

CCL24-Eval 任务3系统报告：基于上下文学习与思维链策略的中文空间语义理解

王士权，付薇薇，方瑞玉，李孟祥，何忠江，李永翔，宋双永
中国电信人工智能研究院

{wangsq23, fuweiwei, fangry, hezj, liyx25, songshy}@chinatelecom.cn
limengx@126.com

摘要

本技术报告详细介绍了我们团队参加第四届中文空间语义理解评测（SpaCE2024）的方法和成果。SpaCE2024旨在全面测试机器对中文空间语义的理解能力，包括空间信息实体识别、空间信息实体识别、空间信息异常识别、空间方位信息推理和空间异形同义识别五个不同的任务。我们团队采用精心设计的prompt并结合微调的方式激发大语言模型的空间语义理解能力，构建了一个高效的空间语义理解系统。在最终的评估中，我们在空间信息实体识别题目中准确率为0.8947，在空间信息实体识别题目中准确率为0.9364，在空间信息异常识别题目中准确率为0.8480，在空间方位信息推理题目中准确率为0.3471，在空间异形同义识别题目中准确率为0.5631，测试集综合准确率为0.6024，排名第一。

关键词： 空间语义理解；大语言模型

System Report for CCL24-Eval Task 3: Chinese Spatial Semantic Understanding Based on In-Context Learning and Chain of Thought Strategy

Shiquan Wang, Weiwei Fu, Ruiyu Fang, Mengxiang Li, Zhongjiang He,
Yongxiang Li, Shuangyong Song

Institute of Artificial Intelligence (TeleAI), China Telecom Corp Ltd
{wangsq23, fuweiwei, fangry, hezj, liyx25, songshy}@chinatelecom.cn
limengx@126.com

Abstract

This technical report provides a detailed introduction to the methods and achievements of our team in the Fourth Chinese Spatial Semantic Understanding Evaluation (SpaCE2024). The SpaCE2024 aims to comprehensively test a machine's ability to understand Chinese spatial semantics across five different tasks: spatial information entity recognition, spatial information entity disambiguation, spatial information anomaly detection, spatial orientation reasoning, and spatial heteronym synonym recognition. Our team employed meticulously designed prompts combined with fine-tuning to enhance the spatial semantic understanding capabilities of large language models, thereby constructing an efficient spatial semantic understanding system. In the final evaluation, our system achieved an accuracy of 0.8947 in spatial information entity recognition, 0.9364 in spatial information entity disambiguation, 0.8480 in spatial information anomaly detection, 0.3471 in spatial orientation reasoning, and 0.5631 in spatial heteronym synonym recognition. The overall accuracy on the test set was 0.6024, earning us a first-place ranking.

Keywords: Spatial semantic understanding, Large Language Model

1 引言

空间语义理解是自然语言处理（NLP）领域中一个极具挑战性的任务，它要求机器不仅要理解语言的表面形式，还要能够构建和推理出语言背后的空间关系和场景。掌握中文空间语义理解，对于提升机器翻译的准确性、增强人机交互的自然性、提高自动摘要和信息检索的相关性，以及推动智能教育、自动驾驶、室内导航等应用领域的发展具有深远的影响。此外，由于中文在表达空间关系时具有独特的语言现象和语法结构，中文空间语义理解的研究还有助于丰富跨语言的NLP理论和技术，促进人工智能对人类语言和认知过程的深入理解。

第四届中文空间语义理解评测（SpaCE2024）更加注重针对大语言模型的空间语义理解能力的测试，宗旨是在一个测试数据集上考察机器中文空间语义理解的综合能力。为此，SpaCE2024 将不再划分子任务，而是以选择题的形式考察五个层次的空间语义理解能力，分别是空间信息实体识别、空间信息角色识别、空间信息异常识别、空间方位信息推理、空间异形同义识别。

在本次评测任务中，我们首先使用精心设计的提示（prompt）获取模型在选择答案时的推理过程。随后，我们结合上下文学习与思维链（chain of thought）的方法，构建推理提示以提高模型在空间语义理解题目上的表现。其次，我们基于上下文学习的方法对模型进行微调，以进一步增强其在该任务中的性能。最后，通过投票策略整合多种模型的输出，获取最优结果。我们的系统在最终的线上评测的空间信息实体识别题目中准确率为0.8947，在空间信息实体识别题目中准确率为0.9364，在空间信息异常识别题目中准确率为0.8480，在空间方位信息推理题目中准确率为0.3471，在空间异形同义识别题目中准确率为0.5631，测试集综合准确率为0.6024，综合排名第一。

