CCL25-Eval任务4系统报告：基于叙实性分类和语境特征的大语言模型叙实性推理

张箫驿, 鲁嘉琪, 张达, 陈笑宇, 卢达威*
中国人民大学文学院
北京, 100872
2023103542@ruc.edu.cn，2484774735@qq.com，zhangda2398@ruc.edu.cn
chxy1993@sina.cn，wedalu@163.com

摘要

叙实性推理是机器理解文本隐含事实的关键能力之一，核心在于结合动词的语义判断动词宾语命题的真值。本研究基于首届中文叙实性推理评测任务4（FIE2025）开展叙实性推理研究，经过前期对不同模型的测验和比对，选择了Deepseek-R1模型为基座模型。提示语的总体撰写思路是：首先将动词叙实性进行分类，从传统的三分法扩展至五分法（叙实、弱叙实、反叙实、非叙实、半叙实），同时，对自然语料与人造语料进行差异化处理，再针对部分语义复杂的动词编写更加细致的判断规则。最终结果显示，自然语料的正确率达到0.9155，人造语料的正确率为0.9541，总正确率达到0.9261。

关键词： 叙实性推理；大语言模型

System Report for CCL25-Eval Task 4: Factuality Reasoning in Large Language Models via Factual Consistency Classification and Contextual Features

Zhang Xiaoyi, Lu Jiaqi, Zhang Da, Chen Xiaoyu, Lu Dawei*Corresponding author
School of Chinese Language and Literature, Renmin University of China
No.59 Zhongguancun Street, Haidian District, Beijing, Postal Code: 100872
2023103542@ruc.edu.cn，2484774735@qq.com，zhangda2398@ruc.edu.cn
chxy1993@sina.cn，wedalu@163.com

Abstract

Factuality reasoning is one of the critical capabilities for machines to comprehend implicit factual information in texts, with its core challenge lying in determining the truth value of verbal complement propositions through semantic analysis of verbs. This study, based on Task 4 of the inaugural Chinese Factuality Inference Evaluation (FIE2025), conducts systematic research on factuality reasoning. After preliminary testing and comparative analysis of multiple models, we selected the Deepseek-R1 model as the foundation model. The prompt design strategy involves: first refining the classification of verbal factivity from the traditional tripartite system to a five-category framework (factual, weakly factive, counterfactive, non-factive, and semi-factive); implementing differentiated processing for natural and artificially constructed corpora; and developing granular judgment rules for verbs with semantically complex profiles. Final results demonstrate accuracy rates of 0.9155 for natural corpora, 0.9541 for artificial corpora, achieving an overall accuracy of 0.9261.

Keywords: Factuality Reasoning , Large Language Models

*通讯作者

1 引言

叙实性推理是根据谓语语义推理出宾语命题的真假性。语句的谓语动词中透露了说话人在命题的事实性或主观态度上的倾向性，即动词本身具有一种词汇预设，谓语的语义结构中有明确指引其宾语所指命题的真假的信息。因此，说话人能够通过选用不同的主句谓语动词来规定宾语小句的真值情况。（袁毓林，2020）例如，对于"他们意识到局面已经不可挽回"这句话，我们首先锁定这个句子中的谓语动词，即"意识到"，随后根据该词的语义来判断出"局面已经不可挽回"这一命题的真值为真。如果将该句的谓语动词进行变换，从"意识到"变成"幻想"，那么"局面已经不可挽回"的命题便是错误的。

人类内在的语言能力使得人们能够抓住语言表达的形式线索来进行真值判断和语义推理，但是大语言模型对部分谓语动词的语义理解不到位，在面对灵活多变的语用环境时，有时难以对叙实性推理做出正确的判断。

第一届中文叙实性推理评测（FIE2025）所关注的主题是大语言模型对中文叙实性推理的能力。此次测评分为了微调赛道和不微调赛道，我们所参加的是不微调赛道。通过编制提示词，调用所选大模型的API，将测评语料中"text"字段值（即背景句）作为依据来判断"hypothesis"字段值（即结论句）的真值情况，并记录模型的返回结果。此次比赛的训练语料和测评语料均分为自然语料和人造语料两部分，并将整体的正确率作为最终的测评指标。

