System Report for CCL25-Eval Task 4: From Plain to Hierarchical —Knowledge-Augmented Prompting for Chinese Factivity Inference

Minjun Park
Duksung Women's University
Seoul, South Korea
karmalet@duksung.ac.kr

Seulki Lee
Chungbuk National University
Cheongju, South Korea
leeseulki0118@gmail.com

Abstract

To improve the factivity inference capability of large language models (LLMs), we adopted a Retrieval-Augmented Generation (RAG) framework using a curated bibliography on Chinese factivity semantics. We compared a baseline without retrieval against two RAG-based strategies, showing that hierarchical prompting with `RAPTOR` yields the highest accuracy. Using recursive summarization from the bottom up, `RAPTOR` allows models to access document context at multiple abstraction levels, resulting in more accurate and stable inference. Our findings contribute to deeper Chinese semantic inference through linguistic knowledge-augmented prompting in factivity inference and textual entailment.

Keywords: Factivity Inference , LLM , RAG , RAPTOR , GPT , Gemini , DeepSeek

1 Introduction

This technical report presents the modeling approach developed for participation in CCL25-Eval4: The First Chinese Factivity Inference Evaluation (FIE2025). The submission was made under the “non-finetuned track (不微调赛道)” and included two RAG-based system variants, built on `GPT-4.1`, `DeepSeek-R1-14B`, and `Gemini-2.5-pro-preview-05-06` tailored to the FIE2025 task.

Factivity Inference (FI) is a semantic task concerned with identifying whether a statement is considered true with given linguistic context. In the Chinese language, predicates such as “知道 (*know*), 意识到 (*realize*), 记得 (*remember*)” often imply a factual proposition regardless of the presence of negation. Mastery of FI is crucial for tasks such as hallucination detection, belief revision, and robust human-agent interaction, especially for Chinese-speaking agents.

The FIE2025 evaluation seeks to assess how well LLMs can infer factivity[1] in Chinese and how prompt design strategies influence their performance. Our team focused on enhancing LLMs' performance without fine-tuning by injecting domain knowledge through retrieval strategies.

Baseline large language models (LLMs) are not optimized for the specific demands of the Factivity Inference (FI) task. To enhance their performance, we applied Retrieval-Augmented Generation (RAG) methods, incorporating domain-specific knowledge on Chinese factivity semantics. Two approaches were developed: (1) Plain RAG and (2) `RAPTOR`-based RAG.

2 Methodology

2.1 Plain RAG

To enhance model performance with domain-specific knowledge, we constructed a curated bibliography based on the FIE2025 task description and five key academic papers on Chinese factivity theory (李新良 and 袁毓林, 2016; 李新良 and 袁毓林, 2017; 袁毓林, 2020a; 袁毓林,

[1] See (Ziembicki et al., 2024) for a lexical-semantic definition of factivity as distinct from event factuality.

2020b; 袁毓林, 2020c). These articles were selected based on their relevance to factivity inference, particularly those discussing epistemic verbs (e.g., “知道”, “忘记”, “记得”) and their associated syntactic-semantic behaviors. The sources were selected from high-impact journals in Chinese linguistics, including *Zhongguo Yuwen*(中国语文), *Contemporary Linguistics*(当代语言学), *Language Teaching and Linguistic Studies*(语言教学与研究), among others.

The texts were parsed into plain text using `LlamaParser`, which is optimized for simplified Chinese recognition. Each text was divided into chunks of `size = 500` and `overlap = 50`, resulting in a total of 240 documents and 7,495,073 tokens. Embeddings for all three models(`Gemini`, `GPT` and `DeepSeek`) were generated using `text-embedding-3-large`, a multilingual model suitable for East Asian languages. Prompt templates were structured as follows:

```
You are an expert reasoning assistant. For each question, explain your reasoning step-by-step
before giving the final answer.
你是一个自然语言推理模型，请根据以下信息，判断“假设（hypothesis）”是否成立。
   - 请严格根据提供的“选项（option）”返回答案，仅填写"answer"字段，不要添加任何解释或多余信息。
# Here is the context that you can use to answer the question:
# Context:
   {context}
# OutputFormat:
   输出格式如下:
   "answer": "T" 或 "F" 或 "U" 或 "R"
# Here is user's question:
   {question}
# Answer:
   - 请你回答:
```

Relevant documents (k=10) were retrieved from `FAISS` vectorstore based on similarity to the given question. Empirical evaluation tested k values of 3, 10, and 20; among them, `k=10` consistently yielded the best performance.