2 相关工作

空间语义理解作为自然语言处理领域的核心任务之一，对于实现类人智能具有重要意义。人类通过空间范畴来组织和理解周围的世界，而在自然语言文本中，空间信息的表达和理解尤为关键。随着自然语言处理技术的不断进步，对机器的空间语义理解能力进行评测变得尤为重要。为了更好的评判模型对于空间语义的理解能力，Kordjamshidi等(Kordjamshidi et al., 2012)在SemEval-2012举行了基于静态空间关系的空间语义角色标注任务，该任务涉及从自然语言中提取空间语义的主要组成部分：轨迹、地标和空间指示器。除了这些主要组成部分外，还针对它们之间的联系以及空间关系的一般类型，包括区域、方向和距离进行了标注。在SemEval-2013中，Kordjamshidi等(Kolomiyets et al., 2013)对Spatial Role Labeling任务进行了完善。Pustejovsky等人(Pustejovsky et al., 2015)在SemEval-2015提出在SpRL任务中使用ISO-Space标注体系，进一步完善空间语义角色标注任务。

SpaCE系列赛事在探索中文空间语义理解评测上做出了持久的贡献，SpaCE2021(詹卫东 et al., 2022)分为三个中文空间语义正误判断、中文空间语义异常归因合理性判断与中文空间语义判断与归因联合任务三个子任务，分别考察模型能否正确区分正常与错误的空间语义表达，模型能否解释空间语义表达错误的原因，模型处理上述两个任务的综合能力。SpaCE2022(Liming et al., 2023)在SpaCE2021的基础上扩大了数据规模，拓宽了语料类型，改进了任务形式，提高了标注质量，共分为中文空间语义正误判断、中文空间语义异常归因与异常文本识别与中文空间实体识别与空间方位关系标注三个子任务。SpaCE2023(Xiao et al., 2023)在SpaCE2022的基础上删除了空间语义正误判断任务，以及异常归因任务；保留了异常文本识别任务，以及空间语义角色标注任务，并改进了数据格式，更新了测试集，提高了标注质量；新增了生成类任务，要求机器判断两段在语言表达形式上有差异的文本是否可以描述相同的空间场景，并说明判断理由。

SpaCE2024将之前的测试题形式改为选择题，并通过设置一定比例的重复题目来测试模型的稳定性，同时提高专业领域的语料占比。SpaCE2024更加注重针对大语言模型(Wang et al., 2024)的空间语义理解能力的测试，以选择题的形式考察以下五个层次的空间语义理解能力：

(1) 空间信息实体识别。要求从四个选项中选出文本空间信息的参照物。(2) 空间信息角色识别。要求从四个选项中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。(3) 空间信息异常识别。要求从四个选项中选出文本空间信息异常的语言表达。

(4) 空间方位信息推理。要求基于文本给出的推理条件进行空间方位推理，从四个选项选出推理结果。(5) 空间异形同义识别。要求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。

3 方法

在本次评测任务中，我们首先使用精心设计的提示获取模型在选择答案时的推理过程。随后，我们结合上下文学习 (In-context Learning) 与思维链 (Chain of Thought) 的方法，构建推理提示以提高模型在空间语义理解题目上的表现。其次，我们基于上下文学习的方法对模型进行微调，以进一步增强其在该任务中的性能。最后，通过投票策略整合多个模型的输出，获取最优结果。

3.1 基于上下文学习与思维链策略的模型推理

上下文学习，是大型语言模型中一种重要的能力。ICL允许模型利用输入中的上下文信息来指导其输出，从而在没有额外训练数据的情况下提高模型的泛化能力和灵活性。在大模型中，ICL的作用尤为显著，因为这些模型通常预训练在大规模的数据集上，能够捕捉到丰富的语言模式和世界知识。当给定一个包含示例和问题的上下文时，大型模型能够快速适应新的任务或领域，展现出对上下文的敏感性和适应性。这种能力使得大模型在处理诸如阅读理解、问题解答和代码生成等任务时表现出色，因为它们可以直接从上下文中提取相关信息来生成答案或解决方案。

