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[1]，杨智惠[1]，胡韧奋[1]

1.北京师范大学，国际中文教育学院

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,其中，非微调赛道人工集得分为0.9817，位列第一。

背景

- 任务目标：评估大语言模型（LLMs）的中文叙实推理能力。
- 任务内容：给定背景句+假设，LLMs需判断假设：真/假/不能确定。
- 核心挑战：优化提示词（Prompt）提升LLMs表现。
- 赛道设置：非微调赛道/微调赛道。

1. 他们**意识到**局面已经不可挽回。
   --假设：局面确实不可挽回。
2. 小张**估计**今年的棉花产量会增加。
   --假设：今年棉花产量会增加。

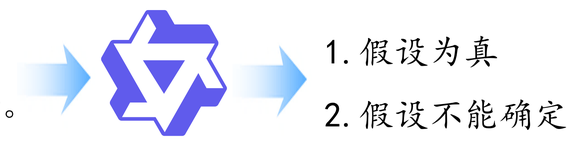

图1. 任务流程示例。LLMs需根据背景句和动词来判断宾语补足语小句所描述的事件的真实性，进而判断假设的成立情况。图中两例答案分别是“假设为真”与“假设不能确定”。

非微调赛道

我们注意到**动词本身会带有叙实性信息**，预设了宾语补足语小句的真实性。基于这种特性，可以把动词分为三类，分别是“正叙实词”“反叙实词”和“非叙实词”。

否定词干扰：前置否定词可能会使**正叙实词**的预设失效。

- “意识到”和“没意识到”都预设小句为真。
- “宣布”预设小句为真；“没宣布”则会使真实性无法确定。

人工集数据特性：人工数据集本身带有动词类型标签，可利用此信息和几类规则帮助LLMs做判断，如图2。

图2. 背景句中动词的类型与宾语补足语的真实性之间的对应关系。

针对人工集，我们共设计了四个版本的提示词：

- 简单版提示词：只用通俗易懂的语言描述任务。
- 思维链版提示词：要求LLMs显式输出思维链，根据推理过程作答。
- 基于动词类型+Few-Shot版提示词：让LLMs根据题目中的动词类型信息结合示例进行作答。
- 基于动词类型+否定词提示+Few-Shot版提示词：在上一版提示词的基础上，补充“否定词干扰” 的判断规则。

自然集数据特性：（1）无动词类型标签。（2）存在动词类型不固定的题目。处理方法如下：

- 我们构建了**动词-类型参照词表**，解决无标注问题。
- 提取出类型不固定的**特殊动词**，总结规律与用法写成**提示文档**，利用RAG方法为LLMs提供参考。
- 设计了一种多策略的分类Prompt方法，如图3（b）。

模型集成：对人工集和自然集各选四个主流LLM，投票表决。

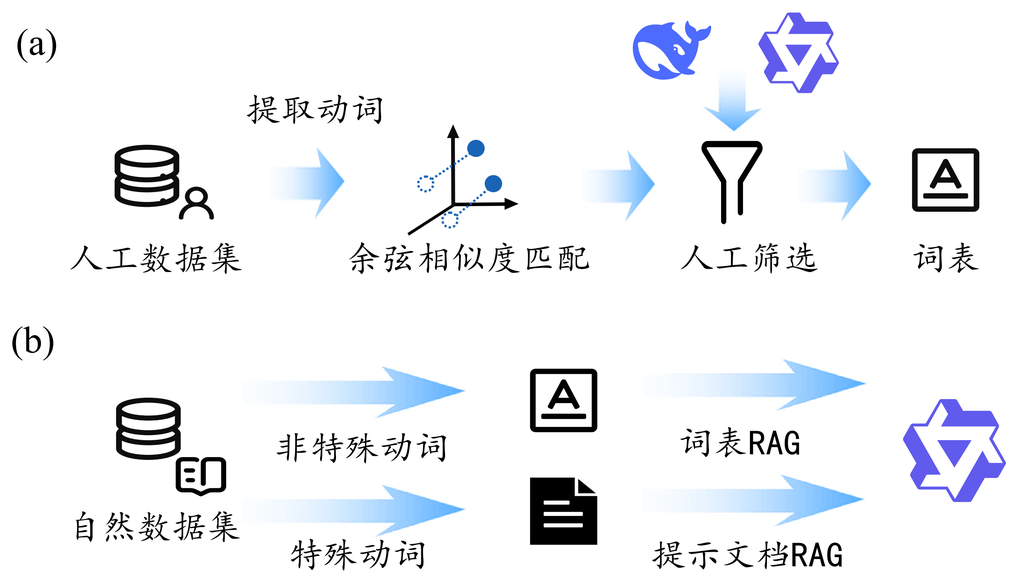

图3. (a) 词表构建过程。从人工数据集中提取动词，作为种子输入中文预训练词向量模型，在余弦相似度较高的候选词中人工筛选；作为补充，额外让LLMs生成一批带标签的候选词并人工筛选。(b) 分类Prompt构建过程。对于非特殊动词，使用词表RAG+人工集Prompt进行构建；对于特殊动词，使用提示文档RAG+基础版Prompt进行构建。

微调赛道

模型选择： GPT-4.1-mini 2025-04-14。

人工集： 训练：基于动词类型版提示词。
推理：基于动词类型+否定词提示+Few-Shot版提示词

自然集： 训练：简单版提示词+Few-Shot。
推理：多策略分类Prompt方法。

结果与讨论

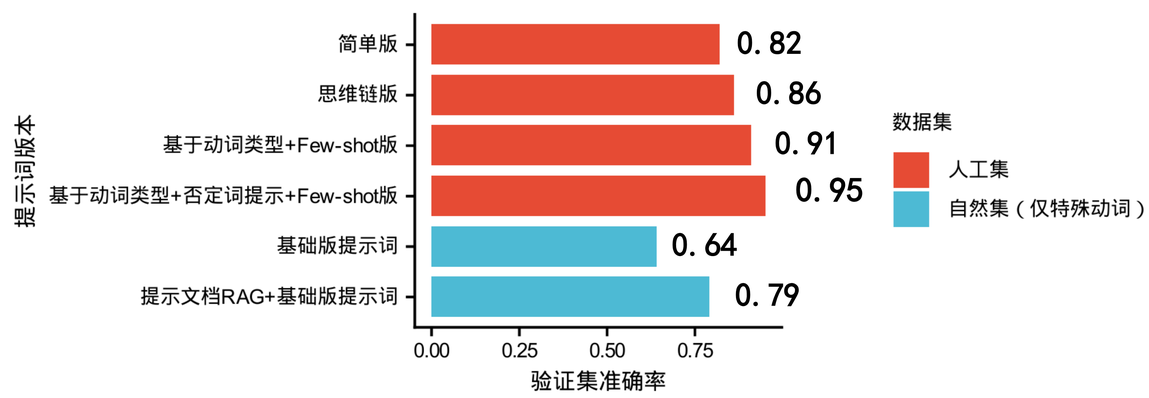

图4. 各版本提示词在非微调赛道中的表现。此结果中人工集使用模型为Qwen-Max，自然集（仅特殊动词）选用模型为DeepSeek-R1。

在人工集上，我们的方法使模型准确率最高达到了95%；在自然集（仅特殊动词）中，提示文档RAG的方法能让模型准确率由64%提升到79%，均证明了方法的有效性。自然集的表现仍有提升空间，在未来可进一步归纳动词类型的规律，探索自动识别特殊动词及动词类型的路径，以提高模型面对新词时的泛化能力。