CCL25-Eval任务四系统报告：基于多策略知识融合的叙实性推理方法研究

李宏宇
北京师范大学
国际中文教育学院
lihongyu
@mail.bnu.edu.cn

杨智惠
北京师范大学
国际中文教育学院
yangzhihui
@mail.bnu.edu.cn

胡韧奋
北京师范大学
国际中文教育学院
irishu
@mail.bnu.edu.cn

摘要

FIE2025任务旨在使用大语言模型对文本及相关假设进行叙实性推理。我们参加了微调和非微调两个赛道，分别在人工数据集和自然数据集上采用提示词优化和词表RAG策略融合语言学知识，并利用模型集成投票方法提升判断准确率。评测结果显示，我们的方法在非微调赛道取得了0.9351的成绩，在微调赛道取得了0.9261的成绩，均位列第三名。

关键词： 叙实推理；提示工程；模型集成

System Report for Task 4 of CCL25-Eval: Research on Factivity Inference Method Based on Multi-Strategy Knowledge Fusion

Hongyu Li Zhihui Yang Renfen Hu
School of International Chinese Language Education, Beijing Normal University
{lihongyu, yangzhihui, irishu}@mail.bnu.edu.cn

Abstract

The FIE2025 task involves using large language models to perform factivity inference on texts and their hypotheses. We participated in both fine-tuning and non-fine-tuning tracks, applying prompt optimization and RAG with verb-type lexicons to integrate linguistic knowledge, and further using the model ensemble voting strategy to enhance the system performance. Evaluation results show that our approach achieved a score of 0.9351 in the non-fine-tuning track and 0.9261 in the fine-tuning track, ranking third in both.

Keywords: Factivity Inference , Prompt Engineering , Model Ensemble

1 引言

叙实性推理（Factivity Inference, FI）是一种与事件真实性判断有关的语义理解任务，主要涉及语言使用中事实性信息的表达。在语言交际中，说话者可以根据某些动词性语言成分（如“相信”“谎称”“意识到”）来判断其宾语补足语小句所描述的事件的真实性，如下例所示：

示例1：他们**意识到**局面已经不可挽回。
示例2：他们**没有意识到**局面已经不可挽回。
判断：示例1和示例2的陈述都可推断出“局面已经不可挽回”的事实。

第二十四届中国计算语言学大会论文集，第157页-第165页，济南，中国，2025年8月11日至14日。

中文叙实性推理评测（Factivity Inference Evaluation 2025, 简称FIE2025）旨在关注大语言模型(Large Language Models, 后文简称LLMs)的叙实推理能力，给定背景句和假设句，模型需要根据背景句判断假设句是否成立，并从“真”“假”“不能确定”三个选项中选择其一。任务要求探索不同的提示词来提升LLMs的中文叙实性推理表现，并根据是否微调模型分为两个赛道。

目前，基于大语言模型的叙实性推理研究较少，现有工作主要集中于理论探讨及个案分析。袁毓林 (2020a)根据叙实性将动词分为正叙实、反叙实和非叙实三类，并指出人们在大部分情况下能自然且轻易地推出宾语的真值，而有时则会十分困难。值得关注的是，部分动词会产生解释歧义或者语义模糊的情况。李新良和袁毓林 (2016)分析了反叙实词的适用条件，并以“假装”为例深入探讨歧义所在。此外，还有研究针对“忘记”类动词(袁毓林, 2020b)、“知道”(李新良和袁毓林, 2017)等词进行个案分析，探讨了多语境下的复杂情况。

大语言模型经过预训练和后训练，掌握了丰富的语言知识和世界知识，在语言理解与生成任务上有出色表现。目前，提示工程领域常采用多种策略提升模型表现，如基于示例的上下文学习(Brown, 2020)、思维链(Wei, 2022)、检索增强生成(Lewis, 2020)等。在本次评测中，我们结合了上述方法优化提示词，向模型注入语言学知识，并利用模型集成提高预测稳定性。最终，该方法在非微调赛道的准确率为0.9351，在微调赛道的准确率为0.9261，均位列第三。