本次测评中，我们参加不微调赛道评测，并将Deepseek-R1作为基座模型，从细化叙实性分类和有针对性的语境特征描写两个方向优化prompt来提高正确率。之所以选用Deepseek-R1，是因为该模型基于回归增强（Regression-enhanced）的训练方法，专注于数学推理、代码生成和逻辑推理任务。相比于Deepseek-V3，该模型更强调推理能力，更加契合本次测评任务的需求。经过测试，最终自然语料判断正确率为0.9155，人造语料为0.9541，总正确率达0.9261。

2 相关工作

目前专门针对大语言模型（LLM）叙实性推理能力的研究不多，相关的工作主要包括以下四类。第一类是叙实性推理的逻辑基础。Allwood et al.（1977）从逻辑学的视角分析了不同语言逻辑特性，为后续比较不同语言的叙实性表达提供了参考；Kato & Kato（1997）对否定极性现象进行研究，为后续LLM理解语境中的否定语义研究奠定基础；Leech（1983）从情感意义等角度进行探讨，强调关注结合语境进行分析。第二类是对自然语言本身的叙实性研究，主要包括Kiparsky& Kiparsky（1970）从叙实性的角度将英语谓词分为叙实（factive）谓词和非叙实（non-factive）谓词；Leech（1983）把谓词分为叙实谓词、非叙实谓词和反叙实（counter factive）谓词三类；袁毓林、李新良、寇鑫等进行了一系列立足于汉语的叙实性研究（李新良,袁毓林，2016等）。第三类是数据集建设工作，即通过标注事件真值状态构建用于后续研究的基准数据集。代表性工作有Saurí, R., & Pustejovsky（2009）基于TimeBank构建了首个大规模事件事实性标注语料库FactBank；曹媛等（2013）以ACE（Automatic Content Extraction）2005中文语料库为基础语料库，标注其中Movement事件的事实性。第四类是在自然语言处理任务中的应用，即利用叙实性在语言形式方面的线索辅助完成自然语言处理任务。李新良，袁毓林（2013）在建立汉语动词蕴涵型式库时，尝试利用动词的叙实性进行相关句子内部的语义推导，从而实现蕴含识别；Jeretič（2020）通过构建IMPPRES数据集评估自然语言推理（NLI）模型对隐含事实的捕捉能力，进而为提高LLM事实性推理能力提供方向。

3 方法与结果

本文将实验结果放在每个方法改进之后，以展现提示语优化过程所取得的进步。提示语优化主要分为两个方向：叙实性动词分类的优化和基于语境特征的规则细化。在提示语撰写的过程中，还使用了few-shot的技术方法。值得注意的是，本文的行文顺序即是我们的实验顺序，其中3.1.1节和3.1.2节是公布测试集前的实验，我们使用的测评语料是700个自然语料及其叙实性判定答案；3.1.3节及后文是公布测试集后的实验，我们使用的是主办方正式发布的1988句测评语料（包括自然语料和人工语料）和官方测评结果。

