Overview of CCL25-Eval Task 4: Factivity Inference Evaluation 2025

Guanliang Cong[1] **Junchao Wu**[1] **Yang Chen**[1] **Tianqi Xun**[1]
Derek F. Wong[1] **Bin Li**[2] **Yulin Yuan**[1,3,*]
[1]University of Macau, MacaoSAR, China 510006
[2]Nanjing Normal University, Nanjing, China 210023
[3]Peking University, Beijing, China 100871
[1]{yc27731,yc47493,yc47706,yc47705,derekfw,yulinyuan}@um.edu.mo
[2]libin.njnu@gmail.com [3]yuanyl@pku.edu.cn

Abstract

This paper presents the results of the FIE2025, a shared task aimed at evaluating the ability of Large Language Models (LLMs) to perform factivity inference on Chinese texts: whether LLMs can correctly discern the veridical information of propositions encoded in the complement clauses. The responses to the task mirror the extent to which LLMs can grasp the implicit truth judgments made by human speakers through texts, as well as their subjective stances. Such a capability is crucial for autonomous inference in intelligent agents and for achieving fluid human–AI interaction. The task was hosted on the Alibaba Tianchi platform and evaluated through two tracks: with and without finetuning. A mixed dataset was constructed, combining both synthetic sentences and authentic corpus instances. The dataset comprises a total of about 3,000 items labeled by expert linguists, including **845** (300+545) manually created items and **2,143** (700+1,443) items selected from existing corpus. **404** results proposed by **74** teams were successfully submitted to Tianchi system. Overall, under current technological conditions, the key to successful factivity inference lies in whether LLMs effectively identify different types of predicates and various contextual conditions from the given texts. Models that support long-context prompt inputs tend to achieve the best inference performance when provided with numerous shots. This shared task deepened our understanding of the factivity phenomenon in Chinese, expanded the influence of factivity research within the field of natural language processing, and provided an exploratory precedent for future activities focusing on factivity inference in Chinese and potentially other languages.

Keywords: Factivity Inference , Natural Language Inference , LLM Evaluation , Chinese , Chinese Information Processing

1 Introduction

In linguistic research, factivity is a complex phenomenon that spans multiple levels of analysis, interacting with the issues in lexical, syntactic, semantic, and pragmatic domains. Broadly speaking, factivity refers to a conventional relationship between the matrix predicate and the truth value of the proposition embedded within the complement clause. For example:

- (a) *Xiaozhou zhidao* [*Xiaobo lai=le*].
  Xiaozhou know Xiaobo come=PRF
  'Jo knows that Bo came.'

- (b) *Xiaozhou huang-cheng* [*Xiaobo lai=le*].
  Xiaozhou lie-claim Xiaobo come=PRF
  'Jo lies that Bo came.'

*Corresponding Author.

- (c) *Xiaozhou xiangxin* [*Xiaobo lai=le*].
  Xiaozhou believe Xiaobo come=PRF
  'Jo believes that Bo came.'

(a) is a complex sentence that contains an embedded event clause. *Xiaobo lai le* (Bo came) functions as the complement of the matrix verb *zhidao* (know) and contains a proposition. One can infer from (a) that the encoded proposition *Xiaobo lai le* is taken to be true in the actual world. This type of inference concerning the veridicality of a proposition appears to be closely tied to the choice of the matrix predicates. When the matrix verb *zhidao* is replaced with *huangcheng* (falsely claim), as in (b), the truth value of the embedded clause is conversely inferred to be false. In contrast, if *zhidao* is replaced with *xiangxin* (believe), forming (c), the truth value of the embedded clause becomes uncertain.

Interestingly, for this kind of predicates, the use of negation operator usually cannot overturn the interpretation of truth value of its complement clause. For example:

- (d) *Tamen yishidao* [*jumian yijing buke-wanhui*].
  they realize situation already ir-reversible
  'They realized that the situation was already irreversible.'

- (e) *Tamen meiyou-yishidao* [*jumian yijing buke-wanhui*].
  they NEG-realize situation already ir-reversible
  'They didn't realize that the situa tion was already irreversible.'

The truth value of the complement clause in (d), *jumian yijing bukewanhui* (the situation was already irreversible), would not be stirred even when its predicate, *yishidao* (realize), is negated by *meiyou* (not), that is, one can infer the same veridical proposition from both (d) and (e). Kiparsky & Kiparsky (1970) consider this phenomenon as a lexical presupposition that the matrix verb presupposes the truth of its complement clause. It shows that the knowledge employed here is a kind of analytical linguistic knowledge, which involves the analysis of semantic relations among linguistic components, and is relatively independent of world knowledge.

Yuan (2020a) proposed that factivity inference (FI) serves as a key navigational mechanism in linguistic inference, which is often accompanied by clear formal linguistic cues in Chinese. Research on factivity helps illuminate the mechanisms underlying human language comprehension and inference. Moreover, it provides syntactic and semantic foundations essential for enabling Artificial Intelligence (AI) to perform Natural Language Processing (NLP) tasks, such as Recognizing Textual Entailment (RTE), hallucination detection, belief revision, etc.

