CCL25-Eval任务1系统报告：使用思维链和投票集成增强大型语言模型空间语义理解

刘海芯，咎红英，宋金旺，李一帆，孔露露
郑州大学，计算机与人工智能学院
郑州，450001
{lhxin，jwsong，kll，lyfan}@gs.zzu.edu.cn，iehyzan@zzu.edu.cn

摘要

本技术报告详细介绍了我们团队在第五届空间语义理解评测（SpaCE2025）中的方法与成果。SpaCE2025 继续聚焦大语言模型在空间语义理解方面的能力评估，涵盖空间语言理解与空间推理两个核心维度，共设置五个子任务：空间信息正误判断、空间参照实体判断、空间异形同义判断、中文空间方位关系推理以及英文空间方位关系推理。我们通过设计结构化提示词并引入思维链推理机制，结合LoRA 微调技术和投票集成方法，有效提升了大语言模型在空间语义理解任务中的表现。在最终评测中，我们团队五个子任务的综合准确率为0.5983，整体排名第五。

关键词： 空间语义理解；提示词；思维链；LoRA微调；大语言模型

System Report for CCL25-Eval Task 1: Spatial Semantic Understanding Using Chain-of-Thought and Voting Ensemble Enhanced Large Language Models

Haixin Liu，Hongying Zan，Jinwang Song，Yifan Li，Lulu Kong
Zhengzhou University，School of Computer and Artificial Intelligence
Zhengzhou，450001
{lhxin，jwsong，kll，lyfan}@gs.zzu.edu.cn， iehyzan@zzu.edu.cn

Abstract

This technical report provides a detailed description of our methods and results in the 5th Chinese Spatial Semantic Understanding Evaluation (SpaCE2025). SpaCE2025 continues to focus on evaluating the capabilities of large language models (LLMs) in spatial semantic understanding, covering two core dimensions: spatial language comprehension and spatial reasoning. The evaluation consists of five sub-tasks: judging the correctness of spatial information, identifying spatial reference entities, distinguishing spatial paraphrases, Chinese spatial position reasoning, and English spatial position reasoning.To enhance the performance of LLMs on these tasks, we designed structured prompt templates and incorporated chain-of-thought reasoning, combined with LoRA-based fine-tuning and a voting ensemble approach. In the final evaluation, our system achieved an overall accuracy of 0.5983 across the five sub-tasks, ranking fifth among all participating teams.

Keywords: Spatial Semantic Understanding , Prompt Engineering , Chain-of-Thought , LoRA Fine-tuning , Large Language Models

1 引言

空间表达描述了物体之间的空间方位关系，是自然语言中的高频现象。实现空间语义理解 (Wu et al., 2024)不仅依赖语言知识，还需要调用空间认知能力，准确构建文本表征的空间场景。空间语义理解是人类认知的核心能力之一，也是自然语言处理中实现场景构建、导航指令解析和多模态交互的关键技术。近年来，大语言模型（LLMs）在文本生成 (Gao et al., 2023)、逻辑推理 (Liu and Zhang , 2024)等任务上展现出接近人类的表现，但大语言模型在SpaCE2024 (Xiao et al., 2024)的评测结果显示，大语言模型的空间语义理解水平与普通人类的平均水平相比，在对空间认知加工要求较高的任务上，存在较大差距。这一差距表明，空间语义理解对大语言模型来说仍然是一项挑战性任务。

SpaCE2025评测体系进行了重要改进：一是聚焦高认知难度任务，舍弃已达标的形式标记任务；二是提升数据多样性和平衡性，新增未考察的空间表达；三是新增跨语言空间推理评估，通过中英文对照数据探究语言与推理能力的关系。这些改进旨在更准确地评估大语言模型的空间认知能力。

本次参与第五届空间语义理解评测任务（SpaCE2025），我们系统地探索了大语言模型在空间语言能力与空间推理能力两个维度上的理解与推理能力。围绕五个子任务，我们采用了三种模型（DeepSeek-R1-Distill-Qwen-7B、Qwen2.5-7B-Instruct、Qwen3-4B）进行评测，通过设计结构化提示词并引入思维链推理机制，结合LoRA 微调技术和投票集成方法，有效提升了大语言模型在空间语义理解任务中的表现。在最终评测中，我们团队五个子任务的综合准确率为0.5983，整体排名第五。

