CCL25-Eval任务四系统报告：宏观模式提示与高效微调在叙实性推理中的应用

李泽群[1], 钟元浩[1], 柴成亮[1]
[1]北京理工大学
{zqli, zyh04, ccl}@bit.edu.cn

摘要

本文研究了利用大语言模型进行谓词引导的叙实性推理任务。在不微调场景下，针对Gemini 2.5 Pro模型，我们构建了基于谓词类型的思维链（CoT）提示，并创新性地让模型学习整个带答案的样本集以归纳宏观模式和规则，最终形成高效的提示词模板。在微调场景下，我们选用Qwen3-32b模型，利用llama factory进行LoRA微调，并使用llama.cpp完成模型向gguf格式的转换、量化及Ollama部署。实验结果展示了所提方法的有效性，其中在不微调赛道上，基于宏观模式提示的方法取得了94.01%的准确率；在微调赛道上，基于微调模型的系统取得了92.61%的准确率。

关键词： 叙实性推理；大语言模型；提示工程；思维链；LoRA微调

System Report for CCL25-Eval Task 4: Application of Macroscopic Pattern Prompting and Efficient Finetuning in Factivity Inference

Zequn Li[1], Yuanhao Zhong[1], Chengliang Chai[1]
[1]Beijing Institute of Technology
{zqli, zyh04, ccl}@bit.edu.cn

Abstract

This paper investigates predicate-guided factivity inference using large language models (LLMs). For the non-finetuning scenario with the Gemini 2.5 Pro model, we constructed Chain-of-Thought (CoT) prompts based on predicate types. Innovatively, the model was made to learn from the entire labeled dataset to induce macroscopic patterns and rules, ultimately forming an effective prompt template. For the finetuning scenario, we selected the Qwen3-32b model, performed LoRA finetuning using llama factory, and utilized llama.cpp for model conversion to gguf format, quantization, and deployment via Ollama. Experimental results demonstrate the effectiveness of the proposed methods, achieving an accuracy of 94.01% in the non-finetuning scenario and 92.61% with the finetuned model on the complete dataset.

Keywords: Factivity Inference , Large Language Models , Prompt Engineering , Chain-of-Thought (CoT) , LoRA Finetuning

1 引言

在自然语言理解的诸多挑战中，精确判断文本陈述的真实性，即“叙实性推理”（Factivity Inference），是一项至关重要且极具挑战性的任务。这项任务的核心在于辨别语言所描述的

第二十四届中国计算语言学大会论文集，第118页-第127页，济南，中国，2025年8月11日至14日。

事件或状态是否与现实相符。这一概念深深植根于语言学和哲学的交叉领域，其现代语言学研究可追溯至Kiparsky 等学者的开创性工作 (Kiparsky et al., 1968)。他们系统地分析了谓词（predicates）如何预设（presuppose）其补足语从句的真实性，从而将谓词分为“叙实性谓词”（factive predicates）和“非叙实性谓词”（non-factive predicates）。例如，“知道”、“意识到”、“后悔”等叙实性谓词，无论其主句是肯定还是否定形式，都预设了其宾语从句所描述的事件为真。从“小王不知道会议取消了”中，我们可以可靠地推断出“会议确实取消了”这一事实。进行叙实性推理所依赖的，正是在很大程度上独立于具体世界知识的、关于语言内部语义关系的“分析性语言知识” (袁毓林, 2020)。

随着大语言模型（LLMs）的飞速发展，其强大的文本理解与逻辑推理能力为解决这一复杂任务开辟了新的道路。然而，当前大语言模型在叙实性推理任务上的表现暴露出一个根本性的矛盾：尽管LLMs在海量文本数据中学习了丰富的统计模式，但它们本质上是概率性的文本生成器，而非严谨的逻辑推理器。这导致它们在处理需要精确语言学知识的任务时，常常表现不佳，甚至产生“幻觉”。模型可能混淆从数据中习得的“世界知识”（即事件发生的普遍概率）与句子结构所蕴含的“分析性知识”（即谓词带来的语义预设） (Bender et al., 2021)。换言之，目前的大语言模型推理链路在很大程度上仍然是一个“黑箱” (Ribeiro et al., 2016)，缺少坚实的语言学理论作为其决策的显式支撑和解释依据。

