种异常空间关系的识别通常需要另外的百科知识或一般推理能力。

478题：经审理查明.....小型客车沿本区亭林镇红梓路由东向西行驶至车亭公路路口时，遇绿灯向北实施左转弯途中.....（题目：异常的空间方位信息是___，要求识别出“小型客车由东向西行驶至车亭公路路口向北左转弯”）

总的来说，ERNIE-4基本可以准确识别相对直观的异常空间关系，但在需要调用推理能力或百科知识判断空间关系是否正常时，模型的表现并不是很好。

5.2.4 在空间推理题目的表现

空间推理类题目，考察模型能否通过简单的推理方式得到正确的实体间空间关系。这类问题中，模型只能在语境里得到条件句命题的前件，后件需要根据实际问题的需要自行推理。

481题：贺知章、李白、陈子昂、骆宾王、王维、孟浩然六个人在海边沙滩上围成一圈坐着，大家都面朝中心的篝火。六人的位置恰好形成一个正六边形。任意相邻两个人之间的距离相等，大约为一米。已知：陈子昂在骆宾王左边数起第1个位置，孟浩然在陈子昂逆时针方向的第5个位置，王维在孟浩然顺时针方向的第1个位置（题目：孟浩然在___的斜对面）

尽管该推理问题并不是很复杂，但ERNIE-4体现出的空间推理能力不是很强。这种推理能力不仅需要正确得到语境中的信息，同时还要调用必要的百科知识（如481题中对“正六边形”的理解等），然后根据这些信息展开推理。空间推理可以理解为复杂的角色识别和实体识别问题，这意味着模型在处理连续实体识别问题上还存在一些问题，无法正确判断这种需要推理和叠加的复杂空间关系。

5.2.5 在同义识别题目的表现

同义识别类题目考察模型对不同空间方位词表达的具体空间关系的认识是否准确。汉语存在一批空间方位词，它们单独使用的语义不同，但和某些空间实体结合时可以表达相同的空间方向。例如在1157题，“回来”和“回去”是两个不同的词，甚至有时被认为是一对反义词，但此时发生替换后空间场景没有明显改变。

1157题：傍晚的时候，宋钢将他带回去的钱用一张旧报纸仔细包好了，放在了枕头下面.....（题目：“回去”替换为___形成的新句可以与原句表达相同的空间场景，要求用“回来”替换“回去”）

由于这类空间方位词单独使用的语义不同，如果要正确替换，模型必须将方位词和与之关联的实体联系在一起替换，这考察了模型对“实体+方位词”序列语义的认识能力，而不是仅仅从语境中找到方位词关联的实体，后者是实体识别的任务。从同义识别题目和实体识别题目的对比表现上看，ERNIE-4可以比较好地找到方位词和方位词关联的实体，但是在其语义的理解上表现得不够出色。

5.2.6 在不同任务类型的任务表现

综合上文分析，虽然测试题目在不同类型空间关系的考察上有所差异，但从ERNIE-4在现有题目的平均表现上还是可以观察到其空间语义理解的特点，具体如表 6。

6 结语

在CCL2024年中文空间语义评测中，我们观察到了不同大模型的空间语义理解能力。本队

任务类别	模型表现	影响因素
角色识别	好	具体实体的具体空间关系不受外界影响，但不容易判断抽象实体（元语言对象）的抽象空间关系
实体识别	较好	表示静态的空间关系，基本不受外界影响，但出现与目标实体有领属或广义领属关系的其他实体时容易受干扰
异常识别	较好	简单异常空间关系容易识别，但在需要百科知识或推理能力的问题上易受干扰
同义识别	较差	表示空间关系的联系，受“实体+方位词”语义的影响
空间推理	差	参与空间主体较多，而且需要百科知识和推理能力，易受干扰项影响

Table 6: 模型表现及其影响因素

伍提交的系统在封闭赛道中排名第6，测试集上的准确率得分为56.20。综合来看，为了提升模型的空间语义理解能力，可以深入优化模型的提示词处理机制，或者设计更具结构化和明确性的提示词。

此外，进一步提升大模型的基座能力，也是未来模型发展的重要方向。这包括对模型架构的优化，如通过更多的数据和更先进的训练算法提升模型的表现；同时，结合外部知识库和信息源，使模型能够在更广泛的知识背景下进行推理和生成。通过结合知识增强方法和交互式决策策略，可以显著提高模型的实际应用效果。

参考文献

Murad D. Akhundov. 1986. *Conceptions of Space and Time*. The MIT Press, Cambridge, MA.

John Bateman, Thora Tenbrink and Scott Farrar. 2007. The role of conceptual and linguistic ontologies in interpreting spatial discourse. *Discourse Processes*, 44(3): 175-212.

Emily M. Bender, Timnit Gebrum, Angelina McMillan-Major and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610-623.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry and Amanda Askell et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877-1901.

Herbert H. Clark. 1973. *Space, Time, Semantics and the Child: Cognitive Development and the Acquisition of Language*. Academic Press, New York.

Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, volume 1, pages 4171-4186.

Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30: 681-694.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Eric Hawkins. 1984. *Awareness of Language: An Introduction*. Cambridge University Press, Cambridge.

Annette Herskovits. 1986. *Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English*. Cambridge University Press, Cambridge.

Ray S. Jackendoff. 1983. *Semantics and Cognition*. The MIT Press, Cambridge, MA.

Mark Johnson. 1987. *The Body in the Mind: The Bodily Basis of Meaning, Imagination and Reason*. The University of Chicago Press, Chicago.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, Jeffrey and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Oleksandr Kolomiyets, Parisa Kordjamshidi, Marie Francine Moens and Steven Bethard. 2013. Semeval-2013 task 3: Spatial role labeling. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 255-262.