This motivation led us to design and organize FIE2025, which is a shared task aimed at evaluating the FI capabilities of Large Language Models (LLMs) on Chinese texts. Throughout this task, we hope to promote growing academic interest in this topic within the NLP community.

The remainder of this paper is structured as follows: Section 2 reviews related work; Section 3 outlines the theoretical foundations and task design; Section 4 describes the data construction; Sections 5 and 6 present and analyze the evaluation results; Section 7 concludes with a discussion of future directions.

2 Related work

2.1 Discussion from theoretical linguistics

In the field of theoretical linguistics, research on factivity primarily treats it as a criterion for classifying predicates, exploring the semantic features of predicates through the lens of factivity and the syntactic-semantic distinctions among predicates with different factivity types. Classic discussions on this issue can be found since Kiparsky & Kiparsky (1970), Karttunen (1971), and Leech (1981). Some studies have investigated the island effects associated with factive structures from the perspective of formal syntax, such as Adams (1985), Rooryck (1992), Oshima (2007), Kastner (2015), and Schwarz & Simonenko (2018). The syntactic realization of factivity also varies cross-linguistically (Kastner, 2015; Altiagoitia

and Elordieta, 2016; Wiemer, 2014; Ohta, 1991). Besides, some studies have addressed factivity phenomena in Chinese, including Yuan (2014; 2020b; 2020c; 2021), Li (2014; 2018; 2020), Zhang (2020), Chen & Zhang (2020), Li & Yuan (2016; 2017), Yuan & Kou (2018), and Lin & Zhang (2024), etc.

2.2 Discussion from computational linguistics

In the field of computational linguistics, FI is considered a task of understanding, specifically a sub-type of textual entailment. In nature, the task involves determining the truth value of one statement based on the content of another. Evaluating language models' ability to perform FI thus falls under Natural Language Understanding (NLU) assessments, and more precisely, under Natural Language Inference (NLI) evaluation. Many existing works have implicitly or explicitly touched upon this issue:

Datasets: Many datasets have incorporated factivity-related sentences into their design, even though few datasets explicitly define factivity as a central concern. In the case of comprehensive benchmarks, MNLI (Williams et al., 2018; Bowman et al., 2015) in GLUE (Wang et al., 2018) collects around 390,000 items including some FI instances. As for targeted evaluations, Saurí & Pustejovsky (2009) built a large-scale corpus, FactBank, focused on event factuality. Several relevant datasets were developed subsequently, such as UW (Lee et al., 2015), MEANTIME (Minard et al., 2016), UDS (White et al., 2016), MegaVeridicality (White and Rawlins, 2018; White et al., 2018; Rudinger et al., 2018), and CGC (Markowska et al., 2023), etc.

Models: In terms of the tested models, most evaluations to date have focused on LSTM and BERT-based classifiers, as seen in works like White et al. (2018); Rudinger et al. (2018); White & Rawlins (2018); Ross & Pavlick (2019); Jiang & de Marneffe (2021); Cohen (2021); Markowska et al. (2023), etc. Evaluations conducted on decoder-only models such as ChatGPT or LLaMA are still relatively rare, though some recent works have begun to address this gap (Basmov et al., 2024; Kosinski, 2024).

Languages: As for corpus languages, the vast majority of NLI datasets are constructed based on English corpora. Only a few non-English datasets have included FI as a component, such as XNLI (Conneau et al., 2018) and a recent Polish factivity dataset (Ziembicki et al., 2024).

3 Task definition and annotation

As mentioned, FI is considered as an NLU task, specifically an RTE task, aiming at determining the truth value of one statement (the proposition of the embedded clause) based on the content of another (the whole sentence). Usually, the sentence used for entailing is named the entailing sentence or premise, while the sentence being entailed is called the entailed sentence or hypothesis.(Dagan et al., 2005; Poliak, 2020) In this shared task, for convenience, we keep up with Yuan & Wang (2009) and use the term *text* and *hypothesis* to entitle the entailing and the entailed sentences.

3.1 Task flow

The shared task proceeds as follows: Participants are first required to choose whether to engage in the finetuning (abbr. FT) track, the non-finetuning (abbr. Non-FT) track, or both, depending on their experimental goals and available resources. There is no restriction on the choice of LLM series or versions; participants are free to use any model, regardless of its architecture or scale. Any external resources or knowledge databases deemed beneficial are allowed in the Non-FT track, provided that they do not involve gradient-based optimization or any modification of the model's weights, which could compromise the fairness and integrity of the competition. Using the dataset provided by the organizers, participants are required to construct input prompts and submit them to the selected models via API access. The models should then generate responses for each item in the dataset. These responses must be collected, formatted according to the specified requirements, and submitted in JSON format to the Tianchi evaluation platform for official assessment.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 166–180, Jinan, China, August 11–14, 2025.