弥合这一差距至关重要。真实性推理作为一种重要的语言导航机制和手段，是机器进行文本蕴涵识别 (MacCartney et al., 2009)、幻觉处理 (Ji et al., 2023)、信念修正 (Zhang et al., 2022)等任务的重要的语义基础和形式依据。同时，它对信息检索、信息抽取、问题回答、情感分析等下游任务都具有重要的价值。随着大语言模型日益具备类人的与外界自主交互的能力，甚至被称为“智能体”（agent），从话语中准确获取事实性信息及说话人对事件真实性判断的主观态度，对于智能体的自主推理和人机交互的顺畅性而言是极为关键的。

本文以CCL25-Eval评测任务四为背景，对谓词引导下的叙实性推理任务进行了深入探索。我们并未局限于单一路径，而是开创性地沿着两条并行的技术路线展开研究：其一是在非微调场景下，我们选用前沿的Gemini 2.5 Pro模型，摒弃了传统繁琐的规则制定，创新性地提出了一种“宏观模式提示”策略，通过让模型从完整的标注数据中学习并归纳出高层级的推理模式与规则，从而构建出高效的提示模板；其二是在微调场景下，我们着眼于开源模型生态，选用了性能卓越的Qwen3-32b模型，并结合低秩适应（LoRA）这一高效微调技术，旨在以更经济的计算资源实现模型的深度任务适配。本研究旨在通过这两条路线的实践与对比，全面揭示当前大语言模型在处理精确叙实性推理任务时的潜力与挑战，并深入分析提示工程策略与模型选择对最终性能的深远影响。

2 相关工作

近年来，利用大语言模型解决各类自然语言处理任务已成为主流趋势。在叙实性推理方面，研究人员发现，通过向模型提供明确的指令和少量示例（Few-shot Learning） (Wei et al., 2022)，可以有效提升其判断的准确性。Chain-of-Thought（CoT） (Kojima et al., 2022)提示策略通过引导模型进行逐步推理，显著增强了其在复杂逻辑问题上的表现。

在模型微调方面，LoRA作为一种参数高效的微调方法，能够在不重新训练整个模型的情况下，通过更新少量参数来适应新任务，从而大幅降低了训练成本。Qwen、Deepseek等开源大模型凭借其优异的中文处理能力，在各类评测中表现出色。Llama.cpp和Ollama等工具的出现，则极大地简化了模型的量化、部署和本地运行流程。

本研究在前人工作的基础上，针对叙实性判断任务的特点，在非微调和微调两个方向上进行了深入探索。非微调方法的核心在于如何设计出能最大限度激发大模型潜能的提示词，而微调方法的重点则是在保证性能的同时，尽可能地提高训练和部署的效率。

3 实验方法

3.1 不微调赛道

3.1.1 模型介绍

我们使用**Gemini 2.5 Pro Preview-0506** 作为评测使用的模型。针对两种不同来源的语料集，我们分别设计了相应的提示策略。

- **自然语料集**: 首先，我们构造一个基本的基于谓词类型的CoT思维链式提示词。随后，依赖Gemini 庞大的上下文窗口和强大的分析推理能力，我们将NatS.json 这个包含了答案的样本集作为整体输入至大模型，由大模型进行整体学习理解，进而推理归纳出预测的**宏观模式和规则**，该规则可以应用于所有叙实性推理的数据类型。随后，我们将该规则和模式合并先前的提示词，作为最终提示词模板输入给Gemini 模型，以得到测试集的最终答案。详细的提示词见附录。

- **人工语料集**: 我们利用人工语料集提供的**谓词叙实类型**构造提示词。通过观察样本集，我们注意到非叙实类型的谓词，背景句一定不能推出结论句，答案是U。而其他类型的谓词，由于明确了叙实类型，我们对每种叙实类型专门构造了一种提示词。接着，再采用在自然语料集上的方法，将ArtS.json 这个包含了答案的样本集作为整体输入至大模型，由大模型进行整体学习理解，进而推理归纳出预测的宏观模式和规则。我们将这两部分结合优化，得到最终的提示词。

3.1.2 评测指标

本次评测采用正确率作为评价指标:

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

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 为自然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。

3.2 微调赛道

3.2.1 模型选择

在非微调赛道中，我们队伍使用了**Gemini 2.5 Pro Preview-0506** 模型进行评测。但是我们注意到使用该模型进行微调需要花费高额的api调用成本，所以最终我们采用了最近推出的开源模型**Qwen3-32b**进行微调的工作。

