CCL25-Eval任务四系统报告：基于层次化思维链构造与推理模型高效微调的中文叙实性推理

闫强[1]　范意兴[1]　钟芸霏[2]
[1]中国科学院计算技术研究所，网络数据科学与技术重点实验室
[2]北京师范大学，人工智能学院
yanqiang@ict.ac.cn, fanyixing@ict.ac.cn, zyfsophieza528@163.com

摘要

本文介绍了我们在第二十五届中国计算语言学大会（CCL 2025）中文叙实性推理评测（FIE2025）中荣获双赛道**第一名**和**第二名**的系统方案。针对中文叙实性推理任务中模型需要从谓词语义正确推断事件真实性的挑战，我们提出了**层次化思维链**（Hierarchical Chain-of-Thought, HCoT）推理框架，通过结构化的多级推理过程引导模型逐步识别关键谓词、分析其叙实性类型及其在否定、疑问等复杂语境下的叙实性变化。在非微调赛道中，我们通过集成多种强大的推理型大模型（如Deepseek-R1-671B、Deepseek-v3-671B、GPT-4o、Gemini-2.5-pro-0506等）的预测结果，并采用自适应投票策略，取得了0.9376的分数。在微调赛道上，我们构建了高质量的思维链指令数据集，发现专注于推理能力的基础模型（如DeepSeek-R1-Distill-Qwen-32B）经微调后在叙实性推理任务上优于同等规模甚至更大参数量的通用大模型（如Qwen2.5-72B-Instruct）。通过伪标签训练进一步优化，最终在官方评测中取得0.9396的最高正确率。实验结果表明，我们提出的层次化思维链结构与推理模型的结合在中文叙实性推理任务中具有显著优势，特别是在处理复杂语境和隐含语义的情况下。

关键词： 层次化思维链；推理型大模型；中文叙实性推理

System Report for CCL25-Eval Task 4: Chinese Factivity Inference Based on Hierarchical Chain-of-Thought Construction and Efficient Fine-tuning of Reasoning Models

Qiang Yan[1]　Yixing Fan[1]　Yunfei Zhong[2]
[1]Key Lab of Network Data Science and Technology,
Institute of Computing Technology, Chinese Academy of Sciences
[2]School of Artificial Intelligence, Beijing Normal University
yanqiang@ict.ac.cn, fanyixing@ict.ac.cn, zyfsophieza528@163.com

Abstract

This paper presents our champion system for the Chinese Factivity Inference Evaluation (FIE2025) at the 25th China Computational Linguistics Conference (CCL 2025), achieving first place in the fine-tuning track and second place in the non-fine-tuning track. To address the challenge of correctly inferring event factuality from predicate semantics in Chinese texts, we propose a **Hierarchical Chain-of-Thought (HCoT)** reasoning framework that guides models through a structured multi-level reasoning process to progressively identify key predicates, analyze their factivity types, and examine their factivity changes under complex contexts such as negation and interrogation. In the non-fine-tuning track, we achieved an accuracy of 0.9376 by integrating inference

第二十四届中国计算语言学大会论文集，第134页-第145页，济南，中国，2025年8月11日至14日。

results from multiple reasoning-focused large language models (including Deepseek-R1-671B、Deepseek-v3-671B、GPT-4o、Gemini-2.5-pro-0506, etc.) using an adaptive voting strategy. For the fine-tuning track, we constructed a high-quality chain-of-thought instruction dataset and discovered that base models specialized in reasoning capabilities (such as DeepSeek-R1-Distill-Qwen-32B) outperformed dialogue models of similar or even larger sizes (such as Qwen2.5-72B-Instruct) after fine-tuning for factivity inference tasks. Through further optimization with pseudo-label training, we ultimately achieved the highest accuracy of 0.9396 in the official evaluation. Our experimental results demonstrate that the combination of our hierarchical chain-of-thought structure and reasoning-focused models offers significant advantages in Chinese factivity inference tasks, particularly when handling complex contexts and implicit semantics.

Keywords: Hierarchical Chain-of-Thought , Reasoning-focused LLMs , Chinese Factivity Inference

1 引言

叙实性推理（*Factivity Inference*, FI）是自然语言理解中的一项核心认知能力，指从特定谓词的语义特性推断相关事件真实性的语言推理过程 (袁毓林, 2020)。作为语义理解的重要组成部分，叙实性推理广泛存在于人类日常交际中，对于构建具备深层语义理解能力的智能系统具有重要意义。第一届中文叙实性推理评测（**FIE2025**）为评估大型语言模型（LLMs）在中文语境下的叙实性推理能力提供了标准化测试平台 (Tianchi-FIE2025, 2025)，推动了该领域的研究进展。

然而，中文叙实性推理任务面临诸多挑战。首先，模型需要准确识别句子中的关键叙实性谓词，这要求对中文语法结构和词汇语义的深入理解；其次，不同类型的谓词（叙实性、反叙实性和非叙实性）具有不同的语义特性，模型需要掌握这些复杂的语义分类知识；最后，在否定句、疑问句等复杂语境下，谓词的叙实性可能发生动态变化，进一步增加了推理难度 (曾et al., 2008)。现有方法主要依赖简单的提示词设计或直接微调，缺乏系统性的推理结构指导，在处理复杂语境和隐含语义时表现有限。