思维链是一种新兴的策略，用于增强大型语言模型的推理能力。CoT的核心思想是模拟人类解决问题时的思维过程，即将复杂问题分解为一系列简单的推理步骤，并清晰地展示这些步骤。在大模型中，CoT的作用体现在能够引导模型逐步构建答案，而不是仅仅依赖于模式识别或记忆匹配。这种方法特别适用于那些需要多步骤逻辑推理的任务，如数学问题解决、科学推理或编程问题调试。通过明确展示思维过程，CoT不仅提高了模型输出的可解释性，还有助于提升模型在面对新颖或复杂问题时的表现。此外，CoT还可以作为一种正则化手段，帮助模型避免在复杂任务中陷入局部最优解，而是鼓励其探索更广泛的解决方案空间。

输入：

下面是一个关于实体识别的单项选择题，请根据下面的文本、问题、选项与答案给出一步一步的推理过程。
 文本：人民大会堂和中国历史博物馆雄踞天安门广场东西两侧，广场中央耸立着人民英雄纪念碑和毛主席纪念堂，北有庄严的天安门城楼，南有巍峨的正阳门，广场面积44公顷，为世界最大广场。地处旧城中心的故宫珍藏了历代的珍奇异宝。北海、中南海、景山、天坛、中山公园、劳动人民文化宫、龙潭湖、先农坛、陶然亭、地坛等处，亭榭建筑精美、风景秀丽，是中外人士游类的集中地。
 问题：()北有庄严的天安门城楼。

选项：

- A:中国
- B:中国历史博物馆
- C:天安门广场
- D:以上选项都不是

答案：C

一步一步的推理过程：

输出：

1. 阅读文本，寻找与问题相关的信息。文本中提到“人民大会堂和中国历史博物馆雄踞天安门广场东西两侧，广场中央耸立着人民英雄纪念碑和毛主席纪念堂，北有庄严的天安门城楼”。
 2. 根据文本中的描述，可以确定天安门城楼位于广场的北侧。
 3. 问题问的是“()北有庄严的天安门城楼”，需要找到天安门城楼所在的地理位置。
 4. 从文本中提取的信息表明，天安门城楼是“北有”的主体，而这个“北有”的参照物是广场。
 5. 因此，正确答案是天安门广场的北侧有天安门城楼。
 6. 选项中C选项表示“天安门广场”，与文本中的信息一致，所以正确答案是C。
- 综上所述，答案是C

Figure 1: CoT数据生成示例

在本次评测中，我们选择Qwen1.5-72B-Chat作为直接推理时使用的模型，具体来说，我们首先依靠训练集与验证集数据得到模型在获取答案时的推理过程，prompt设置与模型的输出如图1所示，通过该步骤我们获取模型在推理答案时的思维过程。其次，在对测试集进行推理时结

合上下文学习与思维链的策略，在prompt设置中添加同类型数据示例，指导模型先对问题进行分析推理，最后总结出答案。

3.2 基于上下文学习的模型有监督微调

随着模型大小和语料库大小的扩展，大型语言模型显示出上下文学习能力，它们从上下文中的几个示例中学习，大量研究表明大型语言模型可以通过ICL有效执行各种复杂任务。形式上，给定一个查询输入文本 x 和一组候选答案 $Y = \{y_1, \dots, y_m\}$ ，预训练语言模型 M 在条件集 C 的条件下选择得分最高的候选答案作为预测。集合 C 包括一个可选的任务说明 I 和 K 个示例，因此 $C = \{I, s(x_1, y_1), \dots, s(x_k, y_k)\}$ 或 $C = \{s(x_1, y_1), \dots, s(x_k, y_k)\}$ ，其中 $s(x_k, y_k, I)$ 表示根据任务用自然语言编写的示例。候选答案 y_j 的概率可以通过模型 M 的整个输入序列的评分函数 f 来表示。

