System Report for CCL25-Eval Task 4: Prompting, Scheduling, and Arbitration Strategies for Chinese Factivity Inference

Daohuan Liu, Lun Xia, Yuxuan Zhang, Xinyu Yang, Fanzhen Kong
School of Foreign Languages,
Huazhong University of Science and Technology,
Wuhan, Hubei 430074, China
{liudh, xiaxx, zhangyuxuan, aceyyyyu, kongfz}@hust.edu.cn

Abstract

This report presents the methodology and findings of prompting large language models (LLMs) for Chinese Factivity Inference (FI). We evaluated five LLMs, among which DeepSeek-R1 demonstrated the best overall performance. A combination of Chain-of-Thought (CoT), few-shot, and system-level instructions were combined for final prompting. Additionally, we introduced a pairwise task scheduling strategy and a multi-agent disagreement arbitration mechanism to further enhance inference quality. Experimental results show that the integration of prompting, scheduling, and arbitration strategies significantly improves performance, with DeepSeek-R1 achieving 91.7% overall accuracy on the evaluation set. The report also highlights findings regarding LLM behavior on FI tasks and outlines potential directions for future improvement.

Keywords: Factivity Inference , Chinese , LLMs , Prompt Engineering , Evaluation

1 Introduction

Factivity refers to the property of certain verbs (*predicate*) to presuppose the truth of their complement clauses (*hypothesis*), thereby reflecting the interactive subjectivity inherent in linguistic expression (*text*); Factivity Inference (FI) is a semantic understanding task that involves determining the factual status of an embedded event(Yuan, 2020). To assess the performance of Large Language Models (LLMs) on FI, we evaluated five reasoning-capable LLMs: DeepSeek-R1 (DeepSeek-AI, 2025), Ernie-X1[1], Qwen3 (Qwen-Team, 2025), Doubao-1.5-thinking[2], and Hunyuan-T1[3]. The evaluation results are presented in Table 1.

Model	Test Set (Art/Nat %)	Eval Set (Art/Nat %)
deepseek-r1	**87.67/82**	**87.52/88.61**
ernie-x1-turbo-32k	88.67/81.86	82.97/86.49
qwen3-235b-a22b	86/82.29	84.95/85
doubao-1-5-thinking-pro-250415	86/84.71	86.74/86.49
hunyuan-t1-20250403	76/78.29	/

Table 1: Performance of five tested LLMs.

Among the five models evaluated, DeepSeek-R1 was ultimately selected as the primary model for further analysis due to its consistently superior performance on both artificial (Art) and natural (Nat) datasets. In contrast, Hunyuan-T1 was excluded from downstream experiments because of its unsatisfactory accuracy on the test set. The remaining three models, i.e., Ernie-X1, Qwen3, and Doubao-1.5-thinking, were retained and employed in a multi-model disagreement arbitration framework. This

[1]https://cloud.baidu.com/doc/qianfan-docs/s/7m95lyy43
[2]https://console.volcengine.com/ark/region:ark+cn-beijing/model/detail?Id=doubao-1-5-thinking-pro
[3]https://cloud.tencent.com/document/product/1729/104753

decision was motivated by the desire to enhance overall inference robustness through cross-model consensus and to better understand inter-model variability in FI.

2 Methods and Results

Our final evaluation on the evaluation dataset yielded a total accuracy of 91.7%, with 94.13% on the Art set and 90.78% on the Nat set. This performance was achieved through the integration of three key strategies: prompt engineering, task scheduling, and multi-agent disagreement arbitration. This section outlines these three methods in detail, along with their respective contributions to model performance.

2.1 Prompt Engineering: CoT, Few-Shot, and System

For the Art set and the Nat set which exhibit distinct characteristics, optimal prompts were identified by combining three techniques: Chain-of-Thought (CoT) (Wei et al., 2022), Few-Shot (Brown et al., 2020), and System instructions (see Appendix A). The performance variations of DeepSeek-R1 under different prompts are summarized in Table 2.

Prompt Version	Art %	Nat %	Total %
Baseline	70.67	79.43	76.8
CoT	87.33	82	83.6
CoT+Fewshot	**94.67**	86	88.6
CoT+Fewshot+System	94.33	**87.42**	**90.4**
CoT+Fewshot+System+Predicates	94.67	86.57	89

Table 2: Performance under different prompts on the test set.

CoT. Our CoT pipeline explicitly incorporated four analytical steps: 1) propositional parsing, 2) predicate categorization and rule application, 3) negation analysis, and 4) semantic verification. Since factivity shifts due to negation were found to be extremely rare in the Nat set, the negation analysis step was omitted when prompting this subset in order to shorten the reasoning chain without compromising accuracy.