Parisa Kordjamshidi, Martijn Van Otterlo and Marie-Francine Moens. 2011. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing (TSLP)*, 8(3): 1-36.

George Lakoff. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. The University of Chicago Press, Chicago.

George Lakoff and Mark Turner. 1989. *More Than Cool Reason: A Field Guide to Poetic Metaphor*. The University of Chicago Press, Chicago.

Ronald W. Langacker. 1982. Space grammar, analysability, and the English passive. *Language*, 22–80. JSTOR.

Ronald W. Langacker. 1987. *Foundations of Cognitive Grammar I: Theoretical Prerequisites*. Stanford University Press, Stanford.

Kenneth C. Litkowski and Orin Hargraves. 2007. SemEval-2007 Task 06: Word-sense disambiguation of prepositions. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 24-29.

Erika Petersen and Christopher Potts. 2023. Lexical Semantics with Large Language Models: A Case Study of English "break". In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 490-511.

Steven Piantadosi. 2023. Modern language models refute Chomsky's approach to language. *Lingbuzz Preprint, lingbuzz*, 7180.

Pustejovsky, J., Kordjamshidi, P., Moens, M. F., et al. 2015. Semeval-2015 task 8: Spaceeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 884-894.

Martin Pütz and René Dirven. 1996. *The Construal of Space in Language and Thought*. de Gruyter Mouton, Berlin.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Jochen Renz and Bernhard Nebel. 2007. Qualitative spatial reasoning using constraint calculi. In *Handbook of Spatial Logics*, pages 161-215. Dordrecht: Springer Netherlands.

Kirk Roberts and Sanda Harabagiu. 2012. UTD-SpRL: A joint approach to spatial role labeling. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 419-424.

Oliviero Stock. 1998. *Spatial and Temporal Reasoning*. Springer Science & Business Media.

Soteria Svorou. 1993. *The Grammar of Space*. John Benjamins, Amsterdam.

Leonard Talmy. 1983. How language structures space. In H. L. Pick and L. P. Acredolo, editors, *Spatial Orientation*, pages 225-282. Springer, Boston, MA.

Daniel Allen Tappan. 2004. Knowledge-based spatial reasoning for automated scene generation from text descriptions. New Mexico State University.

Lindia Tjuatja, Emmy Liu, Lori Levin and Graham Neubig. 2023. Syntax and Semantics Meet in the "Middle": Probing the Syntax-Semantics Interface of LMs Through Agentivity. *arXiv preprint arXiv:2305.18185*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Llion, Aiden N. Gomez, Łukasz Kaiser and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998-6008.

Claude Vandeloise. 1994. Methodology and analyses of the preposition in. *Cognitive Linguistics*, 2: 157-184.

Wei J, Wang X, Schuurmans D, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824-24837.

Zhao Wayne Xin, Zhou Kun, Li Junyi, Tang Tianyi, Wang Xiaolei, Hou Yupeng, Min Yingqian, Zhang Beichen, Zhang Junjie and Dong Zican et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

白丽芳. 2006. “名词+上/下”语义结构的对称与不对称. 语言教学与研究, 4: 58-65.

崔希亮. 2000. 空间方位关系及其泛化形式的认知解释. 载《语法研究和探索（十）》, pages 85-97. 北京: 商务印书馆.

贾红霞. 2009. 普通话儿童空间范畴表达发展的个案研究. 博士学位论文. 中国社会科学院研究生院.

蓝纯. 2003. 从认知角度看汉语和英语的空间隐喻. 外语教学与研究出版社, 北京.

李云兵. 2016. 论苗语空间范畴的认知. 民族语文, 3: 8-34.

李云兵. 2020. 朱坝羌语静态空间范畴的表征与认知. 民族语文, 5: 3-20.

廖秋忠. 1986. 现代汉语篇章中的空间和时间的参考点. 载《廖秋忠文集》, pages 3-15. 北京: 北京语言学院出版社.

刘宁生. 1994. 汉语怎样表达物体的空间关系. 中国语文, 3: 169-179.

陆俭明. 2002. 动词后趋向补语和宾语的位置问题. 世界汉语教学, 1: 5-17+114.

齐沪扬. 1998. 现代汉语空间问题研究. 学林出版社, 上海.

徐丹. 2008. 从认知角度看汉语的两对空间词. 中国语文, 6: 504-510+575.

尹蔚彬. 2014. 拉坞戎语的空间范畴. 语言科学, 3: 268-280.

雍茜. 2013. “在+L”类构式与动词的语义整合. 硕士学位论文. 上海师范大学.

岳朋雪, 王诚文, 孙春晖, 詹卫东and 穗志方. 2023. 中文空间语义理解评测数据集质量评估研究. 语言文字应用, (01): 101-113. DOI: 10.16499/j.cnki.1003-5397.2023.01.006.

曾传禄. 2014. 现代汉语位移空间的认知研究. 商务印书馆, 北京.

张国宪. 2006. 现代汉语形容词功能与认知研究. 商务印书馆, 北京.

詹卫东, 孙春晖, 岳朋雪, 唐乾桐and 秦梓巍. 2022. 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. 语言文字应用, (02): 99-110. DOI: 10.16499/j.cnki.1003-5397.2022.02.005.

张敏. 1998. 认知语言学与汉语名词短语. 中国社会科学出版社, 北京.

赵艳芳. 2001. 认知语言学概论. 上海外语教育出版社, 上海.

祝克懿. 2018. 心理空间范畴与语言生成机制.