2 相关工作

空间语义理解是自然语言处理中高度依赖认知能力的任务。早期研究主要基于规则或浅层机器学习方法，如SemEval-2012 (Kordjamshidi et al., 2012) 和SemEval-2013 (Kolomiyets et al., 2013) 提出的空间关系语义角色标注任务，以及SpaceEval 2015 (Pustejovsky et al., 2015) 引入的ISO-Space 标注体系。

近年来，SpaCE 系列评测系统探索了中文空间语义建模。SpaCE2021 (詹卫东et al., 2022) 首次引入空间语言错误归因评估，SpaCE2022 (Xiao et al., 2023) 扩展了任务范围，而SpaCE2023 (Xiao et al., 2023) 增加了生成任务以考察空间概念建模能力。SpaCE2024 (Xiao et al., 2024) 首次系统性评估大语言模型在中文空间语义任务上的表现，发现模型在低认知负载任务（如语义角色识别）上接近人类水平，但在复杂推理任务上仍有差距。

SpaCE2025 进一步聚焦空间推理与跨语言评估，首次引入中英文对照的方位推理任务，以检验模型的语言中立性。SpaCE2025空间语言能力类评测任务，包括三个子任务：（1）空间信息正误判断。本任务要求机器判断文本的空间信息是否正确。（2）空间参照实体判断。本任务给出可能在句子中充当参照物的实体，要求机器判断该实体是否是正确的参照物。（3）空间异形同义判断。本任务要求机器判断两个文本描述的空间场景是相同还是不同。SpaCE2025空间推理能力类评测任务，包括两个子任务：（1）中文空间方位关系推理。本任务要求机器在中文文本中推理出实体在空间场景中的位置，以及未知的方位关系，选择题。（2）英文空间方位关系推理。与中文推理文本对照的英文文本，要求机器在英文文本中推理出实体在空间场景中的位置，以及未知的方位关系。此外，SpaCE2025显著扩展了语料覆盖与题型均衡性，更全面考察大语言模型在多任务、多语种、多领域下的空间语义理解表现。

整体来看，空间语义理解任务正逐步从信息提取向“空间认知-语言融合”过渡，模型的空间建模能力、语言灵活性及跨语言迁移能力正成为评估焦点。SpaCE2025的设计充分体现了这一趋势，并为后续研究提供了更具挑战性与现实意义的评测基准。

3 方法

在本次评测任务中，空间语言能力类评测任务要求使用不大于7B的模型，因此我们使用DeepSeek-R1-Distill-Qwen-7B、Qwen2.5-7B-Instruct、Qwen3-4B这三个模型进行实验，并对比了零样本采用思维链直接推理，少样本LoRA微调方法的性能。中英文空间方位关系推理任务以DeepSeek-R1-Distill-Qwen-7B作为主模型，将中英文训练集合并进行多任务学习，使用精心设计的提示词并结合思维链的方法，对模型进行LoRA指令微调，根据评测指标设置相应评估函数并使用验证集进行训练中端到端评估，以保存最好的LoRA权重。最后采用投票集成

方法以获取最优结果。此外，我们对五个任务均使用了vLLM框架（Vectorized Large Language Model Inference），通过创新的PagedAttention和连续批处理技术，显著提升推理速度和服务吞吐量。我们的框架图如图1所示。