$$P(y_j|x) \triangleq f_M(y_j, C, x) \tag{1}$$

最终预测的标签 \hat{y} 是具有最高概率的候选答案：

$$y = \arg \max_{y_j \in Y} P(y_j|x) \tag{2}$$

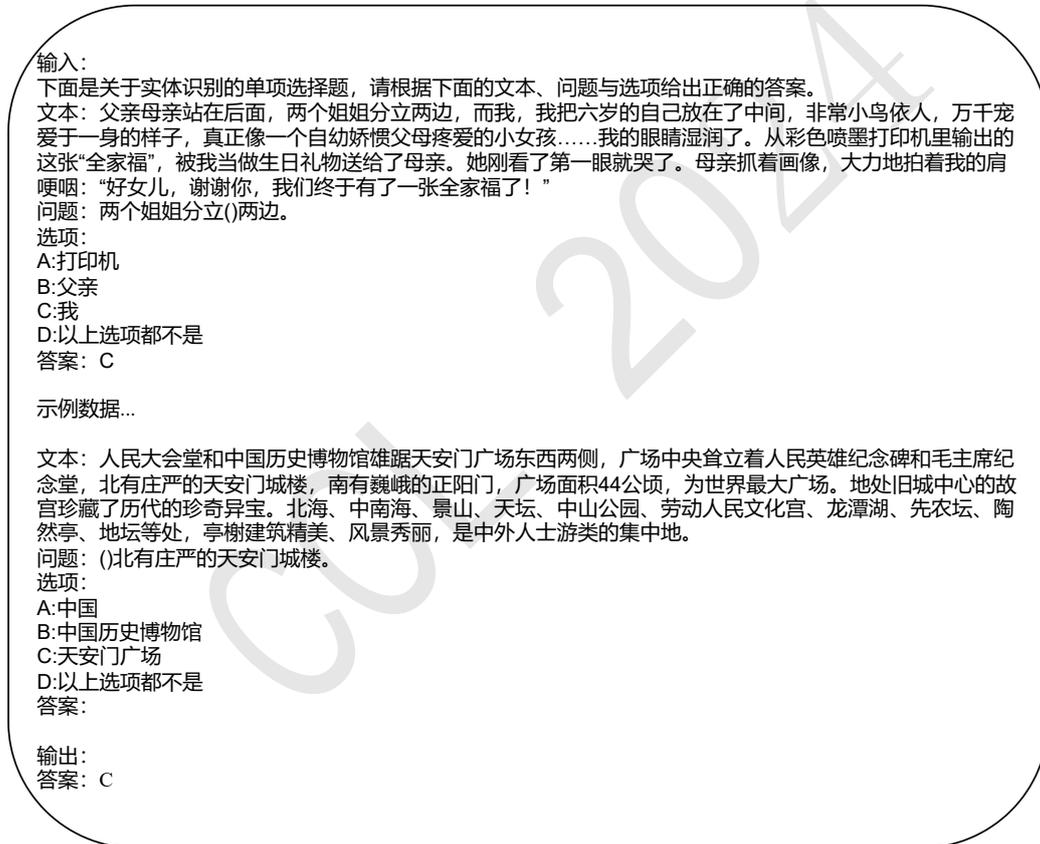


Figure 2: 上下文学习指令微调数据示例

虽然大型语言模型已经展示了强大的上下文学习能力，但一些研究也表明，通过预训练和上下文学习推理之间的持续训练阶段，这种能力可以进一步增强(Wei et al., 2023; Chen et al., 2022)。因此，我们通过构建上下文学习的指令训练数据来增强大型语言模型的上下文能力，并通过监督式指令微调消除预训练任务与下游上下文学习任务之间的差距。具体来说，我们利用随机选择方法选择同类型的10个示例数据，随后通过人工挑选的方式从中选择5个数据作为prompt中的示例数据，构造有监督的ICL训练数据，然后基于有监督的ICL数据训练Qwen1.5-7B-Chat模型，示例数据如图2所示。