Few-Shot. Few-shot exemplars were tailored separately for the Art (Figure 1) and Nat (Figure 2) dataset to better align with their respective characteristics. Each example explicitly follows our designed CoT framework to guide the model through the inference process.

System. System-level instructions were found to improve overall accuracy, with a particularly significant effect observed on the Nat dataset, though little to no impact was observed on the Art dataset.

Predicate Marking. Contrary to expectations, explicitly highlighting the key predicate within the prompt failed to improve the overall performance. We assume such foregrounding may disrupt the full CoT process, prompting the model to rely on the surface-level semantics of the predicate rather than engaging in contextual interpretation. This hypothesis, however, warrants further empirical investigation.

2.2 Task Scheduling: Pairwise Strategy

Inspired by contrastive learning (Robinson et al., 2020), we adopted a pairwise scheduling strategy in which tasks sharing the same *text* but with opposing *hypothesis* statements were submitted concurrently. Experimental results (Table 3) indicate that this approach consistently outperforms conventional single-task scheduling. The findings suggest that simultaneous processing of contradictory hypotheses may help LLMs establish sharper decision boundaries and enhance discriminative reasoning.

Scheduling Strategy	Art %	Nat %	Total %
single (Test/Eval Set)	85.67/87.52	80.71/88.61	82.2/88.31
pairwise (Test/Eval Set)	**87/89.07**	**85.42/89.26**	**85.9/89.21**

Table 3: Impact of pairwise task scheduling on the test set (Prompt: CoT only).

Proceedings of the 24th China National Conference on Computational Linguistics, pages 146–151, Jinan, China, August 11–14, 2025.

2.3 Multi-Agent Disagreement Arbitration

Borrowing the principles of ReAct prompting (Yao et al., 2023b) which integrates reasoning and action steps through iterative feedback, and self-consistency strategy (Wang et al., 2022) which aggregates multiple reasoning outputs to improve final accuracy, we designed a multi-agent disagreement arbitration mechanism to enhance prediction reliability.

In this framework, an arbitrator model re-evaluates samples on which the four primary models produce conflicting predictions. Responses with full agreement among the four are accepted directly, while disagreements are resubmitted to the arbitrator for final adjudication, accompanied by anonymized references to the original outputs. As shown in Table 4, the introduction of arbitration led to a consistent improvement in overall accuracy across all arbitrator configurations. Among the four, Doubao-1.5-thinking achieved the highest performance on the Test set[4]. The gain was particularly notable on the Nat set, while no improvement was observed on the Art set. The prompt used for arbitration is provided in Appendix (Figure 3).

Arbitration Strategy	Art %	Nat %	Total %
Primary Model Average	**87.09**	82.715	83.99
Arbitrator:deepseek-r1	84.67	85.43	85.2
Arbitrator:ernie-x1-turbo-32k	86	85.26	85.49
Arbitrator:qwen3-235b-a22b	86.33	85.14	85.5
Arbitrator:doubao-1-5-thinking-pro-250415	86.33	**86.29**	**86.3**

Table 4: Performance of disagreement arbitration using different arbitrator models on the test set (Prompt: CoT only).

3 Analysis and Discussion

This section analyzes the main findings observed during our experiments, highlighting key patterns, insights, and future directions derived from the results.

3.1 Effects of Reasoning Capabilities

To evaluate the effect of deep reasoning capabilities on factivity inference, we conducted comparative experiments using three representative models series—Qwen3, Hunyuan, and Doubao—each of which offers both standard and reasoning-enhanced versions. While deep reasoning was enabled by default in DeepSeek-R1, Ernie-X1, and Doubao-1.5-thinking, it remained configurable in Qwen3 and across different models of the Hunyuan and Doubao series. As shown in Table 5, enabling reasoning consistently led to substantial accuracy improvements across both the Art and Nat subsets, indicating that deep reasoning significantly enhances model performance in the FI task.

Model	Art %	Nat %	Total %
qwen3-235b-a22b (Reasoning Off)	75	76.67	76.16
qwen3-235b-a22b (Reasoning On)	**86**	**82.29**	**83.26** (↑7.1)
hunyuan-turbos-20250416	68	73.71	72
hunyuan-t1-20250403	**76**	**78.29**	**77.6** (↑5.6)
doubao-1-5-pro-32k-250115	79	81.29	80.6
doubao-1-5-thinking-pro-250415	**86**	**84.71**	**85.1** (↑4.5)

Table 5: Impact of deep reasoning across models on the test set. In each pair, the upper row represents the version or model without reasoning.

[4]Although doubao-1.5-thinking turned out to be the best arbitrator on the test set, our leaderboard score was based on Deepseek-R1 arbitration on the Eval set.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 146–151, Jinan, China, August 11–14, 2025.