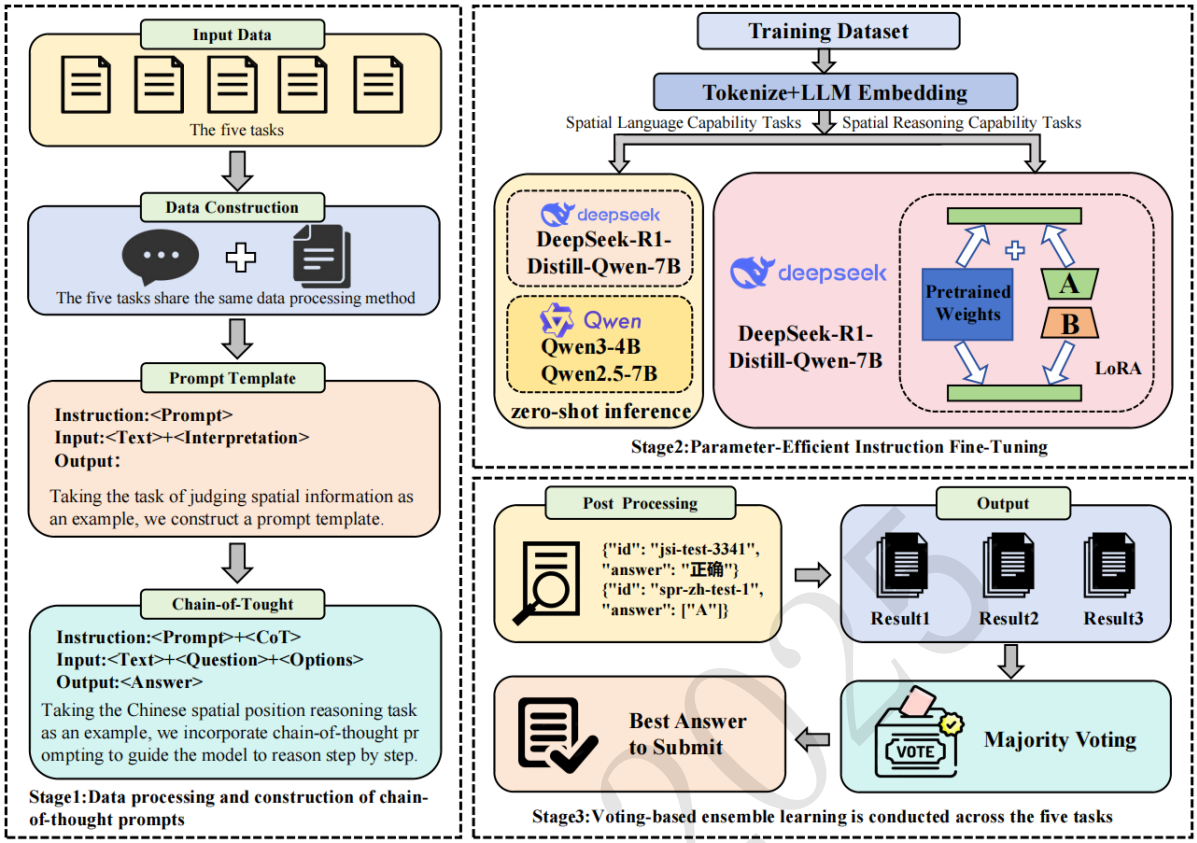

Figure 1: 任务整体框架图

3.1 思维链

思维链（Chain-of-Thought, CoT）（Wei et al., 2022)是一种新兴的人工智能推理方法，旨在通过模拟人类在解决复杂问题时的思维过程，增强机器学习模型的能力。具体而言，CoT方法通过构建多跳（multi-hop）的思维链条，将问题求解过程拆解为一系列具有严格逻辑关联的中间状态，每一个步骤都基于上一步的输出进行进一步的分析与推理，从而逐步逼近最终答案。这种方法不仅能够提高模型在复杂推理任务中的表现，还能增强其决策过程的可解释性，使模型的推理路径更加透明和易于理解。在SpaCE2025任务中，我们也采用了思维链推理方法，以增强大语言模型在处理复杂空间语义任务时的推理能力。

具体实现上，我们为每个空间语义理解子任务人工设计了结构化的CoT提示词模板。这些模板旨在引导模型显式地进行关键的空间认知步骤推理，例如：识别空间场景中的关键实体与潜在参照物、理解方位词（如“上”、“下”、“左”、“右”、“内部”、“相邻”）的确切含义、基于已知条件逐步推导实体间的相对位置关系、构建简化的空间关系示意图、综合所有信息进行判断或选择、以及排除不符合条件的选项。图2展示了一个用于中文空间方位关系推理任务的典型CoT提示词模板示例。针对空间参照实体判断任务，提示词会侧重于引导模型分析给定实体在句子中是否提供了确定其他实体位置所需的参考框架；而对于空间异形同义判断任务，提示词则会引导模型逐项对比两个描述中的实体、方位关系和整体空间布局是否等价。