针对上述挑战，我们提出了基于层次化思维链（**Hierarchical Chain-of-Thought, HCoT**）的推理框架，通过结构化的多级推理过程引导模型逐步识别关键谓词、分析其叙实性类型，并考虑复杂语境下的叙实性变化。在非微调赛道中，我们集成多种推理型大模型[1]的预测结果，采用自适应投票策略，取得了**0.9376**的正确率；在微调赛道上，我们构建高质量的思维链指令数据集，发现专注于推理能力的基础模型经微调后在叙实性推理任务上优于同等规模甚至更大参数量的通用大模型，最终通过伪标签训练优化获得**0.9396**的最佳成绩，在双赛道中分别荣获第二名和第一名。实验结果表明，精心设计的层次化思维链结构与推理能力优化的模型相结合，能够有效提升中文叙实性推理的表现，特别是在处理复杂语境和隐含语义方面具有显著优势。

2 相关工作

2.1 思维链

思维链推理（Chain-of-Thought, CoT）作为一种提升大模型推理能力的关键技术，已被广泛应用于复杂任务求解中 (Chu et al., 2023)。传统CoT方法通常采用单层线性推理路径，难以处理需要多维度分析的复杂语义任务。近期研究开始探索**多级思维链**（Multi Level Chain-of-Thought, MLCoT）方法，通过构建结构化的多级推理过程来提升模型的复杂推理能力 (Tahmid and Sarker, 2024)。在叙实性推理任务中，思维链可以引导模型先识别关键谓词，再分析其叙实性类型，最后考虑语境因素（如否定、疑问等）对叙实性的影响，形成从局部到整体的**递进式结构化分析**。与传统CoT相比，结构化多级思维链能更有效地处理语义复杂度

[1]术语说明：本文中"推理型大模型"特指具备思维链输出能力的大模型（如DeepSeek-R1系列），区别于一般的通用型大模型；"模型推理过程"指模型执行预测的计算过程；"叙实性推理"指从谓词语义推断事件真实性的预测任务。

高、需要多步交互推理的任务，显著提升了模型在真实性推理等依赖语言内部语义关系的任务中的表现 (Wei et al., 2022)。

2.2 推理型大模型

随着大语言模型技术的发展，专注于推理能力的**推理型大模型**逐渐成为研究热点 (Wang et al., 2022)。与传统对话型大模型相比，推理型大模型在架构设计、预训练目标和指令优化上更加注重逻辑推理能力的培养。近期研究表明，即使规模较小的推理型模型（如DeepSeek-R1系列）在特定推理任务上也能优于规模更大的通用非推理大模型 (Guo et al., 2025)。这些模型通常采用**多头自注意力机制**的改进版本，更有效地捕捉长距离依赖关系，并在训练过程中加入针对性的逻辑推理样本 (Reid et al., 2024)。在中文叙实性推理等语义理解任务中，推理型大模型表现出明显优势，特别是在处理需要利用谓词叙实性特征进行事件真实性判断的复杂场景时 (Liu et al., 2025)。

2.3 参数高效微调

针对大语言模型的微调，**参数高效微调**（Parameter-Efficient Fine-Tuning, PEFT）技术已成为平衡计算资源与模型性能的主要策略 (Han et al., 2024)。传统全参数微调方法计算成本高且易导致过拟合，而PEFT方法通过仅更新模型中的小部分参数，显著降低了训练成本并保持了模型的泛化能力。在推理任务微调中，LoRA（Low-Rank Adaptation）等技术通过在原始权重矩阵旁增加低秩矩阵来实现参数高效更新，已被证明在保持模型通用能力的同时能有效提升特定任务性能 (Dettmers et al., 2023)。近期研究进一步探索了**适应性rank配置**和**层次选择性微调**策略，根据不同层的重要性动态调整微调参数量，在叙实性推理等复杂语义任务上取得了更好的效果。结合伪标签训练和数据增强等技术，参数高效微调方法可以进一步提升模型在低资源条件下的推理性能 (Hu et al., 2022)。