3.2 Comparative Evaluation of Prompting Strategies

In addition to the techniques used in our final prompt configuration, we evaluated several alternative prompting methods, including Tree-of-Thought (ToT) (Yao et al., 2023a) and Automatic Prompting Engineer (APE) (Zhou et al., 2022). Table 6 presents the performance of each technique when applied individually to the DeepSeek-R1 model.

Prompting Techniques	Art %	Nat %	Total %
Baseline (Zero-Shot)	70.67	79.43	76.8
APE	82.33	77.57	79
ToT	80	79.14	79.4
Few-Shot	77	81.71	80.3
CoT	**87.33**	**82**	**83.6**

Table 6: Impact of prompt engineering methods on model responses on the test set.

Among all tested methods, CoT delivered the most significant performance gains, closely followed by few-shot. These results highlight the effectiveness of step-wise reasoning and exemplar-based guidance in improving semantic inference accuracy. While prompts generated by ToT and APE exhibited moderate improvements on the Art set, they resulted in performance declines on the Nat set, suggesting that such automated or tree-structured prompting strategies may be less effective in guiding model understanding for real-world, naturally occurring FI tasks.

3.3 Insights from Multi-Agent Arbitration

Disagreement as an Error Signal. We found that predictions receiving unanimous agreement from all four models were highly reliable: 97.79% on the Art set and 92.17% on the Nat set. As the agreement threshold was relaxed—for example, accepting predictions as agreed when only three models concurred—the number of agreed samples increased, but the accuracy of those agreed answers declined noticeably (see Table 7). Meanwhile, the overall accuracy after arbitration also dropped, since fewer incorrect responses remained eligible for correction.

Agreement Threshold	Agreed Count	Agreed Acc (%)	Arbitrated Total Acc (%)
$= 4/4$	226/549	**97.79/92.17**	**86.3**
$\geq 3/4$	276/646	92.75/87.46	86.0
$\geq 2/4$	282/659	91.84/86.49	85.9

Table 7: Statistics under different agreement threshold on the test set (Arbitrator: Doubao-1.5-thinking).

Nevertheless, even under the most lenient condition where only two models agreeing was considered sufficient[5], the accuracy of the agreed subset still exceeded the average performance of all four models individually (see Table 4). This underscores that multi-model consensus is a strong predictor of correctness. Conversely, samples with divergent answers were disproportionately concentrated in lower-accuracy regions, reinforcing the idea that disagreement acts as a reliable proxy for task difficulty or ambiguity. This suggests that disagreement level may serve as a fine-grained indicator for arbitration confidence.

Limitations and Future Work. As shown in Figure 4, arbitration produced negligible improvement on the Art set. On the one hand, we attribute this to a performance saturation effect: the baseline accuracy on Art was already high (87.09%), limiting the margin for correction. On the other hand, our current arbitration prompt simply mirrored the baseline structure without leveraging prompting techniques (Figure 3). Future work may enhance arbitration effectiveness by adopting more specialized prompts tailored for disagreement resolution. Additionally, replacing or augmenting the agent pool, or using a stronger external model as the arbitrator, may further refine final outputs and better handle borderline cases.

[5] A 2-out-of-4 agreement only includes (2 vs. 1 vs. 1) cases. Cases like (2 vs. 2) are still categorized as disagreement.

Acknowledgements

This work is supported by the Fundamental Research Funds for the Central Universities (HUST: No.YCJJ20252111).

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Qwen-Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Yulin Yuan. 2020. Factivity and factuality: Two guiding mechanisms in linguistic reasoning. *Linguistic Research (YU WEN YAN JIU)*, (1):1–9.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