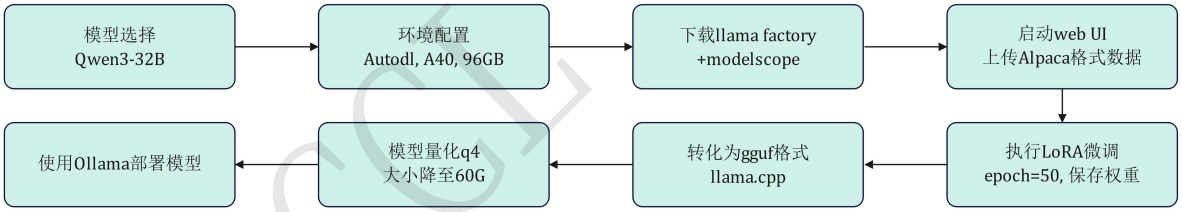

Figure 1: Qwen 微调与部署流程图

3.2.2 LoRA微调

LoRA（Low-Rank Adaptation）是一种高效的微调方法。其核心思想是在预训练模型的权重矩阵旁边，并行地引入两个低秩矩阵（A和B），训练时只更新这两个低秩矩阵的参数，而预训练模型的原始权重保持冻结。数学上，对于原始权重矩阵$W_0 \in \mathbb{R}^{d\times k}$，LoRA引入低秩分解$W_0 + \Delta W = W_0 + BA$, 其中$B \in \mathbb{R}^{d\times r}, A \in \mathbb{R}^{r\times k}$，其中$B \in \mathbb{R}^{d\times r}$，$A \in \mathbb{R}^{r\times k}$，且秩$r \ll min(d, k)$。采用LoRA进行微调的优势包括：显著减少训练所需的参数量、存储高效、不易发生灾难性遗忘等。

3.3 配置介绍

我们注意到在本地微调大模型往往由于显存不够或者配置的原因，无法正常训练，且训练速度相当慢，只能够尝试一些参数很小的模型（2b）以内。我们最终选择使用算力平台Autodl租借云算力资源来完成从微调到使用Ollama完成云部署测评的全过程。我们使用的技术主要包含以下三个部分。

3.3.1 LoRA微调Qwen大模型

为实现大模型的高效微调，我们在具备**A40 GPU**（96GB 显存）、Python 3.10、CUDA 12.1 与PyTorch 2.1.0 的环境中搭建了微调平台，并通过AutoDL 云算力完成部署。微调流程基于开源框架`llama-factory`，该框架支持多种主流大语言模型及包括LoRA 在内的多种训练策略，具备高度灵活性与可扩展性。为获得Qwen3 模型源码，我们额外集成了`modelscope` 平台，并通过前者的Web UI 进行微调流程管理。

为便于远程访问，我们通过AutoDL 提供的SSH 隧道工具实现了本地浏览器对远程JupyterLab 微调界面的访问。在数据准备阶段，我们依据`llama-factory` 的格式规范，采用Alpaca 格式组织训练样本，数据字段包括`instruction`（指令）、`input`（可选输入）与`output`（参考答案），均源自非微调赛道中的提示词模板及样例集合。

模型微调过程中，我们设定训练周期为50 epoch，以兼顾欠拟合与过拟合风险。训练结束后，系统自动保存LoRA 权重文件、损失曲线与训练日志，并提供微调结果的验证接口。在确保生成质量的前提下，我们将模型导出为`safetensors` 格式，为后续格式转换与部署提供基础支持。由于Ollama 推理引擎需加载`gguf` 格式模型，我们进一步执行格式转换以完成最终部署。

3.3.2 使用llama.cpp转换gguf格式并量化

为了支持部署和推理，我们在完成LoRA微调后，采用了开源工具llama.cpp 对模型进行格式转换与量化处理。该工具支持将HuggingFace 格式的safetensors 权重转化为可部署的gguf 格式，同时提供多种精度的权重量化机制。我们首先在具备CUDA 支持的环境下对llama.cpp 进行编译，随后使用其官方提供的convert-hf-to-gguf.py 脚本完成模型格式的转换。由于微调后的模型体积较大，约为240GB，为降低存储与推理成本，我们进一步使用llama-quantize 工具对模型进行4-bit 量化（q4），在保持性能的前提下将模型体积压缩至约60GB。最终，转换与量化后的gguf 模型被用于Ollama 部署，为后续本地或云端推理提供了可行的解决方案。