3 实现方法

图 1 展示了我们方法的整体流程框架，分为思维链设计、不微调赛道与微调赛道两大部分，体现了从推理提示构造到模型训练与集成推理的完整链条。

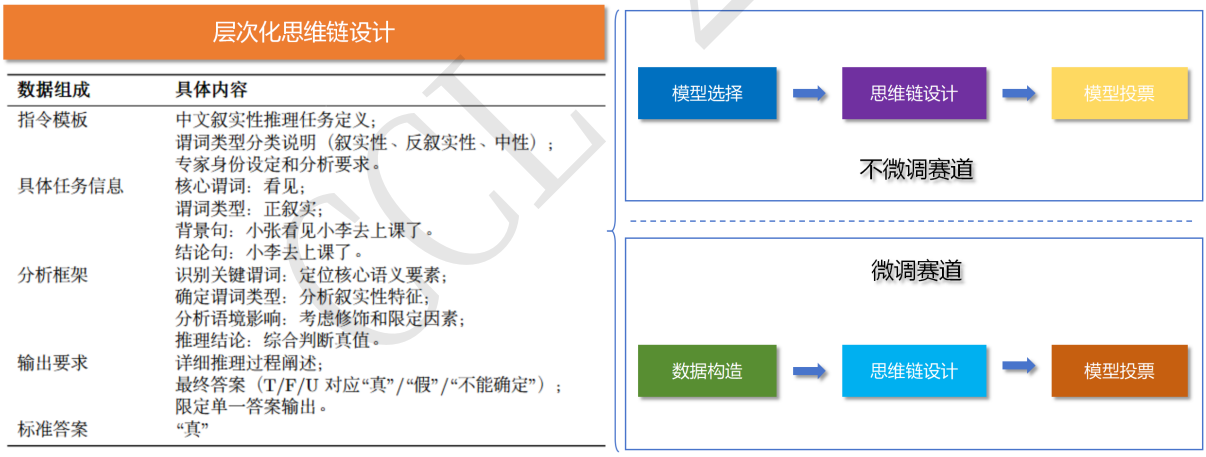

Figure 1: 中文叙实性推理任务的系统框架与处理流程

3.1 层次化思维链设计

针对中文叙实性推理任务的特点，本文提出了层次化思维链（Hierarchical Chain-of-Thought, HCoT）推理框架。该框架通过结构化的多级推理过程，引导大语言模型从复杂的语义分析中逐步提取关键信息，最终得出准确的叙实性判断。

HCoT框架的核心思想是将复杂的叙实性推理任务分解为四个层次化的子任务。首先，模型需要从给定的背景句中识别出核心的谓词成分，这些谓词通常是决定整个句子叙实性特征的关键要素。其次，模型需要根据语言学知识对识别出的谓词进行分类，判断其属于叙实性谓词、反叙实性谓词还是中性谓词。第三个层次要求模型分析否定词、情态词等语境因素对谓词

叙实性特征的影响，因为这些语境元素可能会改变谓词的预设特性。最后，模型需要综合前三个层次的分析结果，推导出结论句的真值判断。

相较于传统的端到端推理方式，HCoT框架的优势在于其推理过程的可解释性和系统性。通过将复杂的语义分析任务分解为可管理的子任务，该框架不仅提高了模型推理的准确性，还保证了推理过程的连贯性和透明性。

为了确保思维链提示的一致性和标准化，我们设计了如表1所示的结构化提示模板。该模板包含了任务背景介绍、叙实性概念定义、具体分析步骤和输出格式要求等关键组成部分，为模型提供了清晰的推理指导。

组成部分	具体内容
任务定义	中文叙实性推理任务的基本介绍，明确任务目标和重要性
概念解释	叙实性谓词、反叙实性谓词、中性谓词的定义和典型例子
输入信息	核心谓词、谓词类型、背景句、结论句等结构化输入
分析步骤	1. 识别关键谓词：找出背景句中的核心谓词 2. 确定谓词类型：判断谓词的叙实性特征 3. 分析语境影响：考虑否定词、情态词等因素 4. 推理结论：综合分析得出结论句真值
输出要求	详细推理过程说明+ 最终答案（T/F/U）

Table 1: 层次化思维链提示模板结构

值得注意的是，HCoT框架的设计充分考虑了中文语言的特殊性。中文句子中的叙实性标记往往更加隐含和复杂，需要模型具备深入的语义理解能力。通过层次化的分析过程，该框架能够有效处理中文语境中的复杂语义现象，特别是在处理否定结构、情态表达和隐含预设等方面表现出色。

3.2 模型选择

在中文叙实性推理任务中，模型的选择直接影响推理效果的优劣。基于任务特性和模型能力的深入分析，我们针对不同赛道制定了相应的模型选择策略。

不微调赛道模型	微调赛道模型
DeepSeek-R1-Distill-Qwen-32B	DeepSeek-R1-Distill-Qwen-32B
DeepSeek-v3-671B	Qwen3-8B
DeepSeek-R1-671B	Qwen3-14B
GPT-4o	Qwen3-32B
Gemini-2.5-pro-0506	Qwen2.5-72B-Instruct
	Qwen3-32B

Table 2: 中文叙实性推理任务中的模型选择

不微调赛道模型选择策略：对于不微调赛道，我们重点关注模型的原生推理能力和语义理解深度。叙实性推理任务要求模型能够准确识别谓词的语义特征并进行复杂的逻辑推断，这对模型的推理架构和预训练质量提出了较高要求。经过广泛的实验验证，我们发现大规模推理型模型在此类任务上表现卓越。Deepseek-R1-671B作为专门优化推理能力的超大规模模型，在处理复杂语义关系时展现出显著优势。GPT-4o凭借其强大的多模态理解能力和精细的语言建模，在叙实性判断的准确性上表现突出。Gemini-2.5-pro-0506则在语境理解和细粒度语义分析方面具有独特优势，特别适合处理中文语境中的隐含语义现象。

微调赛道模型选择策略：微调赛道的模型选择需要平衡推理能力、微调效率和计算资源约束。通过对比实验，我们发现了一个重要现象：专注于推理能力的中等规模基础模型经过精心微调后，其性能可以超越更大规模的通用非推理大模型。DeepSeek-R1-Distill-Qwen-32B作为推理型模型的代表，其架构设计更贴近逻辑推理任务的需求，在叙实性推理的关键环节（如谓