3.2 Annotation

To enable participating teams to efficiently utilize the dataset, we encode all information that is potentially required during the evaluation task into a JSON file. There are 7 keys in each object, which together constitute the basic structure of the data from the sample set of the synthetic corpus. The 7 keys include the data id (*"d_id"*); the word form of predicate (*"predicate"*); the entailing sentence or premise (*"text"*); the entailed sentence or hypothesis (*"hypothesis"*); the formalization of the predication or output of LLMs (*"output"*); the factive type of the predicate (*"type"*), which is concealed in the authentic sets; and the expert answer (*"answer"*), which is only provided in the sample sets. In the key of *"option"*, there are 3 embedded keys {*T(rue), F(alse), U(ncertain)*}, corresponding respectively 3 potential labeling results of textual inference {*Entailment, Contradiction, Neutral*}. An extra key *R(ejection)* was proposed to help to collect the potential bug items during the pilot tests.

4 Data

The sample set and the unlabeled test set are available under a Creative Commons Attribution 4.0 International License.[1]

4.1 Scale and provenance

The dataset of FIE2025 contains two subsets: a sample set used as the validation set during pre-competition stage, and a test set used for the competition stage.

The sample set contains 300 synthetic corpus items and 700 authentic corpus items, while the test set includes 545 synthetic corpus items and 1,443 authentic corpus items.

It's worth noting that there were 36 items about *jiazhuang (pretend)* and 2 items about *zhuangzuo (pretend)* excluded during the pre-competition stage due to potential semantic controversy.[2]

Some scholars pointed out that some components in the authentic context can impact the interpretation of the truth value of the complement clause.(Yuan, 2021; Ju, 2023; Yuan, 2024) To compare the prediction accuracies under different contextual richness, we collect data in different ways. The synthetic corpus is handmade to assure that each premise provides only an idealized human-created textual environment. While all the data in the authentic corpus are extracted and revised from CCL, which is a comprehensive Chinese corpus developed by Peking University.[3]

4.2 Predicates selection

The predicates in the synthetic corpus are derived from Li (2020), while those in the authentic corpus are slight modifications based on the synthetic corpus predicates.

To evaluate the generalization ability of the techniques employed by participating teams across different predicates within the FI task, we deliberately limited the number of predicates allowed in the sample set. Specifically, we randomly selected half of all predicates to be included in the sample set, ensuring that some predicates in the test set also appeared in the sample set, while the other half remained unseen during sample exposure. The quantity of collected predicates is listed in the Table 1.

Quantity of predicates	Synthetic corpus	Authentic corpus	Total
Sample set	66	19	70
Test set	73	70	78
Total	79	71	82

Table 1: The quantity of predicates of different subsets and corpus resources.

[1]Dataset: https://github.com/UM-FAH-Yuan/FIE2025/. License details: http://creativecommons.org/licenses/by/4.0/.

[2]See our explanation in the Appendix A.

[3]See http://ccl.pku.edu.cn:8080/ccl_corpus/.

4.3 Labeling

To simplify dataset construction, we retain the classical three-way classification scheme for labeling. All annotation results are selected from a fixed set of three labels: *T*, *F*, *U*, representing true, false, and uncertain, respectively.

For the human-constructed corpus, the data were handmade by one Ph.D. student and labeled and revised by three expert annotators. While the authentic corpus was extracted from CCL, filtered with rigorous standard and labeled by three expert annotators.

It has been reported that human are facing an issue of low inter-annotator agreement in semantic annotation of NLU datasets.(Nie et al., 2020; Zhou et al., 2021) To ensure both accuracy and consistency of the annotations, all three annotators are PhD students in linguistics with long-term research experience in factivity and a record of collaborative work.

The annotation procedure is as follows: first, the three annotators label the data independently; then, their results are compared and cases of disagreement are discussed; finally, the gold label is determined by majority vote among the annotators. The thinking process behind all annotation decisions is reconstructable enough to explain the assigned labels.

5 Shared task evaluation and results

The shared task was conducted on the Alibaba Tianchi System,[4] which provides a compositive platform of dataset management, output evaluation, and a quickly updated leaderboard.

The sample dataset was made available on the Tianchi platform one month prior to the competition phase, whereas the test set was exclusively accessible during the 7-day competition window. Participating teams were permitted a maximum of two daily submissions while maintaining unrestricted access to the leaderboard.

5.1 Assessment metric

The shared task adopts overall accuracy as the assessment metric. Specifically, the score of a submission is calculated as the total number of correctly answered items in both the set of synthetic corpus and the set of authentic corpus, then divided by the total number of items in the test set. As mentioned in Section 4.1, some bug items were skipped during score calculation.

$$total\_acc = \frac{correct\_art + correct\_nat}{total\_art + total\_nat} \times 100\%$$

This formula shows how we calculate the grades of submitted predictions, where *total_acc* refers to total accuracy, *art* refers to synthetic corpus, and *nat* refers to authentic corpus.

5.2 Baseline

As the task baseline, we adopt the Qwen2-7B-Instruct model to evaluate textual entailment relations between predicates and their complement clauses. In Non-FT, we obtain model responses solely by sending the designed prompts, without any model modifying. In FT, we provided a baseline to perform instruction fine-tuning on Qwen2-7B-Instruct. Specifically, we utilized LLaMA-Factory to achieve efficient instruction fine-tuning with the Low-Rank Adaptation (LoRA).[5] BF16 mixed-precision mode was enabled during training to improve computational efficiency. The LoRA rank was set to 8 , freezing most of the model weights and optimizing only a small number of parameter matrices, thereby significantly reducing memory requirements. The maximum input sequence length was set to 2048, and the learning rate was 0.0001. Finally, our finetuned model was combined with the same prompt template to produce high-quality task-specific responses. All experiments were conducted on a single NVIDIA A100 GPU.