2 非微调赛道

针对人工集，我们设计了一套基于规则的prompt引导大模型重点关注动词的叙实类型。对于自然集，我们首先将题目分为两类：其中，第一类题目中的动词叙实类型固定，我们构建了“动词-类型”对照词表，让大模型查找词表后基于规则回答；第二类题目较难，即动词叙实类型不固定，引入词表会误导模型判断，我们分类总结这些动词的用法，为模型提供决策参考。最后，我们在两个数据集上都使用了模型集成投票的方法，提升整体准确率。模型推理的参数除温度设置为0外，其余参数均为默认参数。

2.1 人工集方法

在实验中，为了探索哪种提示策略下模型表现最优，我们尝试了多种设置，包括：简单版提示词、思维链版提示词、基于动词类型+Few-shot版提示词、基于动词类型+否定词提示+Few-shots版提示词。

简单版提示词是用通俗易懂的语言对任务进行说明，不附示例，不说明任务细节，考验模型本身的推理能力，详见附录A。这样做可以初步定位模型能力，并发现对模型具有挑战性的难题以便进一步优化。**思维链版提示词**是在简单版的基础上，附一个题目的推理过程示例，并要求模型输出显式思维链，依照推理过程输出结果，详见附录B。**基于动词类型+Few-shots版提示词**会利用到人工集给出的`"predicate"`（动词）和`"type"`（动词类型）信息，由于动词类型（正叙实、反叙实、非叙实）与宾语的真实性有强相关性，所以我们在提示词中直接提取题目的`"type"`值，并且让模型遵循以下规则作答（该版提示词详见附录C）：

- 若动词为非叙实词，直接返回U。
- 若动词为正叙实词，且假设与动词后宾语相一致，返回T，否则返回F。
- 若动词为反叙实词，且假设与动词后宾语一致，返回F，否则返回T。

在验证集上应用此方法后，错例主要分为三类：(1)模型对规则或题目信息的误判；(2)部分动词的叙实情况较为复杂，如“哀叹”“假装”等词；(3)部分正叙实词前出现否定词时，叙实性可能发生改变，也可能不变。为了进一步优化，我们设计了**基于动词类型+否定词提示+Few-shot版提示词**，主要是针对第(3)类错例补充规则，这条说明在prompt中附在上述的动词类型规则之下：

> **注意事项**
> 当【正叙实词】前有否定词（如“没有”）时，内容真实性可能改变，主要分两种情况：
>
> 1. 【正叙实词】是施为动词时，否定使得内容是否是事实无法确定，答案为U；施为动词，即说出这个词的同时，就在进行某种行动，如宣布、保证、证实、证明、承认等；
>
> 2. 【正叙实词】不是施为动词时，否定不影响内容的真实性；
>
> 例*1*：”文章没有证明大气污染和地质活动有关系”→”证明”是施为动词，关系无法确定（U）
> 例*2*：”小张没有揭露小李是卧底”→”小李是卧底”仍为事实
> 注：规则不一定能包含所有情况，请仔细学习和参考，并结合语境做出合适的判断。

第二十四届中国计算语言学大会论文集，第157页-第165页，济南，中国，2025年8月11日至14日。

为了适配新的规则，我们从错例中选取四例作为示例，使其更具针对性，详见附录D。

2.2 自然集方法

自然集的难点有两方面：首先，自然集题目没有"type"信息，需要补充标注；第二，自然集涵盖较多动词类型不固定、用法灵活的题目，比如“感觉”在不同语境下，可以是非叙实词，也可以是正叙实词，这类题目无法通过标注动词类型解决，需要设计新的方法。

为了解决第一个问题，我们额外构建了一个包含动词类型的词表用于检索增强生成(Retrieval Augmented Generation，简称RAG)，模型可参考词表中的动词类型辅助判断。我们提取验证集的人工集中的动词及其类型作为基础词表，并从人工集中选取了部分动词作为种子，利用预训练中文词向量(Li et al., 2018)提取词义相似度高的动词作为补充，为进一步扩充词表，我们还使用大模型如Qwen-max、DeepSeek-V3生成了更多动词，经过人工筛选，得到动词和类型标签。由此，对于动词类型固定的题目，采用词表RAG+人工集prompt即可作答。