3.3 模型投票

投票策略通过集成多个模型或同一模型在不同推理路径上的预测结果，可以有效减少随机误差和模型偏差，提高整体的鲁棒性。由于投票策略能够平衡不同模型的优势，特别是在面对复杂或模糊的输入时，可以综合考虑各个模型的判断，从而得出更加全面和可靠的答案。除此之外，投票机制还可以作为一种正则化手段，防止模型过拟合于特定的数据模式，增强模型在未知数据上的泛化能力。同时在微调训练过程中，我们发现不同模型，甚至同结构的模型在不同训练轮次的验证集预测结果上仍存在差异，我们使用投票策略对模型预测结果进行处理，最后取得了最优的提交成绩。

4 实验

4.1 数据介绍

SpaCE2024更加注重针对大语言模型的空间语义理解能力的测试，宗旨是在一个测试数据集上考察机器中文空间语义理解的综合能力。SpaCE2024不再划分子任务，以选择题的形式考察模型在空间信息实体识别、空间信息角色识别、空间信息异常识别、空间方位信息推理与空间异形同义识别五个层次的空间语义理解能力，数据分布如表1所示。

题目类型	Train			Dev			Test		
	单选	多选	Total	单选	多选	Total	单选	多选	Total
实体识别	937	161	1098	226	24	250	513	87	600
角色识别	1074	19	1093	186	4	190	776	24	800
异常识别	1077	10	1077	40	0	40	530	0	530
空间推理	909	301	1210	468	207	675	1533	537	2070
同义识别	4	1	5	44	11	55	541	139	680

Table 1: SpaCE24数据集数据分布

SpaCE2024 使用准确率 (Acc, Accuracy) 作为评价指标和排名依据，如公式3所示：

$$Acc = \text{命中正确答案的题数} / \text{题目总数} \quad (3)$$

除此之外还使用稳定性来衡量机器表现的稳定程度，不作为排名依据。以比赛题及其对应的干扰题为一组，如公式4所示：

$$\text{稳定性} = \text{组内答案完全一致的组数} / \text{总组数} \quad (4)$$

4.2 基于上下文学习与思维链策略的中文空间语义理解模型实验结果

我们首先基于SpaCE2024验证集进行了一系列实验尝试，实验结果如表2所示。

DataSet	Metrics			Accuracy					
	ICL	CoT	Train	Total	实体识别	角色识别	异常识别	空间推理	同义识别
SpaCE24.dev	0-shot	w/o	w	0.5570	0.8560	0.9526	0.8750	0.3170	0.5454
SpaCE24.dev	5-shot	w	w	0.4330	0.6360	0.9316	0.4250	0.2119	0.5091
SpaCE24.dev	5-shot	w	w/o	0.4240	0.6720	0.9263	0.5750	0.1719	0.5455
SpaCE24.dev	5-shot	w/o	w	0.5686	0.8440	0.9737	0.9250	0.3304	0.5818

Table 2: SpaCE24验证集的实验结果

表2展示了在SpaCE24验证集上不同策略的具体得分，其中ICL表示在训练与推理过程中使用上下文学习策略在prompt中加入任务数据示例，CoT表示在训练与推理过程中使用思维链策略在模型输出最后答案前输出推理过程，Train表示是否使用SpaCE24训练集进行模型训练，当不进行模型训练时我们选用的模型为Qwen1.5-72B-Chat，当进行模型训练时我们选用的模型为Qwen1.5-7B-Chat。实验结果表明，在训练和推理阶段使用上下文学习策略与思维链策略对模型效果有较大的提升，值得注意的是，使用思维链策略参与模型训练后会导模型效果下

降。这里我们推测原因为思维链数据是模型根据问题与答案生成的伪推理过程，中间部分步骤存在错误，在训练过程中引入错误的推理过程会影响模型本身的推理能力，抽样检查的思维链推理过程也证明了我们推测的原因。如果有高质量的思维链推理数据即完全正确的推理过程，在模型训练过程中使用思维链数据应当会提升的表现。