As shown in Table 2, in Non-FT, our baseline achieves an accuracy of 53.74% on the synthetic corpus and 69.86% on the authentic corpus, with a weighted overall accuracy of 65.02% . In FT, we observe an accuracy of 94.66% on the synthetic corpus and 88.93% on the authentic corpus, with a weighted overall

[4]See https://tianchi.aliyun.com/.

[5]For details, see https://github.com/hiyouga/LLaMA-Factory/.

Model: Qwen2-7B-Instruct	Non-FT	LoRA-FT
art_acc	53.74%	94.66%
nat_acc	69.86%	88.93%
weighted mean	65.02%	90.65%

Table 2: This table shows the baseline results in two tracks.

accuracy of 90.65%. Obviously, the LoRA fine-tuning significantly enhances model performance on FI, and is therefore introduced as a demonstrative method.

5.3 Results of participants

We received a total of 218 registrations on the Tianchi platform, with 74 teams successfully submitting results. Among them, 30 teams had identifiable affiliations. There are 22 teams who submitted rough reports, and 9 of these were recommended for expansion into detailed system reports.

Number of result submissions During the competition stage, the Tianchi platform received a total of 539 result submission attempts, of which 404 were successfully submitted. Among these, 305 were from Non-FT track, and 99 were from FT track.

Ranking results The highest-scoring submissions in Non-FT and FT tracks are shown in Table 3 and Table 4, respectively. The vast majority of teams outperformed our baseline in the Non-FT track, while in the FT track, about half of the teams exceeded the baseline performance (Non-FT: top 97.14%; FT: top 50%).[6] We believe this difference reflects the high level of enthusiasm and engagement among the participating teams, who improved their best scores through repeated experimentation and multiple submissions.

6 Analysis of evaluation techniques

In this section, we provide a general overview of the methods and strategies reported by the teams. The analysis is based on the 9 system reports and 13 rough technical reports submitted by participants after the competition stage. Teams who did not submit any report are not counted in the statistics. For detailed technical information, please refer to the system reports of this shared task.

Overall, most of the participants realized that the key for LLMs to conduct the FI task lies in identifying the varying properties of predicates and the contextual conditions under which they appear. Most teams adopted the strategy of pre-constructing an external knowledge base through certain automated methods before drawing on it to customize prompts in accordance with different predicate types or specific contextual conditions. In terms of instructing manner, participating teams proposed novel and diverse approaches to enhance the consistency and reliability of model responses.

6.1 Model selection

As shown in Figure 1,the uses of 12 series covering 19 models that were reported among submissions, including Deepseek, Claude, ChatGPT, Qwen, Ernie, LLaMA, Kimi, GLM, Spark, Hunyuan, Doubao, and Gemini. Among them, the most popular large language models came from the Deepseek and Qwen series. The larger fan area indicates that the series or model was selected by more participants.

6.2 Prompt engineering

Our participants employed a wide range of prompt strategies in both tracks, characterized by both quantity and diversity. Most of the reported strategies focus on guiding models to conduct fine-grained recognition and understanding. There are 8 core methods of prompt engineering, covering Chain-of-Thought (CoT), few-shot, role assigning, factive types indicating, contextual components indicating, model-summarized rules, linguistic knowledge providing, and data enhancement. See more detailed information in Table 5 in the Appendix B.

[6] In terms of all submissions, 93.18% surpassed the baseline in Non-FT track, while 30% exceeded the baseline in FT track.