第二十四届中国计算语言学大会论文集，第134页–第145页，济南，中国，2025年8月11日至14日。

词识别、语义分类、语境分析）上具有天然优势。相比之下，虽然Qwen2.5-72B-Instruct等非推理大模型参数规模更大，但其设计目标偏向通用对话生成，在专业语义分析任务上的表现反而不如专业化的推理模型。

针对叙实性推理这类专业语义分析任务，模型的专业化推理能力比纯粹的参数规模更为关键。这一发现为未来的模型选择和优化提供了重要指导。

3.3 模型投票

在不微调赛道中，为了充分利用不同模型的互补优势，我们设计了基于专家仲裁的自适应投票策略。该策略的核心思想是通过多模型预测结果的综合分析，由专业能力较强的模型担任最终仲裁者，从而减少单一模型的判断偏差。

我们的投票机制包含多模型并行推理、结果汇总分析和专家模型仲裁三个关键环节。首先，我们选择了在叙实性推理任务上表现优异的多个模型进行并行预测，包括Deepseek-r1-671b、GPT-4o、Gemini-2.5-pro-0506等。随后，将各模型的预测结果和推理过程整合成结构化的信息，供仲裁模型进行综合分析。最终，我们选择Gemini-2.5-pro-0506作为仲裁模型，基于其在复杂语义理解和逻辑推理方面的卓越表现。

为确保投票过程的标准化和有效性，我们精心设计了专门的投票提示模板，如表 3 所示。该模板不仅包含叙实性推理的理论背景和任务定义，还提供结构化的分析框架，引导仲裁模型进行系统性判断。

模板组成	具体内容
任务背景	叙实性推理的定义与重要性； 不同谓词类型的特征说明（如正叙实、反叙实、中性）； 专家身份设定与任务目标明确。
原始问题信息	背景句（text）； 结论句（hypothesis）； 核心谓词（predicate）； 谓词类型（type）。
多模型预测汇总	列出各模型预测结果（T/F/U）； 对预测一致性进行分析； 识别模型间存在的冲突并加以标注。
分析框架	回顾原始任务信息； 评估各模型推理的主要依据与差异； 识别冲突与一致性的模式； 整合信息，形成专业化最终判断。
输出要求	提供详细的推理过程说明； 输出唯一最终答案（T/F/U）； 给出判断的置信度评分。

Table 3: 模型投票提示语模板结构

3.4 数据构造

针对微调赛道的特殊需求，我们构建了高质量的层次化思维链指令数据集。该数据集的核心目标是引导模型学习叙实性推理的内在逻辑，通过结构化的推理过程提升模型在复杂语义分析任务上的表现。

在初步实验中，我们使用了基于默认提示语[2]的指令作为基础，评估模型的零样本能力和推理倾向。后续我们通过系统性设计，进一步优化了提示结构并构建出具有层次化推理能力的微调数据。

图 2 展示了我们构建的层次化思维链指令数据的典型样例，体现了数据集在结构化设计与推理引导方面的关键特征。

[2]默认提示语参考了公开项目https://github.com/UM-FAH-Yuan/Factivity-LLM 中提供的数据指令。

数据构造过程遵循理论指导、结构化设计、质量优先的原则。我们首先基于语言学理论中的叙实性分类体系，确保每个训练样本都包含明确的谓词类型标注和语义特征说明。随后，为每个样本设计了详细的思维链分析过程，涵盖谓词识别、语义分类、语境分析和逻辑推理四个层次。数据集涵盖了各类谓词类型（叙实性、反叙实性、中性）以及多种语境条件（肯定句、否定句、疑问句等），确保模型能够学习到全面而多样化的推理模式。

为保证数据质量，我们建立了多层次质量控制机制。所有构建的思维链都经过语言学专家的审核，确保推理过程与理论知识一致，推理结论与标准答案匹配。此外，我们还通过交叉验证和一致性检查，进一步提升了数据集的可靠性和有效性。