4.3 基于投票策略的中文空间语义理解模型实验结果

基于验证集的实验结果，我们在SpaCE2024测试集上进行推理并共提交了6次结果，实验结果如表3所示。

DataSet	Metric	Vote	Accuracy					
			Total	实体识别	角色识别	异常识别	空间推理	同义识别
Baseline	\	w/o	0.4792	0.7509	0.8818	0.6860	0.2196	0.4200
SpaCE24_test	TeleAI.test.1	w/o	0.5991	0.8895	0.9312	0.8440	0.3471	0.5538
SpaCE24_test	TeleAI.test.2	w	0.5958	0.8895	0.9273	0.8480	0.3373	0.5631
SpaCE24_test	TeleAI.test.3	w/o	0.5898	0.8912	0.9364	0.8360	0.3265	0.5523
SpaCE24_test	TeleAI.test.4	w/o	0.5885	0.8947	0.9260	0.8440	0.3255	0.5492
SpaCE24_test	TeleAI.test.5	w	0.5958	0.8895	0.9273	0.8480	0.3373	0.5631
SpaCE24_test	TeleAI.test.6	w	0.6024	0.8947	0.9364	0.8480	0.3471	0.5631

Table 3: SpaCE24测试集的实验结果

表3展示了在SpaCE24测试集上每次提交结果的具体得分，其中Vote表示是否使用了投票策略，我们发现不同策略下的模型，甚至同结构的模型在不同训练轮次的验证集预测结果上存在差异，因此我们挑选了验证集中表现较好的模型作为投票模型。TeleAI.test.2采用多数投票策略来汇总各模型的预测结果，TeleAI.test.5引入了加权投票机制，根据各模型在验证集上的性能为其预测结果分配特定的权重，TeleAI.test.6进一步改进了这一方法，不仅为各模型设置权重，还为不同类型的题目分配权重，从而通过更细致且上下文敏感的聚合过程，提高最终预测的精确度和鲁棒性。由实验结果可以看出投票策略在该任务上可以平衡不同模型的优势，有效减少随机误差和模型偏差，提高整体的性能，我们在测试集上最终取得了0.6024的准确率，与基线模型相比取得了25.71%的相对提升，在实体识别任务上与基线模型相比取得了19.15%的相对提升，在角色识别任务上与基线模型相比取得了6.19%的相对提升，在异常识别任务上与基线模型相比取得了23.62%的相对提升，在空间推理任务上与基线模型相比取得了58.06%的相对提升，在同义识别任务上与基线模型相比取得了34.07%的相对提升。我们的模型在实体识别、角色识别、异常识别与同义识别任务中达到了较高的准确率，但是在空间推理任务上准确率较低，可以看出大模型在推理能力上仍具有较大的提升空间。

5 总结与展望

在本次CCL2024中文空间语义理解评测任务（SpaCE2024）中，我们基于上下文学习与思维链策略精心设计了prompt并结合微调的方式激发大语言模型的空间语义理解能力，构建了一个高效的空间语义理解系统。在最终的评估中，我们队伍提交的系统在测试集实体识别与异常识别题目中排名第一，综合准确率为0.6024，综合排名第一。

参考文献

- Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Sridhar Iyer, Veselin Stoyanov, and Zornitsa Kozareva. 2022. Improving in-context few-shot learning via self-supervised training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3558–3573.
- 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, Steven Bethard, and Marie-Francine Moens. 2012. Semeval-2012 task 3: 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)}*, volume 2, pages 365–373. ACL.
- Xiao Liming, Sun Chunhui, Zhan Weidong, Xing Dan, Li Nan, Wang Chengwen, and Zhu Fangwei. 2023. Space2022 中文空间语义理解评测任务数据集分析报告(a quality assessment report of the chinese spatial cognition evaluation benchmark). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 547–558.
- James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. 2015. Semeval-2015 task 8: Spaceeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (semeval 2015)*, pages 884–894. ACL.
- Zihan Wang, Xinzhang Liu, Shixuan Liu, Yitong Yao, Yuyao Huang, Zhongjiang He, Xuelong Li, Yongxiang Li, Zhonghao Che, Zhaoxi Zhang, et al. 2024. Telechat technical report. *arXiv preprint arXiv:2401.03804*.
- Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, et al. 2023. Symbol tuning improves in-context learning in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 968–979.
- Liming Xiao, Weidong Zhan, Zhifang Sui, Yuhang Qin, Chunhui Sun, Dan Xing, Nan Li, Fangwei Zhu, and Peiyi Wang. 2023. Ccl23-eval 任务4 总结报告: 第三届中文空间语义理解评测(overview of ccl23-eval task 4: The 3rd chinese spatial cognition evaluation). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 150–158.
- 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*, 2:99–110.