自然集中还存在许多动词类型不固定的题目，无法通过处理人工集的规则方法判断。通过误例分析，我们发现出错的题目集中于11个动词：以为、假装、哀叹、埋怨、声言、幻想、感叹、感觉、批评、数落、觉着，覆盖0510版自然集中214道题。于是，我们针对这十一个动词详细总结了其叙实用法，为模型提供参考。针对验证集中这批动词类型不固定的题目，推理能力较强的DeepSeek-R1模型使用基础版prompt（详见附录E）仅能达到0.64的准确率，而融合动词用法描述后，准确率达到0.79，证明了补充叙实用法描述的有效性。其中，动词“哀叹”的提示文档见附录F。

2.3 模型集成

进一步地，我们集成了多个模型，并通过投票选出最终答案，选用的模型和数据集参见表 1。平票时，在人工集上，选取验证集上成绩最好的Qwen-Max的预测作为答案；在自然集上，选取验证集上成绩最好的DeepSeek-R1的预测作为答案。

模型名	数据集
Qwen-max-2024-09-19	自然集/人工集
DeepSeek-R1-2025-01-25	自然集[1]
Claude-3.7 sonnet-2025-02-19	人工集[2]
GPT-4.1- 2025-04-14	自然集/人工集
Gemini-2.5 flash review-2025-04-17	自然集/人工集

Table 1: 非微调模型集成配置

3 微调赛道

综合考虑推理速度、性能及API开销，我们选择GPT-4.1-mini 2025-04-14进行微调试验。对于人工集，我们使用基于动词类型的提示词构造训练数据，强化模型对于规则的理解和遵循能力。对于自然集，我们使用简单版提示词加上示例构造训练数据。微调数据集取自验证集全集，其中，人工集共有300条微调训练数据，自然集共有700条微调训练数据。微调使用OpenAI平台，在两个数据集上均训练三轮，其余参数均为默认参数。在推理时，对于人工集,我们使用在非微调赛道表现最好的基于动词类型+否定词提示+Fewshot版提示词；对于自然集，我们使用了与非微调赛道相同的基础版prompt+词表RAG提示词。

4 实验结果与讨论

在非微调赛道，验证集的人工集上，四版提示词的对比结果参见表 2，使用的模型均为Qwen-Max。相较于简单版提示词，加入思维链会让准确率有所提升，但在第四版提示词中加入思维链后，我们发现推理速度会显著变慢，而且会导致模型过度依赖规则做题，泛化能力

[1]由于人工集题目较为简单，而DeepSeek-R1推理速度较慢，且适合复杂题目，因此其仅用于自然集预测。

[2]由于自然集题目数量较多，评测时间内未能完成Claude-3.7对全部自然集数据的预测，因此模型集成时仅使用了其在人工集上的预测结果。

下降，也会增加幻觉的概率，准确率不增反降，所以我们在第四版提示词中未使用思维链策略。

提示词版本	验证集准确率
简单版	0.82
思维链版	0.86
基于动词类型+Few-shot版	0.91
基于动词类型+否定词提示+Few-shot版	0.95

Table 2: 非微调赛道中不同版本提示词在验证集（人工集）上的准确率

最终，我们使用在验证集上表现最好的方法对测试集进行预测，在微调赛道和非微调赛道上的结果参见表 3，均位列第三。值得一提的是，我们的方法在人工集上表现最优，非微调赛道上的准确率为0.9817，在所有队伍中排名第一。