2.2 RAPTOR-based RAG

To overcome the limitations of Plain RAG - specifically, its inability to capture long-range dependencies and multi-level document structures —we adopted the `RAPTOR` (*Recursive Abstractive Processing for Tree-Organized Retrieval*, Sarthi et al. 2024) framework for the Chinese Factivity Inference task. Unlike Plain RAG, which retrieves a flat list of relevant chunks, `RAPTOR` generates a multi-layer abstraction tree of the reference corpus, enabling the model to reason with different levels of semantic granularity.

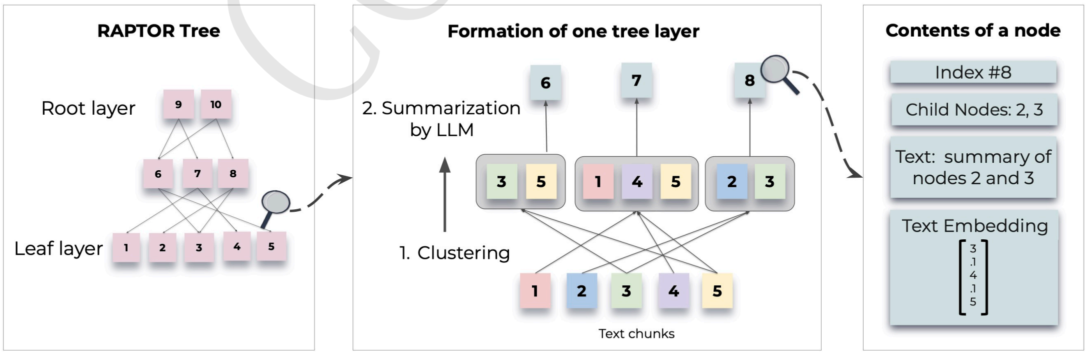


Figure 1: `RAPTOR` Text Clustering Process (Sarthi et al., 2024)

In our implementation, we began by chunking the factivity-specific bibliography documents using a window size of 500, which was an empirically reasonable choice made under memory and processing constraints, though not necessarily optimal. This size allowed for more coherent units of discourse to be preserved while reducing the total number of nodes for recursive summarization.

The resulting document hierarchy included 221 leaf-layer documents (original text segments), 46 cluster-layer documents (low-level abstractions via clustering and summarization), 7 cluster-layer documents (intermediate-level abstractions), and 1 roof-layer document (the highest-level summary of the corpus).

This multi-level document context enabled the model to draw on both surface-level details and abstract generalizations, yielding richer and more accurate factivity inferences. For example, as illustrated in **Appendix**, the RAPTOR mechanism effectively clusters similar factive constructions and generates summaries that reflect conceptual overlap across predicates like “知道”, “意识到”, and “记得”.

All parameters (e.g., chunk size, hierarchy depth, document selection strategy) were optimized specifically for the Chinese Factivity Inference Evaluation (FIE2025). The findings here may not directly generalize to other domains or languages without task-specific adaptation.

3 Experiments

Experiments were conducted on a combined test set of 2,038 artificial and natural instances, following the “non-finetuned track (不微调赛道)” of CCL25-Eval Task 4. We evaluated three large language models—`GPT-4.1`, `DeepSeek-R1-14B`, and `Gemini-2.5-pro-preview-05-06`—under three inference strategies: `Baseline (No-RAG)`, `Plain RAG`, and `RAPTOR-based RAG`. Each configuration was evaluted by its total accuracy across test sets.