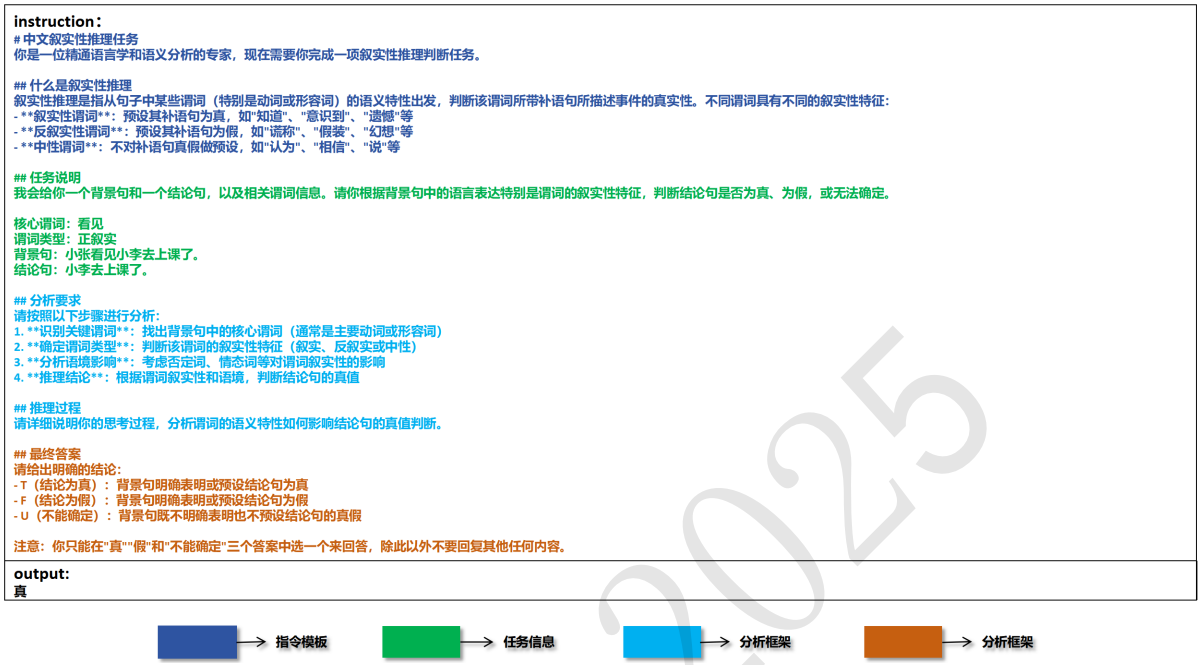

Figure 2: 层次化思维链指令数据样例

3.5 参数高效微调

考虑到计算资源的限制和模型规模的庞大，我们采用了参数高效微调（Parameter-Efficient Fine-Tuning, PEFT）策略来优化模型性能。该策略的核心优势在于仅需更新模型中的少量关键参数，即可实现显著的性能提升，同时避免了全参数微调带来的计算成本过高和过拟合风险。

在具体实现上，我们选择了LoRA（Low-Rank Adaptation）技术作为主要的微调方法。LoRA通过在原始权重矩阵旁引入低秩分解矩阵，实现了参数更新的高效化。相比传统的全参数微调，LoRA不仅大幅降低了显存需求和训练时间，还能更好地保持模型的通用能力，避免在特定任务上的过度拟合。表 4 展示了LoRA微调的具体参数配置。

参数名称	参数值	说明
train_type	LoRA	低秩适应技术，提升微调效率
lora_rank	128	平衡表示能力与资源消耗
lora_alpha	64	LoRA 缩放因子，有助于训练稳定
target_modules	all-linear	对所有线性层进行参数注入

Table 4: LoRA 超参数配置

在微调过程中，我们以最小化交叉熵损失函数为优化目标:

$$J(\theta) = -\sum_{z_i \in V} z_i \log\left(P(z_i)\right) \tag{1}$$

$$P(z_i) = \frac{e^{z_i}}{\sum_{z_i \in V} e^{z_i}} \tag{2}$$

其中，V 是词汇表的大小，z_i 为真实分布中第i 个词的值（对于one-hot 编码，目标词的z_i为1，其余为0），$P(z_i)$ 为模型预测该词的概率。

通过将微调范围设定为"all-linear"，我们对模型中所有线性变换层进行了优化，确保推理链路的全面提升，训练过程中采用了梯度累积和学习率预热等技术，进一步提升了微调的效果和稳定性。

3.6 伪标签学习

为了进一步提升模型性能，我们引入伪标签学习策略作为微调过程的重要补充。该策略利用已微调模型对未标注数据进行高质量标注，通过扩充训练数据规模来提升模型在多样化场景下的泛化能力。我们采用基于置信度的筛选机制，利用微调后的DeepSeek-R1-Distill-Qwen-32B模型对测试集进行叙实性推理预测，根据模型输出的概率分布计算置信度分数，仅选择置信度超过预设阈值的样本作为伪标签数据，有效避免了低质量伪标签的负面影响。在数据融合方面，我们采用渐进式混合训练方法，保持原始标注数据权重不变的同时逐步引入筛选后的伪标签数据进行联合训练。经过实验发现，过多轮数的筛选标注收益递减且变动较小，因此我们最终通过两轮筛选和标注将测试集数据加入训练，在保证标签质量的前提下有效扩充了训练数据规模。

4 实验

4.1 评测指标

本次评测采用总正确率（Total Accuracy）作为主要评价指标，计算公式如下：

$$\text{total\_acc} = \frac{\text{correct\_art} + \text{correct\_nat}}{\text{total\_art} + \text{total\_nat}}$$

其中，total_acc 为总正确率，correct_art 为人造语料集中模型回答正确的数据量，correct_nat 为真实语料集中模型回答正确的数据量，total_art 为人造语料集中的数据总量，total_nat 为真实语料集中的数据总量。

该指标能够综合反映模型在人造语料和真实语料两个子集上的表现，为模型的叙实性推理能力提供全面的量化评估，评测任务是将微调模型方向与不微调模型方向分开评比。

4.2 实验参数设置

4.2.1 训练参数设置

参数名称	参数值	说明
参数精度	bfloat16	使用混合精度进行高效训练
训练轮数	3	保证训练充分，同时避免过拟合
批次大小	1（训练/验证）	考虑显存限制，采用小批次训练
学习率	1e-4	稳定的初始学习率，便于模型收敛
梯度累积步数	8	实现等效大批次训练
最大序列长度	2048	提供充分语境信息以提升推理能力
预热比例	0.05	采用预热策略平滑启动学习率