对于空间语义理解任务，特别是空间方位关系推理与参照实体识别等子任务，准确理解文本中的空间信息往往依赖于多个隐含推理过程，如识别隐式参照系、整合上下文信息、构建合理的空间场景等。传统语言模型在处理这类任务时容易出现跳步或忽略隐含线索的问题，而通过引入CoT方法，我们能够引导模型逐步明确推理路径，从而显著提升模型的推理深度与答案可解释性。

"instruction": "你是一个空间推理专家，需要根据给定的空间场景描述，推理出实体之间的方位关系。题目是多选题，有两个或两个以上的正确答案。答案选项必须与标准答案完全一致才能得分。请逐步思考，并最终输出答案选项。请逐步思考，并最终输出答案选项。\n推理步骤：\n1. 首先，分析场景布局（如正六边形、圆形等），并确定所有实体的初始位置关系。\n2. 根据已知条件，逐一推导每个实体的相对位置。\n3. 绘制一个简化的方位图（如六边形排列），并在图中标注各实体的位置。\n4. 根据问题要求，找出符合题目描述的选项。\n5. 检查每个选项是否满足条件，并排除不符合的选项。",
"input": "text:铁拐李、曹国舅、吕洞宾、张天师四人来到火锅店吃火锅，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：\n(1)吕洞宾在曹国舅右手边挨着坐；\n(2)张天师在铁拐李右手边挨着坐。\nquestion:吕洞宾和()身体朝向不同方向。\noptions:\"A\": 铁拐李,\n\"B\": 张天师,\n\"C\": 曹国舅,\n\"D\": 以上选项都不是\n",
"output": "A,B"

一步一步推理过程：
1.布局分析
长方形餐桌，长边两侧各坐两人 → 形成两排座位（排A和排B）
两排座位面对面布置 → 身体朝向相反
2.位置推导
条件(1)：吕洞宾在曹国舅右手边挨着坐 → 二人同排（排A），曹国舅在左，吕洞宾在右
条件(2)：张天师在铁拐李右手边挨着坐 → 二人同排（排B），铁拐李在左，张天师在右
3.方位图绘制
排A（朝南）：[曹国舅]（左） —— [吕洞宾]（右）
排B（朝北）：[铁拐李]（左） —— [张天师]（右）
4.朝向分析
吕洞宾（排A右侧）朝南
铁拐李（排B左侧）朝北 → 朝向相反
张天师（排B右侧）朝北 → 朝向相反
曹国舅（排A左侧）朝南 → 朝向相同
5.选项验证
A：铁拐李 → 朝北（与吕洞宾朝南相反）→ 正确
B：张天师 → 朝北（与吕洞宾朝南相反）→ 正确
C：曹国舅 → 朝南（与吕洞宾相同）→ 错误
D：以上都不是 → A、B都正确 → 错误
综上所述， 正确答案为A,B

Figure 2: CoT数据生成示例（以中文空间方位关系推理为例）

3.2 参数高效指令微调

在中英文空间方位关系推理任务中，我们采用了LoRA（Low-Rank Adaptation）方法 (Hu et al., 2021)对大型语言模型进行高效微调。LoRA 是一种参数高效微调技术，适用于在计算和显存资源受限的情况下对大规模预训练语言模型（如GPT (Achiam et al., 2023)、BERT (Vaswani et al., 2017)等）进行任务适配。LoRA 的核心思想是：不直接更新预训练模型的原始权重，而是在特定模块（如Attention 层）引入一对可训练的低秩矩阵A和B，对权重更新进行建模：

$$W = W_0 + \Delta W = W_0 + BA \tag{1}$$

其中，$W_0 \in \mathbb{R}^{d\times k}$ 为冻结的预训练权重，$A \in \mathbb{R}^{r\times k}$，$B \in \mathbb{R}^{d\times r}$ 为可训练的低秩矩阵，且秩$r \ll \min(d, k)$。