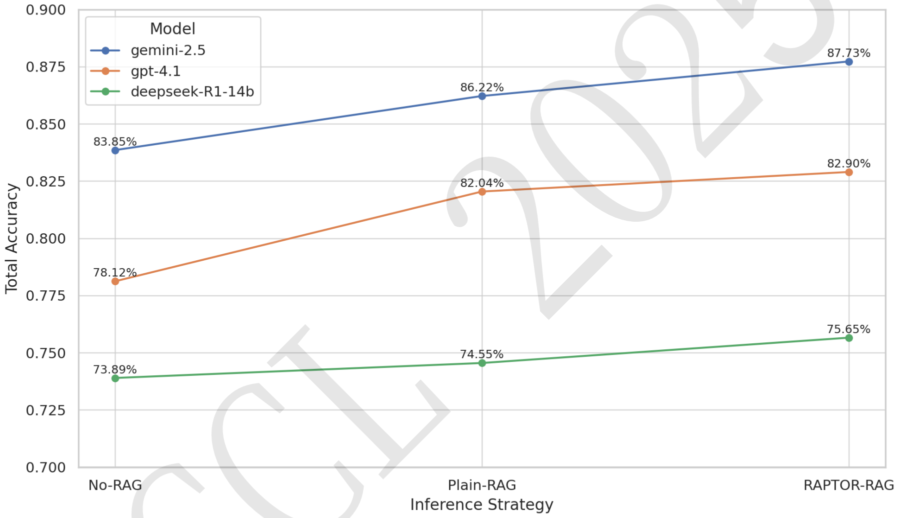


Figure 2: Total Accuracy by Inference Strategy and Model. All experiments were conducted under **a *zero-shot* setting, without incorporating supervised examples**, to prioritize generalizability over dataset adaptation.

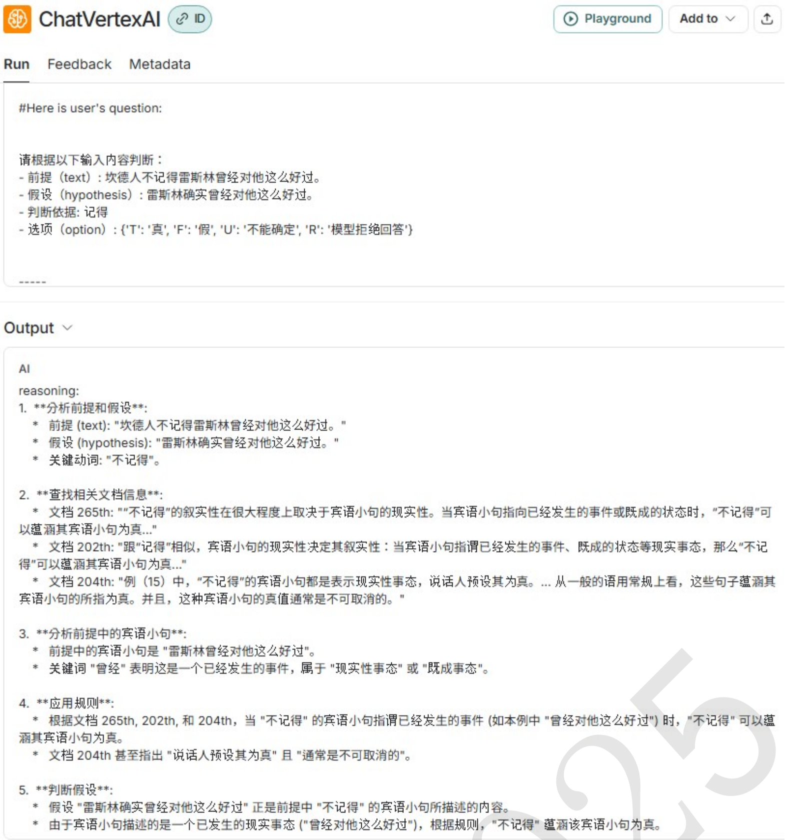
As shown in Fig. 2, performance improves steadily from baseline to RAPTOR-based RAG in all of three models. `Gemini 2.5` consistently outperforms the others under all configurations, achieving the highest overall accuracy.