Table 5: 训练超参数配置

表 5 详细展示了我们在DeepSeek-R1-Distill-Qwen-32B 模型上所采用的参数高效微调配置。这些超参数经过精心调优，有效提升了模型在叙实性推理任务中的性能。

4.2.2 推理参数设置

在推理阶段，我们采用精细调控的解码策略，以降低模型生成结果的随机性、提升输出的稳定性和置信度，同时为思维链式推理保留充足的生成长度预算，从而更好地发挥大模型的推理能力。表 6 汇总了本实验中的主要推理参数配置。

参数名称	参数值	说明
infer_backend	pt	使用PyTorch 后端进行模型推理
temperature	0.1	极低的采样温度，控制输出确定性
Top-p	0.7	保留一定随机性以增强语言自然性
Top-k	10	限定候选词数量以提升输出质量
max_new_tokens	512	保留足够字符预算，支持复杂思维链生成

Table 6: 推理阶段参数配置

在保持生成结果稳定性的同时，较大的最大生成长度（512 tokens）允许模型输出更完整的思维链结构，提升模型在叙实性推理任务中的表现。此外，通过设置较低的温度（0.1）与限制性采样策略（Top-p = 0.7, Top-k = 10），有效降低了推理输出的随机性，确保高置信度、高一致性的响应。

4.3 实验结果

4.3.1 不微调赛道实验结果

在不微调赛道中，我们通过系统性的提示词优化和模型集成策略来提升叙实性预测效果。表7展示了不同模型配置下的详细实验结果，充分验证了我们所提出方法的有效性。

序号	模型	total_acc	correct_art	correct_nat
1	DeepSeek-R1-Distill-Qwen-32B（默认提示语）	0.6766	0.7222	0.6595
2	DeepSeek-R1-Distill-Qwen-32B（HCot）	0.8184	0.8011	0.8250
3	DeepSeek-v3-671B（HCot）	0.8224	0.8029	0.8297
4	Deepseek-r1-671B（HCot）	0.8724	0.8889	0.8662
5	GPT-4o（HCot）	0.9328	0.9588	0.9230
6	Gemini-2.5-pro-0506（HCot+投票）	**0.9376**	**0.9780**	**0.9224**

Table 7: 不微调赛道实验结果

实验结果表明，我们提出的层次化思维链（HCoT）推理框架在不微调场景下取得了显著的性能提升。通过对比序号1和序号2的结果可以发现，相同的DeepSeek-R1-Distill-Qwen-32B模型在采用HCoT提示后，总体正确率从67.66%大幅提升至81.84%，提升幅度达到14.18个百分点。这一结果充分验证了结构化思维链在引导模型进行复杂语义推理方面的重要作用。

进一步分析不同规模模型的表现，我们发现大规模推理型模型在叙实性推理任务上展现出明显的优势。Deepseek-r1-671B凭借其强大的推理架构，在采用思维链提示后达到了87.24%的正确率。更为突出的是GPT-4o模型，其在思维链引导下实现了93.28%的高正确率，特别是在人造语料集上达到了95.88%的优异表现，这表明大规模模型在处理结构化语义分析任务时具有显著优势。

最终，通过引入基于专家仲裁的自适应投票策略，我们成功将系统性能提升至93.76%。该投票机制以Gemini-2.5-pro-0506作为仲裁模型，综合多个高性能模型的预测结果，有效减少了单一模型的判断偏差。值得注意的是，该方法在人造语料集上取得了97.80%的卓越表现，同时在真实语料集上也保持了92.24%的稳定性能，体现了良好的泛化能力。通过这一精心设计的投票策略，我们成功将不微调赛道的性能从单一模型的最高水平87.24%提升至93.76%，充分展现了模型集成方法在复杂语义任务中的有效性。

4.3.2 微调赛道实验结果

微调赛道的实验通过系统性的模型选择、参数优化和数据增强策略来提升性能。表8展示了不同模型配置的详细结果。

序号	模型与方法	total_acc	correct_art	correct_nat
1	Qwen3-8B + 默认提示语	0.7066	0.5932	0.7493
2	Qwen3-14B + 默认提示语	0.8754	0.9140	0.8608
3	Qwen3-32B + 默认提示语	0.8979	0.9337	0.8845
4	DeepSeek-R1-Distill-Qwen-32B + 默认提示语	0.9107	0.9427	0.8986
5	Qwen3-32B + HCot	0.9190	0.9523	0.9064
6	DeepSeek-R1-Distill-Qwen-32B + HCot	0.9259	0.9695	0.9095
7	Qwen2.5-72B-Instruct+HCot	0.9117	0.9606	0.8932
8	DeepSeek-R1-Distill-Qwen-32B + 伪标签	0.9269	0.9677	0.9115
9	模型集成（多模型结果集成）	**0.9396**	**0.9779**	**0.9252**

Table 8: 微调赛道实验结果

实验结果揭示了几个重要现象。首先，模型规模的增大对性能提升具有显著效果，从Qwen3-8B的70.66%到Qwen3-32B的89.79%，展现了参数规模对推理能力的重要影响。其次，优化策略的效果差异明显：HCoT方法在不同模型上均表现出色，将Qwen3-32B的性能从89.79%提升至91.90%，将DeepSeek-R1-Distill-Qwen-32B从基础提示语的91.07%大幅提升至92.59%。