A Appendix

你是一个严谨的文本事实核查助手，请根据用户提供的【文本】内容，判断【假设】的真值情况（真/假/不能确定）。
按照以下步骤分析：
1.命题解构：根据“假设”(H)内容，从“文本”中识别出与H紧密相关的关键谓语动词(V)和命题从句(P)。
2.动词叙实性分类与规则应用：根据V的叙实性类型（正叙实/反叙实/非叙实）判断P的真值。正叙实动词（“看见/发现/认识到/羡慕/抱怨/批评”等）描述客观情况或表达感知状态，P=真；反叙实动词（“污蔑/幻想”等）P=假；非叙实动词（“猜测/声称/怀疑”等）一般是主语的主观推断，P=不能确定。若是非叙实，则直接结束分析，输出“不能确定”。注意叙实性可能受时态影响，实际分类请结合文本语境信息综合判断。例如“幻想”用于已发生的事物时，一般起反叙实作用；而用于描述未来的事物时，则是非叙实动词。
3.否定分析：检查V是否被否定词否定。若被否定，且V是否定敏感词（“承认/证明/证实/表明”），则触发叙实降级，即由正叙实 → 非叙实，P=不能确定。
4.语义检查：检查P与H的语义是否一致。若一致，则P的真值即为最终答案；若不一致，则P的真值取反作为最终答案。
例1【文本】媒体没有披露小李会前往现场。【假设】小李会前往现场。（答案：V=披露 → 正叙实 → 非否定敏感词，无视否定 → H=P=真）
例2【文本】媒体谎称小李私生活没有不检点。【假设】小李私生活不检点。（答案：V=谎称 → 反叙实 → 无否定 → P=假 → H=¬P=真）
例3【文本】小张没有猜到小李会忘带钥匙。【假设】小李带了钥匙。（答案：V=猜到 → “到”字表结果，正叙实 → 非否定敏感词，无视否定 → P=真 → H=¬P=假）
例4【文本】小张并非没有感觉小李喜欢自己。【假设】小李喜欢小张。（答案：V=感觉 → 没有“到/出”后缀，是非叙实 → P=不能确定 → H=不能确定）
例5【文本】小张不记得要给女友买花。【假设】小张不要给女友买花。（答案：V=记得 → 正叙实 → 非否定敏感词，无视否定 → P=真 → H=¬P=假）
例6【文本】小张哀叹公司将会走下坡路。【假设】公司将会走下坡路。（答案：V=哀叹 → 通常是正叙实，但此处描述未来事件且并非主语可控，故为非叙实 → P=不能确定 → H=不能确定）
完成下列任务，回答只能列出<任务ID:答案>格式(例如<Art_999:假>)的答案列表，禁止回复其他解释性内容：

Figure 1: The final prompt used for Art set pairwise questioning.

你是一个严谨的文本事实核查助手，请根据用户提供的【文本】内容，判断【假设】的真值情况（真/假/不能确定）。
按照以下步骤分析：
1.命题解构：根据“假设”(H)的内容，从“文本”中识别出与H紧密相关的关键谓语动词(V)和命题从句(P)。
2.动词叙实性分类与规则应用：根据V的叙实性类型（正叙实/反叙实/非叙实），判断P的真值（真/假/不能确定）。正叙实动词（“看见/发现/认识到/羡慕/抱怨/表明”等）描述客观情况或表达感知状态，P=真；反叙实动词（“污蔑/幻想”等）P=假；非叙实动词（“猜测/声称/怀疑”等）一般是主语的主观推断，P=不能确定。若是非叙实，则直接结束分析，输出“不能确定”。注意叙实性可能受时态影响，实际分类请结合文本语境信息综合判断。例如“幻想”用于已发生的事物时，一般起反叙实作用；而用于描述未来的事物时，则是非叙实动词。
3.语义检查：检查P与H的语义是否一致。若一致，则P的真值即为最终答案；若P与H语义相反，则P的真值取反作为最终答案。
示例任务：
例1【文本】他不禁哀叹，社会经济就将跌入无底深渊中。【假设】社会经济确实不会跌入无底深渊。（答案：V=哀叹 → 可为正/非叙实，此处蕴含主语对外界、未来的推测，是非叙实 → P=不能确定 → H=不能确定）
例2【文本】被访者表示，他感觉这个药物确实让伤口恢复的更快了。【假设】这个药物确实没有让伤口恢复的更快。（答案：V=感觉 → 可为正/非叙实，此处是说话人对自身生理感知的描述，为正叙实 → P=真 → H=¬P=假）
例3【文本】他早已猜到这样的政策会激起民愤。【假设】这样的政策确实激起了民愤。（答案：V=猜到 → 不同于“猜/猜测”，后缀“到”蕴含正叙实作用 → P为真 → H=P=真）
例4【文本】美国外交官批评中国在朝核问题上毫无作为。【假设】中国在朝核问题上确实毫无作为。（答案：V=批评 → 可为正/非叙实，此处表达的是说话者的意见立场，而非批评某个具体行为，是非叙实 → P=不能确定 → H=不能确定）
例5【文本】很多人幻想通过炒股一夜暴富。【假设】通过炒股确实不能一夜暴富。（答案：V=幻想 → 可为反/非叙实，此处蕴含主语对外界、未来的推测，是非叙实 → P=不能确定 → H=不能确定）
完成下列任务，回答只能列出<任务ID:答案>格式(例如<Nat_999:假>)的答案列表，禁止回复其他解释性内容：

Figure 2: The final prompt used for Nat set pairwise questioning.

下面是一个叙实性推理任务，即根据“文本”判断“假设”内容是否为真，但四个大模型对此的判断结果之间存在冲突。
题号：{d_id}
文本：{text}
假设：{hypothesis}
模型判断：{answer}
请你结合文本语义、语言逻辑和常识，同时也参考各个模型的建议，做出最终仲裁（真/假/不能确定）。回答格式为“<题号:最终答案>”（例如<Art_999:真>），不能回复其他内容。

Figure 3: The final prompt used for disagreement arbitration.