The RAPTOR-based approach consistently demonstrates superior performance. This outcome underscores RAPTOR’s ability to incorporate multiple levels of abstraction within document structures, allowing for more refined and accurate language inference.

4 Discussion

To qualitatively assess how well our model utilized the contextual information retrieved, we present the following representative example processed by the RAPTOR-based RAG pipeline.

Several retrieved documents consistently support the interpretation that the predicate “不记得” (*does not remember*) can imply the truth of its complement clause when referring to a past event. For instance, Document 265th explains that such factivity implications typically arise



```
#Here is user's question:


请根据以下输入内容判断：
- 前提（text）：坎德人不记得雷斯林曾经对他这么好过。
- 假设（hypothesis）：雷斯林确实曾经对他这么好过。
- 判断依据: 记得
- 选项（option）：{'T': '真', 'F': '假', 'U': '不能确定', 'R': '模型拒绝回答'}


-----
```

```
AI
reasoning:
1. **分析前提和假设**:
   *  前提 (text): "坎德人不记得雷斯林曾经对他这么好过。"
   *  假设 (hypothesis): "雷斯林确实曾经对他这么好过。"
   *  关键动词: "不记得"。

2. **查找相关文档信息**:
   *  文档 265th: ""不记得"的叙实性在很大程度上取决于宾语小句的现实性。当宾语小句指向已经发生的事件或既成的状态时，"不记得"可
以蕴涵其宾语小句为真..."
   *  文档 202th: "跟"记得"相似，宾语小句的现实性决定其叙实性：当宾语小句指谓已经发生的事件、既成的状态等现实事态，那么"不记
得"可以蕴涵其宾语小句为真..."
   *  文档 204th: "例（15）中，"不记得"的宾语小句都是表示现实性事态，说话人预设其为真。... 从一般的语用常规上看，这些句子蕴涵其
宾语小句的所指为真。并且，这种宾语小句的真值通常是不可取消的。"

3. **分析前提中的宾语小句**:
   *  前提中的宾语小句是 "雷斯林曾经对他这么好过"。
   *  关键词 "曾经" 表明这是一个已经发生的事件，属于 "现实性事态" 或 "既成事态"。

4. **应用规则**:
   *  根据文档 265th, 202th, 和 204th，当 "不记得" 的宾语小句指谓已经发生的事件 (如本例中 "曾经对他这么好过") 时，"不记得" 可以蕴
涵其宾语小句为真。
   *  文档 204th 甚至指出 "说话人预设其为真" 且 "通常是不可取消的"。

5. **判断假设**:
   *  假设 "雷斯林确实曾经对他这么好过" 正是前提中 "不记得" 的宾语小句所描述的内容。
   *  由于宾语小句描述的是一个已发生的现实事态 ("曾经对他这么好过")，根据规则，"不记得" 蕴涵该宾语小句为真。
```

Figure 3: Representative inference case using RAPTOR-based RAG

when the embedded clause denotes an event that has already occurred. Furthermore, Documents 202 and 204 emphasize that the facivity of “不记得”-type constructions depends on the real-world status of the complement, and that, in standard pragmatic usage, these constructions are generally understood to presuppose the truth of the embedded propositions that are typically non-cancelable. This example highlights how RAPTOR’ s multi-layered retrieval strategy allows the model to leverage both detailed and abstract knowledge, which proves especially effective in complex semantic tasks such as Chinese factivity inference.

5 Conclusion

This study demonstrates that structuring domain knowledge on Chinese factivity and integrating it through RAG enables more reliable and interpretable inference with LLMs. In particular, our experiments indicate that the RAPTOR-based RAG approach consistently outperforms both No-RAG and Plain RAG strategy in overall accuracy. It is important to note that the RAPTOR-RAG system is solely based on our curated bibliography, independent of 2,038-item test set provided by the organizers. We aimed to explore the generalizability and applicability of the RAPTOR framework rather than fine-tuning the system to a specific dataset.