3.3.3 Ollama部署Qwen微调大模型

最后我们在云端通过编译源码的方式下载了Ollama，并编写了模型路径、模型模版以及停止词，至此可以直接在Ollama调用微调后的模型。

4 结果分析

4.1 非微调赛道

4.1.1 评测结果

模型	人工语料集	自然语料集	完整数据集
Gemini 2.5 Pro Preview-0506	97.80	92.58	94.01

Table 1: 不微调赛道准确率（%）

4.1.2 前期尝试

前期为了节省时间和计算资源，我们使用本地部署的Qwen2:7b 模型进行实验，temperature设定为默认值1。轻量化的模型必然带来能力较弱的推理效果和不稳定的答案输出，但根据提示词的不同，准确率的显著变化能直接反映出提示词的优劣。我们目标是通过Qwen2:7b 这个轻量化的模型筛选出表现较好的提示词，随后将该提示词放于不同模型进行比较，得到最佳实验结果。下面将详细介绍我们的尝试过程。所有的准确率均基于样例集计算。

Baseline：采用非常直接的指令，要求模型仔细阅读背景句和结论句，并严格依据背景句给出T/F/U 的判断。该提示词直接描述任务指令，没有明确指示思考方式或步骤，完全依赖于预训练模型本身的能力。因此把该准确率作为后续改进的起点。

Zero-Shot-CoT (零样本思维链):在Baseline 的基础上，增加了“请一步步进行推理并得出结论”的指令，引导模型进行显式的推理过程。但该提示词在自然和人工数据集上表现差异明显。ArtS 上更显著的提升(+2%) 表明，对于结构化更强、逻辑可能相对简单（但仍需推理）

的人造数据，CoT 提示词能帮助模型更稳定地遵循其内在逻辑。NatS 上较小的提升(+0.29%)则暗示，虽然CoT 有帮助，但自然语言的复杂性、歧义性以及对更广泛世界知识的需求，可能意味着仅仅通过简单的逐步指令，对于一个7B 规模的模型来说，其影响是有限的。任务本身可能更难，或者需要比简单顺序步骤更细致的推理。

Few-shots+Zero-Shot-Cot (少样本+零样本思维链):进一步在Zero-Shot-CoT 的基础上引入了示例。针对自然语料集和人造语料集分别提供了不同的示例，展示了详细的“推理过程”和期望的输出格式。纯粹的指令可能存在多种解释。示例通过实际案例缩小了这种解释空间，向模型清晰地展示了什么样的推理是被认可的。示例起到了锚点的作用，指导模型在处理新的、未见过的数据时的生成过程。模型会尝试模仿示例中展示的模式。实验显示，模型效果有轻微提升(NatS +0.85% ArtS +0.66%),但整体表现离预期还是有一定距离。

针对谓词类型优化：观察人工语料数据集，我们注意到样本明确指示了谓词的叙实类型。非叙实谓词指出这类谓词后的内容真实性无法确定，因此结论句通常也无法判断，引导模型输出U。而正叙实谓词指导模型将谓词后的内容视为事实，反叙实提取“假的内容”，进行逻辑取反得到“隐含的真实情况”。因此，我们根据背景句中谓词的叙实性设计了三种高度特化的提示词模板，指定明确的叙实类型将一个复杂的叙实性推理任务分解成了几个更简单、规则更明确的子任务，很大程度减少了大模型的幻觉程度。实验效果显著证明进行精细化的模板对预测人工语料数据集正确率提高显著（跃升至92.33%）。

面对自然语料集预测准确率提升缓慢的问题，我们尝试将预测错误的样本进行单独分析，并依此制定预测的附加规则。这些规则针对这几类相似样本的特定特征和用词，可能非常精细，能够完美修正这些特定错误。但当这些高度特异性的规则被应用到全局数据时，它们可能无法适应其他样本的多样性和复杂性。对于那些不完全符合这条规则适用条件的样本，这条规则可能会错误地触发，或者给出不恰当的引导，从而导致新的错误。随着错误样本的积累，制定的规则会越来越多，规则库会变得异常庞大和复杂，难以管理、调试和维护。新规则的加入可能会破坏旧规则的有效性。针对错误样本制定规则，更像是在模型的现有能力基础上“打补丁”，而不是从根本上**提升模型对任务的理解和泛化能力**。外部强加的、过于具体的规则，如果与模型内部的“理解”不一致，可能会导致模型行为混乱，效果反而下降。