赛道	自然集准确率	人工集准确率	总得分
非微调赛道	0.9175	0.9817	0.9351
微调赛道	0.9148	0.9560	0.9261

Table 3: 测试集评分结果

需要指出的是，结合验证集误例分析，我们发现，虽然人工集整体表现较好，但模型仍会在部分问题上“粗心”误判，对否定词出现的两种情况无法准确处理。对于自然集来说，虽然融合详细的动词叙实用法描述能让模型在动词类型不固定题目上的准确率得到提升（0.79），但仍有较大改进空间，一方面需要对动词的多种用法进行更准确的归纳，另一方面是要用更加合适的用语描述规则，以更好地向模型传达信息，最大化其推理能力。

5 总结及展望

我们同时参加了FIE2025叙实性推理评测任务的微调和非微调赛道，针对中文叙实性推理问题，以提示词优化为主要路线，通过逐步分析错例，运用了词表RAG、提示文档RAG和模型集成投票等方法提升判断准确率，在非微调赛道得分0.9351，微调赛道得分0.9261。

展望未来，针对动词类型不固定的情况，有必要进一步优化处理方法。当前方法主要依赖于总结高频错例并编写针对性提示，但该方式在新词泛化和自动识别方面存在不足。后续研究可结合语言学理论，深入系统地分析动词类型，探索自动识别类型不固定动词的有效途径，并进一步挖掘其语义特征和用法。此外，还可尝试将丰富的语言学知识通过微调、强化学习等机制嵌入模型，提升模型对复杂语言学现象的处理能力。

参考文献

Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert van Rooij, Kun Zhang, Zhouchen Lin. 2025. Empowering llms with logical reasoning: a comprehensive survey. arXiv preprint arXiv:2502.15652.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 9459–9474.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical Reasoning on Chinese Morphological and Semantic Relations. In *Proceedings of the 56th Annual Meeting of the*

Association for Computational Linguistics (Volume 2: Short Papers), pages 138–143, Melbourne, Australia. Association for Computational Linguistics.

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

李新良, 袁毓林. 2016. 反叙实动词宾语真假的语法条件及其概念动因. 当代语言学, 18(02):194–215.

袁毓林. 2020b. “忘记”类动词的叙实性漂移及其概念结构基础. 中国语文, (5):515–526+638.

李新良, 袁毓林. 2017. “知道”的叙实性及其置信度变异的语法环境. 中国语文, (1):42–52+127.

附录

附录A. 简单版提示词

简单版提示词
任务角色
你是一名自然语言推理专家。你的任务是根据给定的"句子"和"假设"，判断它们之间的关系，并从下列选项中选择一个：
- T（蕴含）：假设可以从句子推断出来。
- F（矛盾）：假设和句子内容矛盾。
- U（中立）：句子既不能支持，也不能否定假设。

—
输入结构
你会收到一个字典条目，包含以下字段：
- "句子"：待分析的完整句子。
- "动词"：句子中的关键动词。
- "动词类型"：该动词的语义类型，分为正叙实、反叙实、非叙实三种。
- "假设"：需要判断的命题。

输出
请你直接输出单个选项，不要输出其他内容

附录B. 思维链版提示词

思维链版提示词
任务角色
你是一名自然语言推理专家。你的任务是根据给定的"句子"和"假设"，判断它们之间的关系，并从下列选项中选择一个：
- T（蕴含）：假设可以从句子推断出来。
- F（矛盾）：假设和句子内容矛盾。
- U（中立）：句子既不能支持，也不能否定假设。

—
输入结构
你会收到一个字典条目，包含以下字段：
- "句子"：待分析的完整句子。
- "动词"：句子中的关键动词。
- "动词类型"：该动词的语义类型，分为正叙实、反叙实、非叙实三种。
- "假设"：需要判断的命题。

请基于以下规则做出判断：
- 若动词为正叙实词，且假设与宾语相同，返回T，不同则返回F；
- 若动词为反叙实词，且假设与宾语相同，返回F，不同则返回T；
- 若动词为非叙实词，直接返回U。

示例："句子"：我抱怨自己赚的太少。，"假设"：我赚的很多。"动词"：抱怨"动词类型"：正叙实
答案及原因：由于该句中“抱怨”为正叙实词，而且假设与句中宾语相反，故可判断假设与整个句子违背，假设陈述的一定不是事实，所以返回F
请参照上面的过程，输出你所推理的过程和得出答案的依据，并把最终的答案以字母形式呈现(T/F/U/R)