3.1 叙实性动词分类的优化

这一阶段的思路是不断优化叙实性动词分类，从基础的三分法，扩展到四分法和五分法。

第二十四届中国计算语言学大会论文集，第96页-第104页，济南，中国，2025年8月11日至14日。

3.1.1 三分法：对叙实性推理的基线测定和基座模型的选择

在这一阶段，我们主要考察训练语料，根据袁毓林（2020）将谓语动词分为了最基本的三类：叙实动词、非叙实动词和反叙实动词。测评所涉及动词的叙实性具体分类如下：

动词类型	定义	代表词
叙实动词	主语（也可以同时是说话人）意识到其宾语所陈述的某种事态是一种事实，即预设其宾语所指命题为真。	暴露出、表明、猜到、诧异、承认、得知、发现、反映出、感觉到、感叹、高兴、后悔、获悉、记得、觉察出、觉察到、觉出、揭露、看出、看到、看见、领悟到、明白、目睹、碰见、披露、瞥见、瞧了、清楚、听出、听到、听见、望见、闻到、羡慕、想到、意识到、遇见、证明、证实、知道、注意到、庆幸、认识到、忘记、预见到、哀叹、抱怨、埋怨、批评、数落
非叙实动词	主语（也可以同时是说话人）意识到其宾语所陈述的某种事态不是一种事实，而是一种假象或幻觉，即预设其宾语所指命题为假。	猜、猜测、感觉、估计、怀疑、觉得、觉着、认为、声明、声言、听说、相信、想、重申、声称、说、推算、扬言
反叙实动词	主语所指所持有的某种信念，却并不承诺或明示其宾语所指命题为真，也不预设或暗示其宾语所指命题为假。	吹嘘、幻想、谎称、假装、妄称、污蔑、诬陷、想象、以为、装作、污陷

Table 1: 本测评所涉及动词的叙实性分类

实验时，我们首先在系统提示语中，告知大语言模型三类动词叙实性类型的定义和若干示例；在含有待测试样本的用户提示语中，我们以"是否告知大模型谓语动词的叙实性类型"为研究变量。告知谓语动词类型，表示我们会告知大模型当前待测试样本谓语动词的叙实性类型，大模型可据此对"hypothesis"字段的真值进行判断；不告知谓语动词类型，则表示让大模型自行对谓语动词的叙实性进行分类，再判断测试样本中"hypothesis"字段真值。因为Deepseek-V3和R1模型的侧重点不同，且两个模型均可以对Temperature（温度参数）进行调节，从而控制生成文本的随机性。因此，我们一并进行了基座模型（Deepseek-V3和R1）以及参数（temperature）的比较，结果如下表所示：

方法+模型	Temperature=0	Temperature=0.3
不告知谓语动词叙实性类型+V3模型	72.1	71.8
告知谓语动词叙实性类型+V3模型	85.29	84.86
告知谓语动词叙实性类型+R1模型	**87.84**	**87.2**

Table 2: 大语言模型对动词叙实性类型的判断和模型选择

结果显示，明确告知动词的叙实性分类可以显著提升大语言模型的叙实性推理能力。在处理逻辑推理问题时，低temperature数值能够提高推理的正确率。因此，本研究在此基础上进一步细化了对动词分类的研究。模型选择上，我们选择DeepSeek-R1作为基座模型且temperature设为0。由于R1运算耗时较长，在实验中，我们先用V3进行提示语优化，最后再迁移到R1上。

3.1.2 四分法：基于客观性差异的叙实性分类

我们发现，现有的三分法不能揭示部分谓语动词之间的差异。比如"哀叹、抱怨、埋怨、批评、数落"这五个词最初被归为叙实动词，但在实际运用过程中我们发现：该类动词之后的命题的真值受宾语内容客观性的影响，有时答案为"不能判断"，所以需要新增叙实性类型。袁毓林（2014）中提到了"半叙实动词"的概念，指只有肯定式或否定式时才预设宾语所表示的命题是一个事实的动词。上述五个词基本满足这些特点，在此基础上，我们还发现，当宾语内容为客观事件时，无论肯定式还是否定式宾语命题均为真。因此我们在原有三分法的基础上，增设"半叙实动词"，并根据我们的研究进行了重新定义，具体定义和动词归类如下：

动词类型	定义	代表词
半叙实动词	当宾语的内容是观点、态度等主观事件时，宾语小句的命题真值为"不能判断"；当宾语的内容是动作、行为等客观事件，或者动词前有否定成分时，宾语小句的命题为真。	哀叹、抱怨、埋怨、批评、数落（原归为叙实动词）

Table 3: 半叙实动词的定义和代表词

与三分法的测评结果进行对比，四分法的测评结果如下：

方法+模型	正确率
三分法+V3模型	85.29
四分法+V3模型	**86.84**

Table 4: 三分法和四分法的测评结果对比

3.1.3 五分法和对应关系提示

在四分法的基础上，我们发现肯否定对于动词叙实性的判断存在较大影响，具体分为以下两种情况。第一，部分叙实性动词宾语命题的真值受到动词否定的影响，需要新增叙实性类型。袁毓林（2014）将只有肯定或否定形式表达事实的动词均归为半叙实动词，本文对语料的考察发现，部分动词只有肯定形式表达事实，而部分动词只有否定形式表达事实，为明确区