The strength of RAPTOR lies in its ability to represent textual information at multiple abstraction levels, e.g., leaf, cluster, and roof layers, allowing models to access both granular detail and conceptual generalizations simultaneously. This hierarchical context proved particularly effective in handling the nuanced factive and anti-factive constructions in Chinese. The observed gains from integrating linguistic knowledge into RAG prompting affirm the viability of knowledge-augmented prompting as a robust alternative to parameter-level fine-tuning. In the context of Chinese semantic reasoning, our approach integrates domain knowledge from Chinese linguistics to enhance inference accuracy while maintaining generalizability. Future work may also incorporate few-shot prompting strategies to complement the current approach.

Appendix. RAPTOR abstraction example text of 265th document

叙实性推理（Factivity Inference, FI）和反事实推理（Counter-Factual Inference, CFI）是语义理解中与事件真实性判断密切相关的两种推理形式，统称为真实性推理（Factuality Inference, FactI）。叙实性推理主要依赖于谓词（如动词）来表达事件的真实性。例如，通过分析句子“约翰不知道罗昆是中国人”中的动词“知道”，可以推断出“罗昆是中国人”这一事实。而反事实推理则主要通过反事实条件句来表达，例如“要不是消防队来得及时，大火就要烧到顶楼了”这一句子中，可以推断出“消防队确实来得很及时”和“大火确实没有烧到顶楼”这两个事实。

在叙实性推理中，动词“记得”通常被认为具有较为稳定的叙实性，即使在情态和疑问等复杂的语境中也是如此。例如，在句子“母亲，您还记得吧？是谁殴打您、抓住您的头发，将您从二楼拖到楼下？”中，通过“记得”这个动词，可以推断出“有人殴打您、抓住您的头发，将您从二楼拖到楼下”这一事实。

与“记得”相对，动词“不记得”则表现出更大的主观性和模糊性，其对宾语小句的真值蕴涵更加复杂。例如，在最高法庭对江青进行犯罪事实调查时，江青常用“不记得”作为答复，这可能意味着她真的没有印象，也可能是出于否认或抵赖的目的。类似地，在其他例子中，“不记得”也被用来表达对某些事件的模糊记忆或委婉否认。

“不记得”的叙实性在很大程度上取决于宾语小句的现实性。当宾语小句指向已经发生的事件或既成的状态时，“不记得”可以蕴涵其宾语小句为真；而当宾语小句指向尚未发生的事件或可能但尚未出现的状态时，“不记得”通常蕴涵其宾语小句为假。此外，宾语小句中通常会有时体标记等形式手段来显示相关事态的现实性。

在实际应用中，“不记得”常被用作一种委婉的否认方式。例如，在某些政治或社交场合中，使用“不记得”可以比直接否认更加礼貌，或为对方留有余地。这种表达方式在政治语境中尤为常见，因为它可以在不直接否定某一事实的同时，表达出对某一事件或人物的不满或异议。

综上所述，叙实性推理和反事实推理在语义理解中扮演着重要角色。通过对动词“记得”和“不记得”的分析，可以更深入地理解事件的真实性及其在不同语境中的表达方式。这不仅有助于语言学研究，也为实际的语言应用提供了重要的理论支持。

References

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Daniel Ziembicki, Karolina Seweryn, and Anna Wróblewska. 2024. Polish natural language inference and factivity: An expert-based dataset and benchmarks. *Natural Language Engineering*, 30(2):385–416.

李新良 and 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, (2):194–215.

李新良 and 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.

袁毓林. 2020a. “忘记”类动词的叙实性漂移及其概念结构基础. 中国语文, (5):13.

袁毓林. 2020b. “记得”的叙实性漂移及其概念结构基础. 语言教学与研究, (1):12.

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