基于此，我们提出了本任务采用的提示词，即**先将数据集整体输入给大模型进行总结归纳预测模式和步骤，随后将该模式作为提示词的附加规则**。由于该规则是宏观的，模式化的，而非特异性的，局部的，所以我们设计的提示词具有很好的预测效果。

模型	方法	准确率(%)		
		人工语料集	自然语料集	总准确率
Qwen2:7b	Baseline	57.67	70.00	66.30
	Zero-Shot-CoT	59.67	70.29	67.10
	Few-shots+Zero-Shot-Cot	60.33	71.14	67.90
	针对谓词类型优化	92.33	71.14	77.50
	宏观模式提示	**92.39**	**86.89**	**88.54**

Table 2: 不同方法在各个语料集上的准确率对比

4.1.3 模型选择

我们将表现最好的提示词，即宏观模式提示分别在Qwen3：32b，Deepseek-r1：32b进行试验，得益于参数量的增加和训练技术的增强，准确率较Qwen2：7b有明显提升。5月6日发布的**Gemini 2.5 Pro**展示了强大的模型推理能力。Gemini 2.5 Pro 的突出之处在于它能够深入分解问题，而不是简单地重复训练数据。谷歌将其描述为一种思维模型，在提供最终答案之前进行逐步严谨推理。在实际基准测试中，Gemini 2.5 Pro 在编程、数学和科学等领域的表现优于GPT-4、Anthropic 的Claude 和其他领先模型——在GPQA(Rein et al., 2024) 等评估中名列前茅。Gemini 2.5 pro拥有惊人的内存，最多支持100万token的上下文窗口，这意味着它可以处理整本书籍、整个代码库或大型数据集，而不会丢失对话的线索。这些能力与我们当前叙实性推理的需求极其适配，当我们将提示词在该模型上测试时，结果远超其他大语言模型。Gemini

2.5 pro的使用成为我们叙实性推理任务取得巨大成功的重要基石。推理过程中temperature参数统一设置为默认值1。

模型	人工语料集	自然语料集	完整数据集
Qwen2:7b	92.39	86.89	88.54
Deepseek-r1	93.17	88.24	89.72
Qwen3:32b	93.97	90.78	91.73
Gemini 2.5 Pro Preview-0506	**97.80**	**92.58**	**94.01**

Table 3: 不同模型使用最佳提示词的准确率（%）对比

4.1.4 讨论

尽管Qwen2-7B 等轻量级模型的绝对性能有限，但其对提示词变化的敏感反应使其成为快速筛选和验证提示词有效性的理想平台，显著降低了实验成本和时间。轻量级模型上的性能差异能够为提示词设计的方向提供早期且有价值的信号，帮助我们在投入更强模型前优化策略。Zero-Shot-CoT和Few-Shot Learning等通用引导策略是提升模型表现的基石，它们教会模型基本的思考框架和任务范式。针对特定数据特征（如人造语料中的谓词类型）设计高度特化的规则，能够在特定场景下实现性能的显著跃升，这揭示了将领域知识或数据结构信息融入提示词的巨大潜力。认识到特异性规则的局限性后，转向让模型自身从数据集整体中学习并总结出宏观的、模式化的预测规则和步骤，是实现提示词泛化能力和鲁棒性的关键一步。这种“授人以渔”而非简单“授人以鱼”的策略，更能适应复杂多变的真实数据。

同时，实验清晰地表明，模型自身的基础能力（如Gemini 2.5 Pro所展现的强大推理和长上下文处理能力）是决定任务性能上限的核心要素。精心设计的提示词能够有效地引导模型，激活其特定能力，使其更聚焦于任务的核心需求，从而在特定任务上逼近甚至达到其性能上限。优秀的提示词与强大的模型相结合，能产生“1+1 >2”的效果。

基于本次研究在叙实性推理任务中积累的经验，我们认为未来提示词工程的发展将聚焦于提升其智能化、自适应性和构建效率，从而更深层次地释放大规模语言模型在复杂推理任务上的潜能。通过深入探究大模型从海量数据中学习并凝练出宏观预测模式和推理步骤的内部工作机制。理解这一“顿悟”过程，将为设计更高级的“元提示（meta-prompting）”——即引导模型自主进行高效归纳的提示——奠定理论基础。同时，考虑探索端到端的自动化方法，使模型能够根据任务目标和反馈自主生成、评估并迭代优化宏观规则型提示词。这可以借鉴强化学习（以任务性能为奖励）、进化算法（模拟提示词的“优胜劣汰”）乃至利用大模型自身进行指令生成与优化（如APE框架）。总之，未来的提示词工程将不再仅仅是“手艺活”，而是会演变成一门更系统、更智能、更自动化的科学与工程学科。通过上述方向的深入探索，我们期待构建出能与大语言模型深度协同、共同进化的提示词系统，从而将AI在复杂认知任务上的能力推向新的高度。