模型在训练过程中仅更新A 和B，损失函数采用因果语言建模（Causal Language Modeling, CLM）形式，其目标为最大化下一个token 的预测概率，对应的损失函数为：

$$\mathcal{L}_{\text{CLM}} = -\sum_{i=1}^{n} \log p(x_i \mid x_{<i}; \theta) \tag{2}$$

其中x_i 表示第i 个token，θ 为包括LoRA 参数在内的可训练参数集合。LoRA 微调不仅显著减少了所需参数与显存使用，还能保持预训练模型原有的表达能力。在本实验中，我们使用DeepSeek-R1-Distill-Qwen-7B 模型在中英文空间推理子任务中进行LoRA微调。

3.3 数据处理

我们将评测任务提供的数据集处理为instruction-input-output的格式，以中文空间方位推理任务为例，由于测试集已包含题型(多选/单选)等信息，我们使用精心设置的提示词作为instruction,使用训练数据中的text、question和options作为微调数据的input，以保证训练和推理时的指令一致性。我们将中文方位关系推理和英文方位关系的数据集合并进行训练，共4000条训练集。数据处理的形式如图2所示。

3.4 训练中端到端评估及vLLM加速推理

我们在对中英文空间方位关系推理任务的训练中通过调用model.generate()进行生成式的端到端评估，根据评测设置的指标(本次评测指标为Accuracy)设置相应评估函数在训练中进行每50步一评估，每当下一次的权重指标超过上一次，则保存LoRA权重，以此保存最佳指标的LoRA权重，最终使用最佳权重进行推理。为提升模型推理效率，我们在推理阶段引入vLLM框架。vLLM通过优化的显存管理机制与并行化注意力计算，显著减少推理延迟与显存占用。实验表明，使用vLLM后，Qwen3-4B模型的单次推理时间从平均1.2秒降至0.8秒，显存占用减少25%，有效加速了推理过程，提升了整体性能表现。

3.5 投票集成

单一模型的预测结果存在语义不准确和信息短期记忆的问题，投票机制是一种基于多数原则的策略，通过汇聚多个模型的预测结果来增强模型的鲁棒性。在空间语言能力类的三个子任务中，我们使用准确率相近的3-5份结果进行投票。针对空间推理任务中模型预测不稳定的问题，我们在经典投票集成框架 (Zhu et al., 2023)的基础上提出三项改进：首先，采用多权重投票机制，通过融合同一模型不同训练阶段的LoRA检查点预测结果，增强决策多样性；其次，设计阈值自适应规则，根据题型动态调整通过阈值，对单选直接进行投票集成，严格输出多数票结果。对于多选题，我们使用3个不同的LoRA权重进行投票，对每一个选项，统计在三份结果中被选择的次数。如果某个选项在三份结果中被选择的次数大于或等于两个，那么该选项就被最终选择，如果某个选项的选择次数小于两个，则该选项不会被最终选中。最终选择的选项为所有满足“大于或等于2票”的选项；最后，引入冲突消解策略，当出现平票时优先采纳验证集表现最优的检查点输出。该方案在保持投票集成鲁棒性的同时，显著提升了复杂空间场景下的预测一致性。

4 实验

4.1 数据集

SpaCE2025总数据量为18,423 题。空间信息正误判断、空间参照实体判断、空间异形同义判断任务在多种不同类型的真实语料上进行改写工作，包括：报刊语料、文学作品语料、中小学课本语料、交通事故描述文本、人体动作文本、地理百科文本。空间方位关系推理任务则是运用基于知识库的数据合成方法生成的高质量合成数据。数据分布如表1所示。

子任务	示例集	训练集	验证集	测试集	数据总量
空间信息正误判断	20	0	0	3500	3520
空间参照实体判断	20	0	0	1763	1783
空间异形同义判断	20	0	0	1100	1120
中文空间方位关系推理	0	2000	500	3500	6000
英文空间方位关系推理	0	2000	500	3500	6000
合计	60	4000	1000	13363	18423