附录C. 基于动词类型+Few-shot版提示词

基于动词类型+Few-shot版提示词
您是一个语言分析助手，任务是分析句子中目标动词和其后宾语的叙实关系，判断假设是否成立。动词的类别规则如下：
1. 正叙实词（如“意识到”、“猜到”、“披露”、“看见”、“抱怨”）一般表示宾语描述为事实且发生。示例：“他猜到我已经迟到了”，“我迟到了”是事实
2. 反叙实词（如“妄称”、“谎称”、“污蔑”）一般表示宾语描述的一定不是事实，一定未发生。示例：“我吹嘘我可以搞定这件事”的“我能搞定这件事”是反事实，实际上我并不能搞定这件事；
3. 非叙实词（如“估计”、“认为”、“相信”）一般表示无法确定宾语描述的事实性。示例：“我猜测小王喜欢小丽”中的“小王喜欢小丽”的事实性是无法推测的。
在任务中会提供动词的类别信息，标记为type，你需要利用这个信息做后续判断。
你需要遵循以下规则做判断：
如果动词是【非叙实词】，直接返回U。
如果不是，你还需要判断【假设】和【宾语】之间是否相同：
-若动词为正叙实词，且假设与宾语相同，那么返回T，不同则返回F；
-若动词为反叙实词，且假设与宾语相同则返回F，不同则返回T。
比如“我抱怨自己赚的太少”，“抱怨”为正叙实词，若假设为“我赚的很少”，那么返回T。若假设为“我赚的很多”，那么应当返回为F。
示例1：
"动词类型": 正叙实
"动词": 表明
"句子": 小张的做法表明感情出现了问题。
"假设": 感情出现了问题。
答案：T，因为“表明”是正叙实词，而且描述的事情和假设相同。
示例3：
"动词类型": 反叙实
"动词": 谎称
"句子": 小张谎称不会帮助小李。
"假设": 小张不会帮助小李。
答案：F，因为“谎称”是反叙实词，描述的事情和假设相同，表明假设一定不会发生。