Ranking	Name code	Team No.	Affiliation	Total_acc	Art_acc	Nat_acc
1	BIT-1 (Li et al., 2025b)	37	Beijing Institute of Technology	**94.01%**	97.80%	**92.58%**
2	CAS (Yan et al., 2025)	49	Chinese Academy of Sciences	93.76%	97.80%	92.24%
3	BNU (Li et al., 2025a)	12	Beijing Normal University	93.51%	**98.17%**	91.75%
4	KU-HNU (Zhao et al., 2025)	30	Kunming University; Hunan Normal University	93.41%	97.61%	91.82%
5	CAS-UCAS	44	Chinese Academy of Sciences; University of Chinese Academy of Sciences	92.66%	97.80%	90.71%
6	RUC (Zhang et al., 2025)	15	Renmin University of China	92.61%	95.41%	91.55%
7	HKPU-SU (Wang et al., 2025)	9	The Hong Kong Polytechnic University; Sichuan University	92.40%	96.15%	90.99%
8	HUST (Liu et al., 2025)	25	Huazhong University of Science and Technology	91.70%	94.13%	90.78%
9	UCASS-CASS	3	University of Chinese Academy of Social Sciences; Chinese Academy of Social Sciences	91.60%	96.15%	89.88%
10	BISTU-CNU	29	Beijing Information Science & Technology University; Capital Normal University	90.95%	93.21%	90.09%
11	CASS	28	Chinese Academy of Social Sciences	90.09%	94.44%	88.45%
12	JLU	7	Jilin University	89.99%	95.05%	88.09%
13	XJU	48	Xi'an Jiaotong University	89.84%	93.55%	88.45%
14	BIT-2	43	Beijing Institute of Technology	89.19%	97.25%	86.14%
15	UW	6	University of Washington	88.43%	92.84%	86.76%
16	BJU	51	Beijing Jiaotong University	88.37%	96.95%	85.14%
17	CMIT (Gu et al., 2025)	2	China Mobile (Hangzhou) Information Technology Co., Ltd.	87.63%	88.71%	87.23%
18	JNU	35	Jinan University	87.27%	91.74%	85.59%
19	CTBU	53	Chongqing Technology and Business University	85.28%	89.61%	83.65%
19	NUPT	41	Nanjing University of Posts and Telecommunications	84.00%	90.14%	81.69%
20	DWU-CNU (Park and Lee, 2025)	16	Duksung Women's University; Chungbuk National University	82.70%	80.55%	83.51%
21	UIR	42	University of International Relations	82.34%	80.18%	83.16%
			Baseline method		**65.02%**	

Table 3: This table shows the best scores of participants in Non-Finetuning track. The score is weighted average of *Art_acc* and *Nat_acc*.

Ranking	Name code	Team No.	Affiliation	Total_acc	Art_acc	Nat_acc
1	CAS (Yan et al., 2025)	49	Chinese Academy of Sciences	**93.96%**	**97.80%**	**92.52%**
2	CMIT (Gu et al., 2025)	2	China Mobile (Hangzhou) Information Technology Co., Ltd.	93.41%	97.06%	92.03%
3	BIT-1 (Li et al., 2025b)	37	Beijing Institute of Technology	92.61%	94.86%	91.75%
3	BNU (Li et al., 2025a)	12	Beijing Normal University	92.61%	95.60%	91.48%
5	XJU	48	Xi'an Jiaotong University	92.40%	96.51%	90.85%
6	BJU	51	Beijing Jiaotong University	91.61%	95.70%	90.07%
			Baseline method		**90.65%**	
7	BIT-2	43	Beijing Institute of Technology	87.78%	91.38%	86.42%
8	BITSU-CNU	29	Beijing Information Science & Technology University; Capital Normal University	80.96%	79.75%	81.42%

Table 4: This table shows the best scores of participants in Finetuning track. The score is weighted average of *Art_acc* and *Nat_acc*.

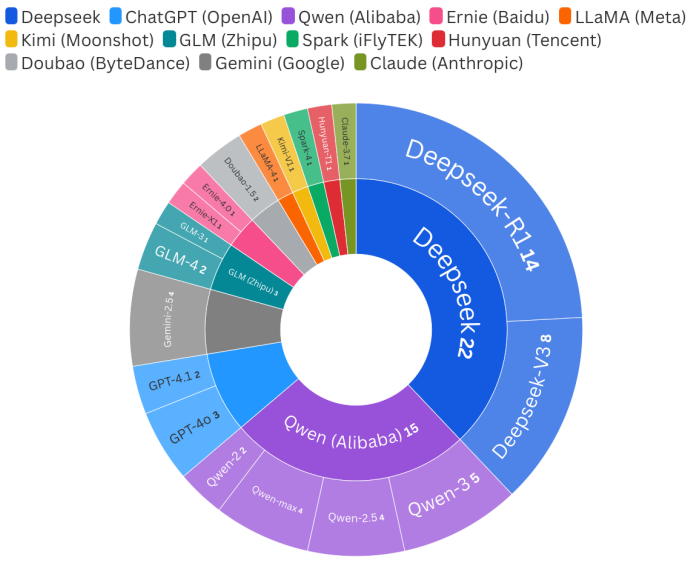

Figure 1: Distribution of teams selecting models, categorized by series and versions. The inner segments represent different model series, while the outer segments detail the corresponding model versions.

6.3 Structured conduction strategies

To cooperate with fine-grained prompt designing, most of the participants also improved the manner of acquiring responses of models, through preprocessing of the corpus and consistency voting. We summarize the reported structured conduction strategies into 8 core methods, including repeated questions voting, multiple models voting, multiple prompts voting, text revision, distinctive questions, word vector extraction, syntactic parsing by knowledge graph, and quantified labels. See more detailed information in Table 6 in the Appendix C.

6.4 Finetuning Method

There are 18 teams that participated in the competition of finetuning track. According to the rough report, most of participants adopted the FT technique of Low-Rank Adaptation (LoRA) as we used in baseline test. The reported FT methods include LoRA, full-parameter finetuning, Pseudo-label learning, data filtering, and model fusion.

6.5 Outstanding Combos

When reviewing the ranking results, we observed that teams achieving high scores in one track often tended to perform well in the other track as well, provided that they participated in both. Given that different teams exhibited varying preferences in model selection, this suggests that certain combined strategies adopted by top-ranking teams were indeed more effective at eliciting strong FI performance from LLMs compared to others. Below, we briefly describe the evaluation techniques used by the top-ranking teams in each track.