值得注意的是，专门优化推理能力的DeepSeek-R1-Distill-Qwen-32B模型在应用HCOT后的表现（92.59%）超越了参数规模更大的Qwen2.5-72B-Instruct模型（91.17%），这表明针对性的推理能力训练比纯粹的参数规模更为关键。通过精细的参数微调策略，我们成功将DeepSeek-R1-Distill-Qwen-32B模型在叙实性推理任务中的正确率从81.84%显著提升至92.59%，实验结果充分验证了该方法在资源受限场景下的有效性与实用性。

伪标签学习策略进一步将DeepSeek-R1-Distill-Qwen-32B的性能提升至92.69%。实验结果表明，伪标签学习策略能够有效提升模型在叙实性推理任务上的表现。通过引入高质量的伪标签数据，模型接触到了更加多样化的语言表达模式和推理场景，增强了其对复杂语境的理解能力。虽然提升幅度相对有限，但这一改进为最终的模型融合策略奠定了坚实基础。更重要的是，伪标签学习提升了模型对边界样本和复杂语境的处理能力，这对于叙实性推理这类依赖细粒度语义理解的任务而言具有重要意义。

最终，通过多模型多数投票集成策略，我们在官方评测中取得了93.96%的最高正确率，其中人工文本识别准确率达到97.79%，自然文本识别准确率为92.52%，验证了所提方法在中文叙实性推理任务中的有效性。

4.3.3 思维链设计效果验证

为了验证层次化思维链（HCoT）设计的有效性，我们在DeepSeek-R1-Distill-Qwen-32B模型上进行了对比实验。实验结果表明，采用HCoT框架后，模型在叙实性推理任务上的正确率从67.66%显著提升至81.84%，在微调赛道同样可以从91.07%提升到92.59%,充分验证了该思维链框架的有效性。

这一显著提升主要体现在以下几个方面：首先，HCoT框架通过结构化的推理过程，引导模型从复杂的语义分析中逐步提取关键信息，提高了推理的系统性和准确性。其次，该框架将复杂的叙实性推理任务分解为谓词识别、语义分类、语境分析和逻辑推理四个层次化的子任务，使模型能够更好地处理中文语境中的复杂语义现象。最后，HCoT框架的设计充分考虑了中文语言的特殊性，特别是在处理否定结构、情态表达和隐含预设等方面表现出色。

4.4 结果分析

通过对实验结果的深入分析，我们得出以下关键发现：1. **思维链的有效性**：层次化思维链设计显著提升了模型的叙实性推理能力，这表明引导模型进行结构化的分析对于复杂的语义理解任务至关重要；2. **模型类型与任务匹配**：在叙实性推理这类需要深度语义理解的任务上，专注于推理能力的模型（如DeepSeek-R1系列）比通用大模型更具优势，这说明模型的专业能

力比纯粹的规模更为关键；3. **人造语料与真实语料表现差异**：大多数模型在人造语料上的表现优于真实语料，反映出真实语境中的叙实性推理更为复杂，包含更多隐含语义和上下文依赖；4. **微调与不微调效果比较**：经过精心微调的中等规模模型（如32B级别）可以达到甚至超越不微调的大规模模型（如671B级别），展示了高效微调策略的价值；5. **集成方法的优势**：在两个赛道中，模型集成和投票策略都能有效提升最终性能，表明不同模型之间存在互补性，综合多模型结果可以降低单一模型的误判风险。

5 总结

本研究针对中文叙实性推理任务，提出了层次化思维链（HCoT）框架和一系列模型优化策略，在CCL 2025 FIE评测中取得了双赛道第一名和第二名的优异成绩。

本文的主要贡献包括设计了结构化的层次化思维链框架，发现专注推理能力的基础模型在叙实性推理任务上具有显著优势，提出了有效的模型集成策略，并通过构建高质量思维链指令数据集和伪标签学习进一步提升模型性能。实验结果表明，我们的方法在官方评测微调赛道中取得了93.96%的最高正确率，为中文叙实性推理和相关语义理解任务提供了有价值的参考。

未来工作将探索叙实性推理与其他语义任务的联系，研究更通用的语义理解框架，以及如何将语言学知识更有效地融入大型语言模型中。

致谢

本研究工作得益于开源微调框架ms-swift (Zhao et al., 2024)所提供的支持，该工具为我们实验过程中的模型微调推理提供了高效灵活的实现。

参考文献

D. Ziembicki, K. Seweryn, and A. Wróblewska. 2024. Polish natural language inference and factivity: An expert-based dataset and benchmarks. *Natural Language Engineering*, 30(2):385–416.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, D. Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *Neural Information Processing Systems*.

X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, D. Zhou. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *International Conference on Learning Representations*.

E. Zelikman, Y. Wu, N.D. Goodman. 2022. STaR: Self-Taught Reasoner Bootstrapping Reasoning With Reasoning. *Neural Information Processing Systems*.

A.K. Lampinen, I. Dasgupta, S.C.Y. Chan, K. Matthewson, M.H. Tessler, A. Creswell, J.L. McClelland, J.X. Wang, F. Hill. 2022. Can language models learn from explanations in context?. *Conference on Empirical Methods in Natural Language Processing*.