4.2 微调赛道

4.2.1 评测结果

同样地，该部分推理过程中temperature参数统一设置为默认值1。我们的评测结果如下：

模型	人工语料集	自然语料集	完整数据集
Llama:7b	84.09	77.23	79.29
Phi4:14b	85.66	80.81	82.27
Mistral:7b	85.71	83.01	83.82
Qwen2:7b	94.09	86.89	89.05
Deepseek-r1	94.22	91.05	92.00
Qwen3:32b	**94.86**	**91.75**	**92.61**

Table 4: 不同微调模型使用最佳提示词的准确率（%）对比

4.2.2 结果分析与讨论

在微调实验的初期阶段，我们自小参数量模型起步，系统性对比了多种不同开源模型在本任务数据集上的表现。实验结果显示，在参数规模相近的条件下，Qwen 与Deepseek 系列模型在中文语义理解方面的表现普遍优于LLaMA、Phi、Mistral 等国外开源模型。其中，Qwen3 系列在本任务设定下展现出更高的响应准确率与更低的平均生成延迟，相对Deepseek-R1 系列具有一定优势。

5 结论

本研究围绕谓词引导的叙实性推理任务，成功地从非微调与高效微调两个维度验证了大语言模型的应用潜力。我们的探索证明，无论是依赖尖端闭源模型的强大推理能力，还是通过对优秀开源模型进行高效适配，都能有效解决这一复杂的自然语言理解问题。

在非微调的路径上，我们提出的“宏观模式提示”策略被证明是一种极为有效的方法，它超越了传统依赖人工制定具体规则的局限性，通过“授人以渔”的方式，充分激活了Gemini 2.5 Pro这类先进模型从海量数据中自主归纳宏观规律的“顿悟”能力，最终实现了卓越的判断准确率。这一成功凸显了模型自身的基础能力与精心设计的提示工程相结合所能产生的“1+1>2”的协同效应。

而在微调的路径上，我们基于Qwen3-32b模型与LoRA技术的实践，不仅验证了开源模型在中文语义理解上的强大实力，也完整地展示了一套从训练、量化到部署的高效、可行的技术流程，为在特定任务上构建高性能、低成本的专用模型提供了宝贵的实践范例。总而言之，本次研究积累的经验揭示了未来提示工程的发展方向，即从当下的“手艺活”演进为一门更加智能化、自适应的科学，通过探索让模型自主生成和迭代优化宏观推理策略的元提示（meta-prompting）方法，从而更深层次地释放大语言模型在复杂认知任务上的潜能，推动人工智能向着更深度的理解与推理迈进。

致谢

感谢北京理工大学计算机学院辛欣老师在《知识工程》课程中对本工作的指导，包括课程中关于知识表示与逻辑推理的研究。感谢CCL评测组委会和任务组织者澳门大学袁毓林老师、南京师范大学李斌老师以及任务联系人丛冠良的支持。该论文得到了北京市自然科学基金（QY25265、QY25259）的项目经费资助。

参考文献

Paul Kiparsky and Carol Kiparsky. 1968. *Fact.* Linguistics Club, Indiana University.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.*

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. *Proceedings of the Eighth International Conference on Computational Semantics.*

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Jungo Kasai, Junjie Hu, W.L. Hamilton, Yejin Choi, and Noah A. Smith. 2022. GreaseLM: Graph reasoning enhanced language models for question answering. *arXiv preprint arXiv:2201.08860.*

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

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

David Rein, Stas Gendler, T.J. Lu, Simran Arora, Daniel M. Ziegler, Vatsal Marwah, Nataniel Ruiz, Alex Kendall, Zong- 2024. Gpqa: A graduate-level google-proof q&a benchmark. *First Conference on Language Modeling.*

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, (01):1–9.