Table 1: 数据概况

4.2 参数设置

在训练过程中，我们使用LoRA 框架对DeepSeek-R1- Distill-Qwen-7B 模型的部分参数进行了微调，训练了4 个epoch。训练过程中设置了端到端评估机制，每隔50 步进行一次评估，并根据评估结果保存性能最优的LoRA 权重。当模型出现过拟合迹象时，手动停止训练，因此我们设置的epoch 数相对较大。训练时的学习率（Learning rate）设置为2e-5，批大小（batch size）为1。LoRA 微调时的超参数包括：LoRA rank 为64，LoRA alpha 为512，dropout 为0.05。具体的模型参数配置列于表2 中。

DeepSeek-R1-Distill-Qwen-7B	Learning rate	2e-5
	epoch	4
	batch_size	1
	LoRA_rank	64
	LoRA_alpha	512
	LoRA_dropout	0.05

Table 2: 超参数设置

4.3 评估指标

本次评测的排名依据为两大类任务的综合得分S ，S1 代表空间语言能力类评测任务的得分，S2 代表空间推理能力类评测任务的得分，Acci 代表各子任务的准确率（Accuracy，Acc）。公式如下：

$$S = 0.5 \cdot S1 + 0.5 \cdot S2 \tag{3}$$

$$S1 = \frac{1}{3}\sum_{i=1}^{3} Acc_i \tag{4}$$

$$S2 = \frac{1}{2}\sum_{i=1}^{2} Acc_i \tag{5}$$

$$Acc_i = \frac{\#correct}{\#total} \tag{6}$$

4.4 实验设置

实验中使用的超参数已封装在HyperParameters 类中。我们启用了梯度累积（gradient accumulation）和检查点保存机制（checkpointing）以降低内存占用。在微调过程中，仅计算指令输出部分对应的交叉熵损失（cross-entropy loss）。在评估与推理阶段，统一采用贪婪解码（greedy decoding）策略。

4.5 实验结果

表3展示了不同模型在SpaCE2025空间语言能力类任务（信息正误判断、异形同义判断、参照实体判断）上的表现。其中，“Zero-shot CoT”指仅在提示中添加“请逐步思考”的零样本思维链策略，不提供任何示例；“CoT”表示在推理阶段向模型展示少量含推理过程的示例，引导其生成中间步骤；“Train”则表示是否对模型进行LoRA微调（“w”为微调，“w/o”为直接推理）。

实验结果表明，Qwen3-4B模型在未微调（Train=w/o）且结合CoT策略时表现最优，语言能力得分达0.717，显著优于其他模型。思维链策略对任务提升具有显著作用：Qwen2.5-7B-Instruct添加CoT后，语言能力得分从0.647升至0.658。此外，通过多数投票融合策略，模型综合语言能力得分达到0.7315，参照实体判断得分0.7856为子任务最高值，表明不同模型在空间语言理解上存在互补性，融合策略有效提升了整体稳健性与性能。

值得注意的是，表3显示在空间语言能力类任务（信息正误判断、异形同义判断、参照实体判断）上，对Qwen2.5-7B-Instruct和Qwen3-4B进行LoRA微调（Train=w）后的模型性能，普

遍低于直接使用CoT策略进行推理而不进行微调（Train=w/o）的同一模型。我们分析认为，主要原因在于以下两点：（1）训练数据匮乏：如表1所示，这三个子任务未提供官方的训练数据集（训练集数量均为0）。LoRA微调作为一种参数更新方法，在缺乏充分、高质量训练数据的情况下，极易导致模型过拟合到有限的微调样本上，或者学习到数据中的噪声模式，反而损害了模型在预训练阶段获得的知识和泛化能力，无法应对测试集中复杂多样的样本。（2）任务特性与微调适配性：空间语言能力类任务（尤其是正误判断和异形同义判断）的核心挑战在于复杂的常识推理、细粒度的语义理解和上下文依赖的空间逻辑判断。这些任务更依赖于模型固有的世界知识和零样本/少样本的泛化与推理能力。预训练好的大语言模型（特别是Qwen3-4B）本身在这些方面已具备较强的基础。在数据不足的情况下进行微调，可能干扰了模型原有的知识结构和推理模式，未能带来性能提升，甚至产生负面影响。相比之下，空间推理能力类任务拥有相对充足的训练数据（2000条），LoRA微调结合CoT策略可取得稳定的性能提升，这进一步佐证了充足且高质量的微调数据对于LoRA方法发挥效用至关重要。