U. Katz, M. Geva, J. Berant. 2022. Inferring Implicit Relations with Language Models. *Conference on Empirical Methods in Natural Language Processing*.

X. Ye and G. Durrett. 2022. The Unreliability of Explanations in Few-Shot In-Context Learning. *International Conference on Machine Learning*.

T. Kojima, S.S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. *Neural Information Processing Systems*.

D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, O. Bousquet, Q. Le, E. Chi. 2022. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. *International Conference on Learning Representations*.

A. Creswell, M. Shanahan, I. Higgins. 2022. Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning. *International Conference on Learning Representations*.

Y. Li, Z. Lin, S. Zhang, Q. Fu, B. Chen, J.G. Lou, W. Chen. 2022. On the Advance of Making Language Models Better Reasoners. *Conference on Empirical Methods in Natural Language Processing*.

J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E.H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*.

E. Dyer and G. Gur-Ari. 2022. Minerva: Solving Quantitative Reasoning Problems with Language Models. Google Research Blog.

W.X. Zhao, K. Zhou, Z. Gong, B. Zhang, Y. Zhou, J. Sha, Z. Chen, S. Wang, C. Liu, J.R. Wen. 2022. JiuZhang: A Chinese Pre-trained Language Model for Mathematical Problem Understanding. *Conference on Empirical Methods in Natural Language Processing*.

S. Zhang, R. Shuttleworth, D. Austin, Y. Hicke, L. Tang, S. Karnik, D. Granberry, I. Drori. 2022. A Dataset and Benchmark for Automatically Answering and Generating Machine Learning Final Exams. *Neural Information Processing Systems*.

X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, D. Zhou. 2022. Rationale-Augmented Ensembles in Language Models. *Neural Information Processing Systems*.

Zeyu Han, Shizhe Diao, Liang Pang, Xirong Li. 2024. Parameter-efficient Fine-tuning for Large Models: A Comprehensive Survey. *arXiv preprint arXiv:2403.14608*.

E.J. Hu, Y. Shen, P. Wallis, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

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

A. Yang, A. Li, B. Yang, et al. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

D. Guo, D. Yang, H. Zhang, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Z. Chu, J. Chen, Q. Chen, et al. 2023. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*.

H. Liu, J. Zhang, X. Wang, et al. 2025. Logical Reasoning in Large Language Models: A Survey. *arXiv preprint arXiv:2502.09100*.

S. Tahmid and S. Sarker. 2024. Qwen2.5-32B: Leveraging Self-Consistent Tool-Integrated Reasoning for Bengali Mathematical Olympiad Problem Solving. *arXiv preprint arXiv:2411.05934*.

R. Patil and V. Gudivada. 2024. A review of current trends, techniques, and challenges in large language models (llms). *Applied Sciences*, 14(5):2074.

A. Shypula, M. Chen, K. Thompson, et al. 2025. Evaluating the diversity and quality of llm generated content. *arXiv preprint arXiv:2504.12522*.

M. Lewis, Y. Liu, N. Goyal, et al. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

J. Achiam, S. Adler, S. Agarwal, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

S. Murugesan. 2025. The Rise of Agentic AI: Implications, Concerns, and the Path Forward. *IEEE Intelligent Systems*, 40(2):8–14.

H. Que, Y. Zhang, L. Wang, et al. 2024. HelloBench: Evaluating Long Text Generation Capabilities of Large Language Models. *arXiv preprint arXiv:2409.16191*.

K. Li, J. Chen, M. Zhang, et al. 2025. LaRA: Benchmarking Retrieval-Augmented Generation and Long-Context LLMs–No Silver Bullet for LC or RAG Routing. *arXiv preprint arXiv:2502.09977*.

DeepSeek-AI, R. Xie, K. Dong, et al. 2024. DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*.

Y.H. Liu, L. Wang, S. Chen, et al. 2023. Summary of ChatGPT-Related Research and Perspective Towards the Future of Large Language Models. *Meta-Radiology*.

M. Reid, N. Savinov, D. Teplyashin, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

A. Zeng, X. Liu, Z. Du, et al. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *arXiv preprint arXiv:2406.12793*.

P. Sahoo, A. Singh, S. Karn, et al. 2024. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *arXiv preprint arXiv:2402.07927*.

T. Dettmers, A. Pagnoni, A. Holtzman, et al. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv preprint arXiv:2305.14314*.

Y. Zhao, J. Huang, J. Hu, et al. 2024. SWIFT: A Scalable lightWeight Infrastructure for Fine-Tuning. *arXiv preprint arXiv:2408.05517*.

袁毓林. 2020. 叙实性和事实性：语言推理的两种导航机制. 语文研究, 0(1):1–9.

Tianchi Platform. 2025. CCL25-Eval 任务4：第一届中文叙实性推理评测（FIE2025）. https://tianchi.aliyun.com/competition/entrance/532342/introduction.

曾新红, 林伟明, and 明仲. 2008. 中文叙词表本体一致性检测机制研究与实现. 现代图书情报技术, (5):1–9.

袁毓林, 崔玉珍, 孙竞, and 游豪. 2023. 怎样构建面向事实性表达研究的法律专题语料库？当代修辞学, (2):16–28.