附录D. 基于动词类型＋否定词提示＋Few-shot版提示词

基于动词类型＋否定词提示＋Few-shot版提示词
您是一个语言分析助手，任务是分析句子中目标动词和其后宾语的叙实关系，判断假设是否成立。动词的类别规则如下：
1. 正叙实词（如“意识到”、“猜到”、“披露”、“看见”、“抱怨”）一般表示宾语描述为事实且发生。示例：“他猜到我已经迟到了”，“我迟到了”是事实
2. 反叙实词（如“妄称”、“谎称”、“污蔑”）一般表示宾语描述的一定不是事实，一定未发生。示例：“我吹嘘我可以搞定这件事”的“我能搞定这件事”是反事实，实际上我并不能搞定这件事；
3. 非叙实词（如“估计”、“认为”、“相信”）一般表示无法确定宾语描述的事实性。示例：“我猜测小王喜欢小丽”中的“小王喜欢小丽”的事实性是无法推测的。
在任务中会提供动词的类别信息，标记为type，你需要利用这个信息做后续判断。
你需要遵循以下规则做判断：
如果动词是【非叙实词】，直接返回U。
如果不是，你还需要判断【假设】和【宾语】之间是否相同：
-若动词为正叙实词，且假设与宾语相同，那么返回T，不同则返回F；
-若动词为反叙实词，且假设与宾语相同则返回F，不同则返回T。
比如“我抱怨自己赚的太少”，“抱怨”为正叙实词，若假设为“我赚的很少”，那么返回T。若假设为“我赚的很多”，那么应当返回为F。
注意事项*
- 当【正叙实词】前面出现否定词（如“没有”）时，该词后面内容的真实性可能发生改变，具体分为两种情况：
（1）【正叙实词】是施为动词时，否定使得内容是否是事实无法确定，答案为U；
施为动词，即说出这个词的同时，就在进行某种行动，如宣布、保证、证实、证明、承认等；
如“文章没有证明大气污染和地质活动有关系”，其中，“证明”是施为动词，“大气污染和地质活动”是否有关系是无法确定的；
（2）【正叙实词】不是施为动词时，否定不影响内容的真实性；
如“小张没有揭露小李是卧底。”中“小李是卧底”仍然是事实。
这两种规则不一定能包含所有情况，请你仔细学习和参考，并结合语境做出合适的判断。
示例1：
"动词类型": 正叙实
"动词": 哀叹
"句子": 小张哀叹经济发展将会停滞。
"假设": 经济发展将会停滞。
答案：U，虽然“哀叹”是正叙实词，但小张哀叹的是对未来的主观判断，无法确定经济发展是否将要停滞
示例2 "动词类型": 正叙实
"动词": 意识到
"句子": 小张没有意识到小李没哭。
"假设": 小李哭了。
答案: F，尽管小张没有意识到小李没哭，但是小李哭了是一个客观事实，动词仍是【正叙实词】，宾语为“小李没哭”，与假设不同因此返回F。
示例3：
"动词类型": 正叙实
"动词": 承认
"句子": 小张不承认小李没打过小王。
"假设": 小李打过小王。
答案：U，承认为【施为动词】，此处小张不承认小李没打过小王，不能判断出小李是否能打过小王。
示例4：
"动词类型": 正叙实
"动词": 证明
"句子": 文章没有证明大气污染和地质活动没有关系。
"假设": 大气污染和地质活动没有关系。
答案：U，证明为【施为动词】，此处文章未证明大气污染和地质活动的关系，不能推出二者一定没关系，也不能推出二者有关系，所以返回U

附录E. 自然集基础提示词

自然集基础提示词
请分析：根据“句子”的描述分析“假设”中的内容是否是事实。
选项说明
- 输出
T：“假设”中的内容一定是事实。示例：句子：“他猜到我已经迟到了”,假设：“我迟到了”，是事实，返回T;
F：“假设”中的内容一定不是事实，一定未发生。示例：“句子”：“我吹嘘我可以搞定这件事”，“假设”：“我能搞定这件事”，违背事实，实际上我并不能搞定这件事，返回F;
U：无法确定“假设”中的内容是不是事实。示例：“句子”：“我猜测小王喜欢小丽”，“假设”：“小王喜欢小丽”，不一定是事实，返回U。

数据格式
- 输入格式
"动词": 感叹,
"句子": 民生银行的招股说明书共213页，约20万字左右，连券商都感叹，这是他们做过的最长的招股说明书。
"假设": 这确实是他们做过的最长的招股说明书。
- 输出格式
仅返回大写字母，如：T

附录F. “哀叹”提示文档

“哀叹”提示文档
【哀叹】在不同语境下情况不同
-哀叹的对象是【主观判断】内容时，其真实性大概率无法判断，如“英国的一个组织却在哀叹英国媒体存在着严重的性别失衡。”，该媒体是否存在“严重的性别失衡”是无法判断的；如“他们为此气愤，哀叹他们死得冤枉。”，死的是否冤枉在该语境下也是无法判断的；如“我们一方面，总是哀叹当今时代出不了文艺大师...”，”当今时代出不了文艺大师“是【主观判断】，无法判断真实性，应返回U。
-哀叹的对象是基于【自身状况】的判断时，大概率是真实的，因为自己对自身的状况的判断是可信的，如“她的灵活步伐也让塞莱斯频频受骗,哀叹自己腿脚不灵活。”，“自己的腿脚不灵活”可认为是事实。
-哀叹的对象是【客观事实】时，大概率是真实的，比如“所有的领导都对互联网有所忌惮，哀叹没有网络的好日子一去不复返了...”，是对没有网络的好日子的哀叹，说明没有网络的好日子确实是一去不复返。
提示可能无法包括全部情况，因此请你【参考】以上提示，并判断出合适的答案。