- **Team #37 (Ranking #1 in Non-FT track):** The Non-FT approach of team #37 can be characterized by three key steps: First, they used a pilot lightweight model to filter and select high-quality prompt templates. Next, they employed a recently released high-performance model, Gemini 2.5 Pro Preview-0506, to elaboratively analyze the data in the training set and summarize response rules tailored to different types of predicates. Finally, they arranged these rules into thinking processes as a part of CoT.
- **Team #49 (Ranking #1 in FT track):** The FT approach of Team #49 can also be characterized by three key steps: First, they selected a reasoning model through comparative experiments. Second, they designed prompts using a Hierarchical Chain-of-Thought (HCoT) strategy to guide the

model in gradually extracting key information. Third, they applied a Parameter-Efficient FineTuning (PEFT) approach to train a LoRA module, and used pseudo-label learning for data augmentation to expand the training dataset. Finally, the trained LoRA module was integrated into the backbone model for final evaluation.

By cross-comparing the evaluation strategies and results of these two top-performing teams, we observe the following:

A. Different types of models exhibit significant differences in FI performance, while the model's inherent capabilities may be the key factor determining its FI performance ceiling. Non-finetuned LLMs that support long-context prompt inputs, such as Google's latest Gemini 2.5 Pro Preview-0506, can achieve optimal FI performance (94.01%) under specific prompting conditions, especially when provided with a large number of shots. Based on this crucial discovery, we presume that: As long as being enlightened by enough high-quality shots, LLMs are already capable of performing excellently; Furthermore, as LLM pretraining techniques continue to advance, it is likely that LLMs will reach the level of human-expert on FI tasks, even under 0-shot conditions or with other simple prompts.

B. When comparing different evaluation techniques, it becomes evident that model selection has a greater impact on FI performance than other evaluation strategies. According to our statistics, designing predicate-specific prompts is pretty common (12 out of 21 teams adopted this approach); however, both the top two teams in the Non-FT track, Team #37 and Team #49, used the latest high-performing model (Gemini 2.5 Pro Preview-0506). We presume this advantage may stem from such models' ability to process longer prompt contexts. A longer input context allows for the inclusion of more guiding information, enabling the model to better learn how humans make factivity judgments in similar contexts, thereby producing responses that more closely align with expert annotations.

C. Finetuning techniques can make up for performance defects caused by poor prompt design, but they do not significantly improve FI performance when high-quality prompts or advanced models are used. This suggests that, although parameter finetuning can serve as a shortcut or remedial resort to temporarily boost FI in mediocre models or with poorly designed prompts, it may not be a necessity for SOTA or other high-performance models.

7 Conclusions and future directions

FIE2025 is the first large-scale shared task dedicated to evaluating factivity inference capabilities of LLMs in Chinese. Through this comprehensive evaluation involving 74 teams and 404 valid submissions, we have gained valuable insights into the current state and limitations of this issue.

Our key findings can be summarized as follows: **A.** Models supporting long-context inputs, particularly Google's Gemini-2.5-pro-preview, demonstrate superior performance when provided with comprehensive contextual information and well-designed prompts, achieving up to 94.01% accuracy in the Non-FT track. **B.** Predicate-type-specific prompting strategies, combined with few-shot learning and CoT, constitute the most effective approaches. The majority of high-performing teams adopted external knowledge base construction methods to customize their prompts according to different predicate categories and contextual conditions. **C.** Though finetuning techniques may compensate for suboptimal prompt design, they do not provide significant improvements when applied to well-designed prompts or advanced models, suggesting that parameter-level adaptation may serve as a remedial measure rather than a fundamental requirement for achieving optimal performance in FI tasks.

From a theoretical linguistics perspective, FIE2025 validates the importance of factivity as a linguistic phenomenon that requires sophisticated semantic and pragmatic understanding. The task demonstrates that FI involves complex interactions between lexical semantics, syntactic structures, and contextual pragmatics, which align with classical theoretical discussions from Kiparsky & Kiparsky (1970) to contemporary Chinese linguistics research. From the standpoint of computational linguistics, our findings suggest that the current LLMs have developed substantial capabilities in handling analytical linguistic knowledge, though their performance remains constrained by the quality of prompt engineering and model architecture. The excellent performance of long-context models indicates that models may benefit significantly from abundant shots when conducting FI tasks.

Acknowledgements

This work was supported by the Research & Development Grant for Chair Professor (CPG2025-00008-FAH) and the Start-up Research Grant (SRG2022-00011-FAH) of the University of Macau. We also gratefully acknowledge the Faculty of Arts and Humanities of the University of Macau for its generous financial support of this shared task.

References

Marianne Adams. 1985. Government of empty subjects in factive clausal complements. *Linguistic Inquiry*, 16(2):305–313.

Xabier Altiagoitia and Arantzazu Elordieta. 2016. On the semantic function and selection of Basque finite complementizers. *Complementizer Semantics in European Languages*, pages 379–411.