分，避免大模型混淆，本文在前文半叙实动词的基础上增设了弱叙实动词，具体定义和动词归类如下：

动词类型	定义	代表词
弱叙实动词	弱叙实动词前紧接着”没”、”没有”、”不”、”未”等否定成分时，则无法对宾语小句的命题真假进行判断，为不能判断；当该弱叙实动词前没有否定成分或者前一个词语不是否定词时，则可以对宾语小句的命题真假进行判断，为真。注意，否定成分只考虑紧接在弱叙实动词前的否定成分。	承认、看出、遇见、证明、证实、表明（原归为叙实动词）

Table 5: 新增第五类叙实性类型——弱叙实动词

第二，受到肯否定的影响，大语言模型在宾语小句和假设之间对应关系的判断方面存在困难，当宾语小句和假设的肯否定形式存在差异时，大语言模型有时会出现错误。针对这一问题，我们在提示语中增设了宾语小句和假设之间对应关系的提示。如下：

你需要注意”text”里宾语小句和”hypothesis”之间的对应关系，举一个例子：”predicate”为”预见到”，”text”为”小张没有预见到第二天会下雨。”，根据叙实性类型，预见到是叙实动词，其肯定式和否定式都预设宾语所表示的命题是一个事实，所以”第二天会下雨”是一个事实，如果”hypothesis”为”第二天会下雨”，则为真，如果”hypothesis”为”第二天不会下雨”，则为假。

在五分法和对应关系提示的基础上，我们对主办方的2038句测试样例进行了测评，包括自然语料和人造语料。经过测验发现，自然语料更适合四分法，而人造语料因谓语动词前出现否定的情况更多，更适合五分法。因此，我们以V3模型的实验结果为基础，在R1模型中，当测试语料为自然语料则使用四分法作为提示模板，当测试语料为人造语料则使用五分法作为提示模板，结果如下：

模型\方法	四分法	五分法
V3模型	Nat_acc:**0.8858** Art_acc:0.8943	Nat_acc:0.8838 Art_acc:**0.9086**
R1模型	Nat_acc:**0.8899**	Art_acc:**0.9301**

Table 6: 四分法和五分法的测评结果对比

3.2 基于语境特征的规则细化

在对动词进行叙实性分类后，虽然能够高效地提高模型判断的正确率，但是部分动词因其复杂的语义和语用环境，对应的正确率有很大的提升空间。对于这一问题，我们采取了两种思路。

3.2.1 基于语境特征对动词叙实性分类进行校正

针对在自然语料和人工语料中语境差异明显的词语，我们采用了差异化归类的方法。半叙实动词的叙实性受到后续宾语内容的影响，但大语言模型对于这一问题的判断存在困难。3.1.3的结果也表明自然语料和人工语料存在系统性的差异。在这种情况下，针对这两类语料的特点将同一个词分别归为不同的类别可以有效解决大语言模型的缺陷。因此，我们对自然

语料和人工语料中五个半叙实词后续的宾语类型进行了对照分析，发现部分半叙实词后续宾语的类型差异明显，分析结果和修改情况如下所示：

词语	宾语客观性类型（自然语料）	宾语客观性类型（人工语料）	动词类型修改结果（自然语料）	动词类型修改结果（人工语料）
批评	主观事件较多	主客观频率差异不明显	非叙实	半叙实（原类型）
埋怨	主客观频率差异不明显	客观事件较多	半叙实（原类型）	叙实

Table 7: 半叙实动词宾语语境特征的分析结果和动词类型修改结果

3.2.2 对复杂动词建立专门知识库

针对在自然语料和人工语料中语境差异不明显的词语，我们采用了人工分析具体语境并总结规律的方法。在经过审查后，我们筛选了”表明、怀疑、觉着、感叹、哀叹、感觉”进行具体语用环境分析，并有针对性地建立了专门的知识库和提示语。