Model	Zero-shot CoT	CoT	Train	语言能力得分	信息正误判断	异形同义判断	参照实体判断
DeepSeek-R1-Distill-Qwen-7B	w	w/o	w	0.550	0.513	0.596	0.541
DeepSeek-R1-Distill-Qwen-7B	w/o	w	w	0.562	0.525	0.605	0.556
DeepSeek-R1-Distill-Qwen-7B	w/o	w	w/o	0.581	0.554	0.621	0.567
Qwen2.5-7B-Instruct	w	w/o	w	0.647	0.588	0.633	0.719
Qwen2.5-7B-Instruct	w/o	w	w	0.658	0.596	0.653	0.723
Qwen2.5-7B-Instruct	w/o	w	w/o	0.680	0.624	0.670	0.749
Qwen3-4B	w	w/o	w	0.666	0.609	0.652	0.737
Qwen3-4B	w/o	w	w	0.686	0.637	0.664	0.758
Qwen3-4B	w/o	w	w/o	0.717	0.675	0.698	0.778
Voting	**-**	**-**	**-**	**0.7315**	**0.6889**	**0.7200**	**0.7856**

Table 3: 空间语言能力类任务实验结果

表4展示了我们在空间推理能力子任务上的实验结果，包括中文方位关系推理和英文方位关系推理任务。我们基于DeepSeek-R1-Distill-Qwen-7B 模型进行了多组实验，并考察了CoT对模型性能的影响。Lora_acc为在训练中评估生成的Accuracy指标结果，评估数据为中英文关系推理任务的1000条验证集。实验结果表明：推理任务总体困难显著高于语言任务。在不使用CoT提示且未进行微调的基础设置下，模型的空间推理能力得分为0.4271，Lora_acc 为0.436，中文方位推理和英文方位推理的准确率分别为0.4242 和0.4256，反映出在缺乏认知引导和参数调整的情况下，大语言模型在复杂空间推理任务中能力有限。加入CoT 后，推理过程被显式引导，模型推理得分最高提升至0.4493，模型表现明显提升。中文方位推理得分提升至0.4520，英文方位推理也达到0.4466，表明引入链式思维策略与参数调节机制有助于增强模型的空间场景建模与方位关系理解能力。通过将多个LoRA 微调模型结果进行投票集成，最终在三个指标上均取得最优结果：推理能力得分0.4651，中文推理0.4694，英文推理0.4609。这说明多模型融合在处理认知负载较高的任务中具有较强的性能增益与鲁棒性。我们在五个子任务上的最终实验结果如表5所示。

Model	CoT	Lora_acc	推理能力得分	中文方位推理	英文方位推理
DeepSeek-R1-Distill-Qwen-7B	w/o	0.436	0.4271	0.4242	0.4256
DeepSeek-R1-Distill-Qwen-7B	w	0.451	0.4446	0.4467	0.4425
DeepSeek-R1-Distill-Qwen-7B	w	0.453	0.4434	0.4446	0.4423
DeepSeek-R1-Distill-Qwen-7B	w	0.456	0.4493	0.4520	0.4466
Voting	-	-	0.4651	0.4694	0.4609

Table 4: 空间推理能力类任务实验结果

综合得分	语言能力得分	推理能力得分	信息正误判断	异形同义判断	参照实体判断	中文方位推理	英文方位推理
0.5983	0.7315	0.4651	0.6889	0.7200	0.7856	0.4694	0.4609

Table 5: 五个任务在测试集上最终结果

5 总结与展望

本次参与第五届空间语义理解评测任务（SpaCE2025），我们系统地探索了大语言模型在空间语言能力与空间推理能力两个维度上的理解与推理能力。围绕五个子任务，我们采用了三种模型（DeepSeek-R1-Distill-Qwen-7B、Qwen2.5-7B-Instruct、Qwen3-4B）进行评测，并引入了LoRA参数高效微调、CoT推理提示、以及投票集成等策略来提升模型的空间语义处理能力。