附录A. 提示词模板

叙实性推理任务：完整提示词模板

最高指令：

你将获得三项信息：**背景句(P)**、**结论句(H)**、以及作为线索的**谓词(V)**。你的核心任务是：**严格依据背景句(P)中由谓词(V)所引导或关联的目标陈述(S)的真实性，来判断结论句(H)的真实性。**

请进行详细的逻辑推理，并最终输出T（真）、F（假）或U（无法确定）。

—

一、核心推理步骤与谓词分类指导(重点强化版)

1. **定位目标陈述(S):** 首先，在背景句(P)中，找到由谓词(V)直接引出或紧密关联的核心陈述或事件，我们称之为目标陈述**S**。

2. **判断S在P语境下的真实性(S_{P_truth}):** 这是最关键的一步。根据谓词(V)的类型和P句的整体上下文，判断S是真是假，还是不确定。

* **A. 强事实性谓词(Strongly Factive Predicates):** * **示例:** ‘看见‘、‘发现‘、‘知道‘、‘意识到‘ (当S为可直接感知或已发生的事实时)、‘清楚‘、‘听出‘、‘证实‘、‘目睹‘、‘揭露‘、‘获悉‘、‘披露‘、‘记得（S实际发生）‘、‘（正确）猜到‘、‘领悟到‘、‘注意到‘、‘承认（事实/过错）‘、‘表明‘ (当P基于客观数据或明确逻辑推导时)、‘暴露出‘ (当P基于事实揭示问题时)、‘反映出‘ (当P基于客观现象推断时)。* **规则:** 这类谓词强烈预设其关联的**S** 在背景句(P)的语境下为**真(TRUE)**。* **”没有承认S”**: 这不等同于S为假。它只是否认了“承认”这个行为。S本身的真实性需要根据P中其他信息判断，若无其他信息，则S为U。*（此条针对“没有承认”特别补充）*

* **B. 强反事实性谓词(Strongly Counter-Factive Predicates):** * **示例:** ‘谎称‘、‘幻想‘、‘污蔑‘、‘诬陷‘、‘妄称‘. * **规则:** 这类谓词强烈暗示其关联的**S** 在背景句(P)的语境下为**假(FALSE)**。* **”假装S”**: S本身是假的。说话者做出的行为是“假装S”，但S描述的状态或行为并未真实发生。* *训练集NatS_259:* ”佛罗多依旧假装<u >听不懂神行客的暗示</u >。” S=”听不懂暗示” (FALSE)。H=”佛罗多确实听懂了” (TRUE)。* *训练集NatS_273:* ”...假装<u >是瞎子</u >...” S=”是瞎子” (FALSE)。H=”我确实不是瞎子” (TRUE)。* **注意**: 如果S描述的是一个动作，如“假装把鞋带解开”(NatS_265)，P没有明确说他是否真的解开了，则“把鞋带解开”这个动作的完成与否是U。我们只知道他在“假装”进行这个动作。

* **C. 主观/非事实性/弱指示性谓词(Subjective/Non-factive/Weakly Indicative Predicates):** * **规则总览:** S的真实性**默认为U**，除非P通过**上下文、后续描述、叙述者立场或特定句式**提供了明确的佐证或反驳。

* **详细分类与案例指导:**

* **1. ‘哀叹S‘ / ‘感叹S‘:** * 若S是说话者对一个**已被P中其他信息描述或暗示为事实的状态/事件的反应**，则S_{P_truth}为T。* *训练集NatS_012:* ”...让塞莱斯频频受骗, 哀叹<u >自己腿脚不灵活</u >。” (P中“频频受骗”佐证了S) =>S为T。* *训练集NatS_013:* ”...哀叹<u >没有网络的好日子一去不复返了</u >...” (P将其作为领导们的普遍共识和既成事实来陈述) =>S为T。

* 若S是**纯粹的观点、评价、或对未证实情况的感伤**，且P无其他佐证，则S<sub>P_truth </sub >为U。* *训练集NatS_001, NatS_003, NatS_005, NatS_007, NatS_019.*

* **2. ‘抱怨S‘ / ‘埋怨S‘ / ‘批评S‘ / ‘指责S‘:** * 若P仅转述抱怨/批评内容，S<sub>P_truth </sub >为U。* *训练集NatS_383:* ”埋怨<u >太贵了</u>” (主观感受) =>S为U。* *训练集NatS_431:* ”批评<u>军演破坏了气氛</u>” (单方面指责，P未证实) =>S为U。* 若P通过**”其实”、”但”、”然而”等引出对S的反驳或不同事实**，则S_{P_truth}(作为绝对事实)为F。

