System Report for CCL25-Eval Task 4: A Two-Stage Inference Framework Based on Predicate Semantics and Context Reasoning

Iskar Deng, Nathalia Xu
Fish NLP
University of Washington
{hd49, mx727}@uw.edu

1 Overview

We propose a two-stage factivity inference framework that leverages predicate semantics as the core signal for inference. This is made possible by the datasets' inclusion of explicit predicates and, in the artificial test set, their semantic types.

Our framework includes two stages:

- **Stage 1:** Initial inference classification based on the predicate's type.
- **Stage 2:** Adjustment based on contextual clues, if necessary.

All predictions are generated using the **Qwen-Max** language model (Team, 2024).

2 Methods

2.1 Stage 1: Predicate-Based Initial Classification

Predicate Types and Default Labels We categorize predicates into three semantic types based on their inherent factual bias, and assign default factuality labels accordingly:

- 事实型 **Factives → T**
  Examples: “看见”, “意识到”, “记得”
  These predicates imply that the embedded proposition is **true (T)**, as they represent confirmed observations, knowledge, or memories.

- 反事实型 **Anti-factives → F**
  Examples: “否认”, “抹黑”, “假装”
  These express disbelief, falsification, or fictional constructions, indicating that the embedded proposition is **false (F)**.

- 主观型 **Subjectives → U**
  Examples: “认为”, “怀疑”, “感觉”, “感叹”
  These reflect personal feelings, assumptions, or subjective views. They cannot determine the truth value of the embedded proposition, and are marked as **uncertain (U)**.

For the artificial test set, we use the provided predicate type directly. For the natural test set, we use a prompt to guide the model to infer the predicate type.

2.2 Stage 2: Contextual Correction and Consistency Voting

Predicate-based labeling alone may overlook important context. To improve factivity judgment:

- If Stage 1 result is U, we try to revise it to T or F if context allows.

- If Stage 1 result is T/F, we revise it to U when uncertainty is supported by context.

We adopt the **Auto-CoT** (Automatic Chain-of-Thought) approach (Zhang et al., 2022). Using an initial prompt with simple instructions (e.g. “pay attention to context and judge”), we identify successful conversions from stage 1 judgments. Through clustering and manual inspection, we identify correction patterns and extract the model's own reasoning to create **few-shot exemplars**.

2.2.1 U→TF Corrections

We identify four major U→TF correction patterns:

1. Context provides supportive evidence

2. Context explicitly contradicts subjective claims (e.g., “认为是...但是...”)

3. Subjective predicate with negation implies confirmation (e.g., “我不再怀疑...”, “他没有感觉...”)

4. Subjective hyperbole not supported by facts (e.g., “脑子被狗吃了”)

We include a full reasoning trace for Rule 1 in Appendix A. For the artificial test set, due to lack of broader context, only rule 3 (negation within

Accuracy	Stage 1	Stage 2	Voting
Artificial test set	92.84%	91.58%	N/A
Natural test set	–	85.88%	86.35%

Table 1: Accuracy on artificial and natural test sets across inference stages. Stage 1 results on the natural test set were not submitted individually due to low performance, but we observed 79.86% accuracy on the sample set.

subjectives) is used. For the natural test set, the prompt uses all four rules as natural examples contain nuanced context.

2.2.2 TF→U Corrections

We also apply revision in the reverse direction when the initial T/F judgment is overly confident or lacks contextual support. We identify the following correction patterns:

1. Hypothetical or fictional context （e.g., "梦里我看见…"，"如果我看见…"）
2. Predicate is negated （e.g., "没看见"，"未发现"，"无法确认"）
3. Speaker self-doubt after assertion （e.g., "虽然觉察到了，但不能确定..."）

As with U→TF correction, for the artificial test set, only Rule 2 (predicate negation) is used due to limited context. For the natural test set, the prompt uses all three rules.

2.2.3 Consistency Voting (Natural Set Only)

Stage-2 corrections often involve uncertain or borderline decisions. To improve robustness, we apply consistency voting on **U→TF** transitions in the natural set:

- Each candidate is passed through the model **five times**
- The final result is determined by majority vote **(if tied, U>T>F)**

This method helps reduce variance and mitigates low-confidence overfitting from a single pass.

3 Results

As shown in Table 1, we applied the same two-stage method to both test sets for consistency. On the artificial set, Stage 1 alone performs strongly (92.84%), while Stage 2 shows a slight drop (91.58%)—likely due to the lack of broader context in the simpler sentences in the artificial set.

In contrast, the natural set benefits substantially from contextual reasoning. Although Stage 1 results were not submitted (as each submission counts and Stage 1 alone underperforms), sample set evaluation of stage 1 shows 79.86% accuracy. Stage 2 raises this to 85.88%, and applying consistency voting further improves it to 86.35%, likely due to the richer contextual cues in natural sentences.

References

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *Preprint*, arXiv:2210.03493.

A Full Example: U→T Reasoning (Rule 1)

text: 1997年10月，他感觉身体不适，经检查确诊为胃癌，由于发现得早，在河南医科大学附属医院治疗后癌块消失，病情大为好转。
predicate: 感觉
hypothesis: 他身体确实不适。
reasoning:
判断过程：
1. 命题分析：命题为"他身体确实不适"，需要判断是否可以从文本中推出。谓词是"感觉"，表明这是一个主观感受的描述。
2. 文本信息提取： - 文本提到"他感觉身体不适"。 - 后续内容提到"经检查确诊为胃癌"，这表明他的身体不适并非仅仅是主观感受，而是有医学诊断支持的事实。 - "由于发现得早，在河南医科大学附属医院治疗后癌块消失，病情大为好转"进一步验证了身体不适的真实性，并且与胃癌相关联。
3. 间接证据分析： - 虽然"感觉身体不适"本身是一个主观描述，但后续的医学检查和确诊结果（胃癌）提供了客观支持。 - 胃癌作为一种严重疾病，通常会伴随身体不适的症状，这在逻辑上与"感觉身体不适"形成呼应。 - 治疗后的病情好转也间接证明了之前的身体不适是真实存在的。
4. 排除歧义或多义性： - 文本没有其他可能解释"感觉身体不适"的语境，也没有暗示这是虚假或夸大的描述。 - 上下文逻辑清晰，不存在歧义。
5. 结论更新： - 尽管"感觉身体不适"是主观描述，但结合后续的医学诊断和治疗结果，可以合理推断他的身体确实存在不适。 - 因此，命题可以从文本中推出。

结论：T

result: T