以”感觉”为例，我们专门细化了该词对应命题的判断标准，将语境分为五种(详见:附录A. 感觉的提示语模板)，每种语境对应的判断结果如下表所示：

语境情况	判断结果
语境中只是某个人单纯地表达意见。	真值无法判断
句子里明确表达了说话人在运用”夸张”的手法，或者表达的内容是夸张的。	真值无法判断
当”感觉”之后的小句主语是人，且和”感觉”的主语不一致。	真值无法判断
上下文有相关的证据来支持”感觉”的内容，可以对”感觉”的内容提供证据，则说明”感觉”之后的内容是正确的命题。	真值可以判断，宾语小句命题为真
”感觉”后的命题与明确的客观事实相关联。	真值可以判断，宾语小句命题为真

Table 8: ”感觉”的语境分类和判断结果

综合前期对动词的叙实性分类和后期语境特征对叙实性影响的研究，本文模型最终的提示语模板如下：

语料类型	提示语策略
复杂动词语料	该类动词包括”哀叹、表明、感觉、感叹、怀疑、觉着” 1.构建专门的叙实性知识提示模板： 2.”感叹、怀疑”模板仅用于自然语料； 3.”哀叹、表明、感觉、觉着”模板用于自然语料和人造语料
其他自然语料	四分法（叙实、反叙实、非叙实、半叙实）

第二十四届中国计算语言学大会论文集，第96页-第104页，济南，中国，2025年8月11日至14日。

语料类型	提示语策略
其他人造语料	五分法（叙实、弱叙实、反叙实、非叙实、半叙实）

Table 9: 本文模型最终提示语模板

最终自然语料判断正确率提升至**0.9155**，人造语料为**0.9541**，总正确率达**0.9261**。

4 讨论

在测评过程中，我们发现prompt对于大语言模型的引导有着重要的作用。四分法和五分法的配合运用能够高效地帮助大语言模型做出判断。在处理部分语义复杂的词上需要结合具体语境，但大语言模型的语境理解能力依然较弱，需要提供明确的具有形式特征的判断规则。因此面对这部分词，我们所采用的方案是分析这类词出现的语境，并从中抽取出一定的规律、细化判断规则，这有助于提升大模型判断的准确率。

参考文献

Allwood, S.Jens, Lars-Gunnar Andersson&sten Dahl. 1977. Logics in Linguistics. Cambridge: Cambridge University Press.

Jeretič, P., Warstadt, A., Bhooshan, S., & Williams, A. 2020. Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition. *Annual Meeting of the Association for Computational Linguistics*.

Kato, Natsuko&Yasuhiko Kato. 1997. Negative Polarity: A Comparative Syntax of English, Japanese, and Spanish. Paper presented at the 16th international Congress of Linguistics, Paris, July 21, (section 11).

Kiparsky, Paul and Carol Kiparsky 1970. Fact. In Bierwisch, Manfred, Karl Erich Heidolph (eds.). Progress in Linguistics, The Hague: Mouton, 143-173.

Leech, G. 1983. Semantics: The Study of Meaning, 2nd edition. Harmondsworth: Penguin Books.

Levinson, Stephen. 1983. Pragmatics. Cambridge: Cambridge University Press.

Saurí, R., & Pustejovsky, J. 2009. FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227-268.

曹媛,朱巧明,李培峰. 2013. 中文事件事实性信息语料库的构建方法. 中文信息学报, 27(06):38-44.

寇鑫,袁毓林. 2018. 汉语叙实反叙实名词的句法差异及其认知解释. 语言研究集刊, (01):1-14+372.

李新良,王明华. 2015. 汉语动词的叙实性研究的应用前景. 对外汉语研究, (02):120-129.

李新良,袁毓林. 2013. 面向计算的汉语动词蕴涵关系研究和型式库建设. 中国社会科学, (12):120-135+207.

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.

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, (01):42-52+127.

袁毓林. 2014. 隐性否定动词的叙实性和极项允准功能. 语言科学, 13(06):575-586.

袁毓林,寇鑫. 2018. 现代汉语名词的叙实性研究. 语言研究, 38(02):1-13.