* 若P**后续的描述或行动强烈佐证了S的内容**，则S_{P_truth}为T。* *训练集(类似逻辑):* ‘NatS_387‘ (”大家都埋怨我们拖欠工资” ->包工头自己也承认了这个情况) =>S为T。* *训练集NatS_441:* 埃尔巴坎...批评<u >青年学生在上课前不背诵《古兰经》</u>。H:青年学生在上课前确实会背诵《古兰经》。答案是T，这意味着S(“青年学生

不背诵")是F，因为H是 S且为T。这个例子特殊，可能暗示了"批评X不Y"时，X实际上是Y。或者，H的判断超出了P的信息，应为U。**（鉴于此例的特殊性，除非H直接针对"批评"这一行为，否则对S的判断应更依赖P的明确信息。）** * **修正/细化:** 对于"批评/抱怨/埋怨X 不Y"，如果H是"X 确实Y"，则H与S(Y)矛盾。如果P只说有人批评X不Y，而H问X是否Y，S(Y)为U，则H(Y)也为U。'NatS_441'的答案T可能基于外部知识或对"批评"在特定语境下的强反事实解读。**在缺乏强P内部证据时，S的内容应判U。**

* **3. '感觉/觉察出/觉出/觉得/觉着S':** * 若S是主体**直接的、可感知的物理状态、环境状况或已发生的简单行为**，且P无反驳，S_{P_truth}为T。* *训练集NatS_133:* "感觉<u>身体不适</u>" (直接感受) =>S为T。* *训练集NatS_313:* "觉出<u>有一点儿不妥</u>" (直接感受/判断) =>S为T。* 若S是关于**他人意图、复杂评估、未来预测或纯粹主观的比喻/联想**，S_{P_truth}为U。* *训练集NatS_141:* "感觉<u>他的脑袋已经被摔碎了</u>" (夸张的主观感受) =>S为U。* *训练集NatS_143:* "感觉<u>现在国内球员...不是太好</u>" (主观评价) =>S为U。

* **4. '相信S' / '认为S' / '声称S' / '估计S' / '猜测S' / '怀疑S' / '听说S' / '声言S' (以及其他纯观点/主张/信念类):** * S_{P_truth}**一律为U**，除非P中出现**明确的、独立的上下文佐证或反驳S**，或者P的叙述者通过特定修辞（如"不敢相信S竟是真的"）来断言S。* *训练集(U的例子):* 'NatS_045' (猜S), 'NatS_203' (怀疑S), 'NatS_321' (觉着S是观点), 'NatS_483' (声言S)。* *训练集(T/F的例外):* 'NatS_201' (怀疑S，但P的整体叙述暗示S为F), 'NatS_205' ("谁也不再怀疑S" =>S为T), 'NatS_495' (声言对判决不满，这是对其立场的陈述，为T), 'NatS_499' (声言S，但P叙述者评价为"恬不知耻" = >S为F)。

* **5. "以为S":** * **"以为S，但/其实/后来才发现/没想到B/ S"**: S_{P_truth}为F, B/ S为T。* *训练集NatS_623, NatS_625, NatS_629, NatS_635, NatS_637, NatS_641.* * **P仅陈述"X 以为S" (无后续修正)**: S_{P_truth}为U。

3. **对比S_{P_truth}与结论句(H):** * 如果S_{P_truth}为**TRUE**: 若H 肯定S ->**T**; 若H 否定S ->**F**。* 如果S_{P_truth}为**FALSE**: 若H 肯定S ->**F**; 若H 否定S ->**T**。* 如果S_{P_truth}为**UNCERTAIN**: H 对S 的任何判断->**U**。

—

二、重要注意事项与启发式规则

1. **严格依赖背景句(P)：** 唯一依据是P提供的信息。

2. **关注否定词与转折词。**

3. **"U"的倾向性：** 对于主观/非事实性谓词引导的S，除非P有**明确的、P内部的佐证信息**，否则倾向于U。

4. **聚焦核心事件S。**

5. **区分"事件/状态本身"与"对事件/状态的描述/态度"。**

6. **"记得/不记得S（事件）"**: "记得S发生过" =>S为T。"不记得S发生过" =>S为F。

7. **"想到/没想到S（事件）"**: "没想到S发生了" =>S为T。

8. **P叙述者使用"（不）敢相信S"或"或许你（不）会相信S"修辞时**，通常S为T。

—

三、输入:

背景句：{text}

结论句：{hypothesis}

谓词：{predicate}

— **请一步一步进行推理**

四、输出格式

* **仅输出T, F, 或U 中的一个字母。**