Victoria Basmov, Yoav Goldberg, and Reut Tsarfaty. 2024. Simple linguistic inferences of large language models (llms): Blind spots and blinds. *arXiv preprint arXiv:2305.14785*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zhenyu Chen and Xinhua Zhang. 2020. *Factivity and Facticity (Xushixing yu Shishixing)*. Shanghai Educational Publishing House, Shanghai.

Michael Cohen. 2021. Exploring roberta's theory of mind through textual entailment. *philpapers.org: https://philpapers.org/archive/COHERT.pdf*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Sunyan Gu, Taoyu Lu, Siqi Liu, Kan Guo, and Yan Shao. 2025. System report for ccl25-eval task 4: Factivity inference based on dynamic few-shot learning. In *the 24th China National Conference on Computational Linguistics (CCL)*, Jinan, Shandong, August 11-14.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2021. He thinks he knows better than the doctors: Bert for event factuality fails on pragmatics. *Transactions of the Association for Computational Linguistics*, 9:1081–1097.

Chen Ju. 2023. *The Expression and semantic inference of factivity in Chinese Mandarin (Xiandai Hanyu Xushixing de Biaoda Xingshi he Xiangguan de Yuyi Tuili Jizhi Yanjiu)*. Ph.D. thesis, Peking University.

Lauri Karttunen. 1971. Some observations on factivity. *Paper in Linguistics*, 4(1):55–69.

Itamar Kastner. 2015. Factivity mirrors interpretation: The selectional requirements of presuppositional verbs. *Lingua*, (164):156–188.

Paul Kiparsky and Carol Kiparsky. 1970. Fact. In M. Bierwisch and K. Heidolph, editors, *Progress in Linguistics*, pages 143–173. The Hague: Mouton.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648.

Geoffrey Leech. 1981. *Semantics: the study of meaning*. Middlesex: Penguin Books, England, second edition.

Xinliang Li and Yulin Yuan. 2016. Conditions on the Interpretation of the Complement Clauses of the Counter-factive Verbs (Fanxushi Dongci Binyu Zhenjia de Yufa Tiaojian Jiqi Gainian Dongyin). *Contemporary Linguistics (Dangdai Yuyanxue)*, 18(2):194–215.

Xinliang Li and Yulin Yuan. 2017. On the factivity of zhidao and its confidence variation under different grammatical environments ("Zhidao" de Xushixing Jiqi Zhixindu Bianyi de Yufa Huanjing). *Studies of the Chinese Language (Zhongguo Yuwen)*, (1):42–52+127.

Hongyu Li, Zhihui Yang, and Renfen Hu. 2025a. System report for task 4 of ccl25-eval: Research on factivity inference method based on multi-strategy knowledge fusion. In *the 24th China National Conference on Computational Linguistics (CCL)*, Jinan, Shandong, August 11-14.

Zequn Li, Yuanhao Zhong, and Chengliang Chai. 2025b. System report for ccl25-eval task 4: Application of macroscopic pattern prompting and efficient finetuning in factivity inference. In *the 24th China National Conference on Computational Linguistics (CCL)*, Jinan, Shandong, August 11-14.

Xinliang Li. 2014. *Research on Factivity of verbs based on Modern Chinese Xiandai Hanyu Dongci de Xushixing Yanjiu*. Ph.D. thesis, Peking University.

Xinliang Li. 2018. On factivity shift of verbs of ganjue ("Ganjue"-lei dongci de xushixing jiqi piaoyi wenti yanjiu). *Language Teaching and Linguistic Studies (Yuyan Jiaoxue yu Yanjiu)*, (5):65–75.

Xinliang Li. 2020. *Research on Factivity of verbs based on Modern Chinese*. Peking University Press, Beijing.

Dingfan Lin and Heyou Zhang. 2024. Factivity and its floating: Evidence from three types of Mandarin verbs (Xushixing Jiqi Piaoyi: Laizi Sanlei Dongci de Zhengju). *Foreign Language Teaching and Research (Waiyu Jiaoxue yu Yanjiu)*, 56(03):359–369+479.

Daohuan Liu, Lun Xia, Yuxuan Zhang, Xinyu Yang, and Fanzhen Kong. 2025. System report for ccl25-eval task 4: Prompting, scheduling, and arbitration strategies for chinese factivity inference. In *the 24th China National Conference on Computational Linguistics (CCL)*, Jinan, Shandong, August 11-14.

Magdalena Markowska, Mohammad Taghizadeh, Adil Soubki, Seyed Abolghasem Mirroshandel, and Owen Rambow. 2023. Finding common ground: Annotating and predicting common ground in spoken conversations. *arXiv preprint arXiv:2311.01273*.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begona Altuna, Marieke Van Erp, Anneleen Schoen, and Chantal Van Son. 2016. Meantime, the newsreader multilingual event and time corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4417–4422.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? *arXiv preprint arXiv:2010.03532*.

Amy Snyder Ohta. 1991. Evidentiality and politeness in Japanese. *Issues in Applied Linguistics*, 2(2):211–238.

David Y. Oshima. 2007. On factive islands: Pragmatic anomaly vs. pragmatic infelicity. In *New Frontiers in Artificial Intelligence: JSAI 2006 Conference andWorkshops*, volume 4384. Springer.