袁毓林. 2020. "忘记"类动词的叙实性漂移及其概念结构基础. 中国语文, (05):515-526+638.

第二十四届中国计算语言学大会论文集，第96页-第104页，济南，中国，2025年8月11日至14日。

附录A. 感觉的提示语模板

”感觉”对应的提示语模板分为五种情况：

谓语动词中透露了说话人在命题的事实性或主观态度上的倾向性，谓语的语义结构中有明确指引其宾语所指命题的真假的信息，即动词本身具有一种词汇预设。说话人通过主句谓语动词来规定宾语小句的真值情况，谓语动词被分为五类，分别是”叙实动词”、”弱叙实动词”、”反叙实动词”、”非叙实动词”、”半叙实动词”。

”感觉”是一个非叙实词，”感觉”对应的命题”hypothesis”的判断分为五种情况。首先判定”text”是否符合以下三种情况，如果是，则命题”hypothesis”直接归为不能判断，结果为”U”。

情况一：语境中只是某个人单纯地表达意见，那么”感觉”之后的内容归为无法判断。如Text里只单纯写了某某人感觉+命题，此时”hypothesis”里的命题归为无法判断。

情况二：句子里明确表达了说话人在运用”夸张”的手法，表达的内容是夸张的，与事实不想符合，则”感觉”后面的内容为无法判断。

情况三：注意当”感觉”之后的小句的主语是人时，采用这条规律。text里的内容”感觉”之后小句的主语是人的情况下，且和”感觉”的主语是不一致，则无论”hypothesis”栏里的命题是什么样子的，答案都是”U”。举例来看：

”did”：
”predicate”：感觉
”type”：非叙实谓词
”text”：我感觉她很快向我跑了过来。
”hypothesis”：确实她很快向我跑了过来
”answer”：U
”did”：
”predicate”：感觉
”type”：非叙实谓词
”text”：我感觉她很快向我跑了过来。
”hypothesis”：确实她没有很快向我跑了过来
”answer”：U
”did”：
”predicate”：感觉
”type”：非叙实谓词
”text”：小明感觉班长讨厌他。
”hypothesis”：确实班长讨厌他
”answer”：U
”did”：
”predicate”：感觉
”type”：非叙实谓词
”text”：小明感觉班长讨厌他。
”hypothesis”：确实班长没有讨厌他
”answer”：U

如果不符合以上三种情况，则说明”hypothesis”的命题可以判断真假，根据以下内容判定输出为”F”还是”T”。

情况四：上下文有相关的证据来支持”感觉”的内容，可以对”感觉”的内容提供证据，则说明”感觉”之后的内容是正确的命题。在这种情况下，如果”hypothesis”里的命题所表达的意思是”对感觉后的内容做出了一个肯定”或者”确实+感觉之后的命题”或者”确实+感觉后命题的部分内容”，则说明其命题为真，输出为”T”；如果”hypothesis”里的命题所表达的意思跟”text”里的命题相矛盾，则输出”F”。举例来看：

”did”：
”predicate”：感觉
”type”：非叙实谓词

”text”：9月T7日正式比赛，经过认真准备，我感觉自己思路清晰，心情平静，举枪瞄准也很稳。

”hypothesis”：自己确实思路清晰，心情平静，举枪瞄准也很稳。

”answer”：T

”did”：

”predicate”：感觉

”type”：非叙实谓词

”text”：9月T7日正式比赛，经过认真准备，我感觉自己思路清晰，心情平静，举枪瞄准也很稳。

”hypothesis”：自己确实思路不清晰，心情不平静，举枪瞄准也不稳。

”answer”：F

情况五：我所”感觉”的命题是与明确的客观事实相关联，此时可以对”hypothesis”里的命题真假进行判断。在这种情况下，如果”hypothesis”里的命题所表达的意思是”对感觉后的内容做出了一个肯定”或者”确实+感觉之后的命题”或者”确实+感觉后命题的部分内容”或者”去掉’确实’后的内容和命题基本一致”，则说明其命题为真，输出为”T”；如果”hypothesis”里的命题所表达的意思跟”text”里的命题相矛盾，则输出”F”。