在空间语言能力类任务中，Qwen3-4B表现出较强的综合能力，在异形同义判断和参照实体判断任务中取得了优异成绩。而在空间推理能力类任务中，我们使用DeepSeek-R1-Distill-Qwen-7B结合LoRA微调和多数投票策略，实现了在中英文方位推理子任务上的稳定性能提升。在最终的评估中，我们在空间信息正误判断题目中准确率为0.6889，在空间异形同义判断题目中准确率为0.7200，在空间参照实体判断题目中准确率为0.7856，在中文空间方位关系推理题目中准确率为0.4694，在英文空间方位关系推理题目中准确率为0.4609，测试集综合准确率为0.5983，最终排名第五。

尽管我们的方法在整体上取得了较好的效果，但实验也暴露出当前大语言模型在空间认知建模以及复杂方位场景理解方面仍存在明显挑战。空间语义理解不仅需要强大的语言表达建模能力，更依赖对物理空间关系、参照系统和场景逻辑的深入把握。未来可结合视觉信息、多场景交互任务与跨语言对比学习，增强模型的空间理解与泛化能力。

本次评测不仅加深了我们对大语言模型空间语义能力的理解，也为后续空间语言处理系统的设计与研究提供了坚实的实验基础与方法参考。

致谢

本研究由国家自然科学基金重点项目资助（项目编号：U23A20316）。

参考文献

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. *Enabling Large Language Models to Generate Text with Citations.* In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Liu Hanmeng and Zhang Yue. 2024. 大模型逻辑推理研究综述*(Survey on Logical Reasoning of Large Pre-trained Language Models)*. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 48–62, Taiyuan, China. Chinese Information Processing Society of China.

Wu Hongyan, Lin Nankai, Ceng Peijian, Zheng Weixiong, Jiang Shengyi, and Yang Aimin. 2024. 基于上下文学习的空间语义理解. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 113–121, Taiyuan, China. Chinese Information Processing Society of China.

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)*, pages 365–373, Montréal, Canada. Association for Computational Linguistics.

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.

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.

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, Harbin, China. Chinese Information Processing Society of China.

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, Harbin, China. Chinese Information Processing Society of China.

Xiao Liming, Hu Nan, Zhan Weidong, Qin Yuhang, Deng Sirui, Sun Chunhui, Cai Qixu, and Li Nan. 2024. *The Fourth Evaluation on Chinese Spatial Cognition*. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 122–134, Taiyuan, China. Chinese Information Processing Society of China.

Qinyuan Ye, Mohamed Ahmed, Reid Pryzant, and Fereshte Khani. 2024. *Prompt Engineering a Prompt Engineer*. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 355–385, Bangkok, Thailand. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. *A Survey on In-context Learning*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Shizhe Diao, Pengcheng Wang, Yong Lin, Rui Pan, Xiang Liu, and Tong Zhang. 2024. *Active Prompting with Chain-of-Thought for Large Language Models*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1350, Bangkok, Thailand. Association for Computational Linguistics.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, F. Xia, Q. Le, and D. Zhou. 2022. *Chain of Thought Prompting Elicits Reasoning in Large Language Models*. *ArXiv*, abs/2201.11903.

J. E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen. 2021. *LoRA: Low-Rank Adaptation of Large Language Models*. *ArXiv*, abs/2106.09685.

Josh Achiam, Steven Adler, Sandhini Agarwal, et al. 2023. *GPT-4 Technical Report*. OpenAI. https://cdn.openai.com/papers/gpt-4.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. *Attention Is All You Need*. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 5998–6008. Curran Associates, Inc.

Zhiyuan Zhu, Yusheng Liao, Zhe Chen, Yu Wang, and Yunfeng Guan. 2023. *Towards Optimizing Pre-trained Language Model Ensemble Learning for Task-oriented Dialogue System*. In *Proceedings of the Eleventh Dialog System Technology Challenge*, pages 144–149, Prague, Czech Republic. Association for Computational Linguistics.

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.