Minjun Park and Seulki Lee. 2025. System report for ccl25-eval task 4: From plain to hierarchical—knowledge-augmented prompting for chinese factivity inference. In *the 24th China National Conference on Computational Linguistics (CCL)*, Jinan, Shandong, August 11-14.

Adam Poliak. 2020. A survey on recognizing textual entailment as an nlp evaluation. *arXiv preprint arXiv:2010.03061*.

Johan Rooryck. 1992. Negative and factive islands revisited. *Journal of Linguistics*, 28(2):343–374.

Alexis Ross and Ellie Pavlick. 2019. How well do nli models capture verb veridicality? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2230–2240.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. *arXiv preprint arXiv:1804.02472*.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43:227–268.

Bernhard Schwarz and Alexandra Simonenko. 2018. Factive islands and meaning-driven unacceptability. *Natural Language Semantics*, (26):253–279.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Yu Wang, Qian Yang, Ke Liang, Yiheng Yang, Yu Zhai, and Churen Huang. 2025. System report for ccl25-eval task 4: A factivity detection agent based on rag and predicate similarity methods. In *the 24th China National Conference on Computational Linguistics (CCL)*, Jinan, Shandong, August 11-14.

Aaron Steven White and Kyle Rawlins. 2018. The role of veridicality and factivity in clause selection. In *Proceedings of the 48th annual meeting of the north east linguistic society*, volume 3, pages 221–234.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on universal dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723.

Aaron Steven White, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2018. Lexicosyntactic inference in neural models. *arXiv preprint arXiv:1808.06232*.

Björn Wiemer. 2014. On the semantic function and selection of Basque finite complementizers. In Elisabeth Leiss and Werner Abraham, editors, *Modes of modality: Modality, typology, and Universal Grammar. Studies in Language Companion Series*, pages 127–166.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*.

Qiang Yan, Yixing Fan, and Yunfei Zhong. 2025. System report for ccl25-eval task 4: Chinese factivity inference based on hierarchical chain-of-thought construction and efficient fine-tuning of reasoning models. In *the 24th China National Conference on Computational Linguistics (CCL)*, Jinan, Shandong, August 11-14.

Yulin Yuan and Xin Kou. 2018. On the factivity of zhidao and its confidence variation under different grammatical environments ("Zhidao" de Xushixing Jiqi Zhixindu Bianyi de Yufa Huanjing). *Studies in Language and Linguistics (Yuyan Yanjiu)*, 38(2):1–13.

Yulin Yuan and Minghua Wang. 2009. The type hierachy and inference machanism of textual entailment (Wenben yunhan de leixing cengji he tuili jizhi). In Xiliang Guo and Guoyao Lu, editors, *Chinese Linguistics Volomn 3(Zhongguo Yuyanxue (Di 3 Ji))*, pages 123–138. Peking University Press.

Yulin Yuan. 2014. On the Factivity and NPI-licensing Fuction of the Implicit Negative Verbs in Mandarin Chinese (Yinxing-fouding dongci de xushixing he jixiang-yunzhun gongneng). *Linguistic Sciences (Yuyan Kexue)*, 13(6):575–586, November.

Yulin Yuan. 2020a. Factivity and facticity: Two nevigation mechanisms of language inference (xushixing he shishixing: Yuyan tuili de liangzhong daohang jizhi). *Linguistic Research (Yuwen Yanjiu)*.

Yulin Yuan. 2020b. Factivity variation of jide and its underlying conceptual structure ("Jide" de Xushixing Piaoyi Jiqi Gainian Jiegou Jichu). *Language Teaching and Linguistic Studies (Yuyan Jiaoxue yu Yanjiu)*, (1):36–47.

Yulin Yuan. 2020c. Factivity variation of Wangji and its underlying conceptual structure ("Wangji"lei Dongci de Xushixing Piaoyi Jiqi Gainian Jiegou Jichu). *Studies of the Chinese Language (Zhongguo Yuwen)*, (5):515–526+638, September.

Yulin Yuan. 2021. A Study on Ambiguous Interpretations of the Pretend-Sentence in Chinese: From the Perspective of Linguistic "Polyphony" (Cong Yuyan de "Duoshengxing" Kan "Jiazhuang"ju de Jiedu Qiyi). *Chinese Journal of Language Policy and Planning (Yuyan Zhanlue Yanjiu)*, 6(05):77–90.

Yulin Yuan. 2024. Factivity reversion and magic effect in non-factive construction "X bugan xiangxin Y": "Unwilling suspension of disbelief due to the contradiction with fact" ("X bugan xiang xin Y"goushi de xushixing nizhuan gongenng yu moshu xiaoying—-biaoshi "shishi dianfu xinnian hou bu-qingyuan-de xuanzhi bu-xinren"de xinli). *Studies of Chinese Language (Zhongguo Yuwen)*, (04):387–399+510.

Xiaoyi Zhang, Jiaqi Lu, Da Zhang, Xiaoyu Chen, and Dawei Lu. 2025. System report for ccl25-eval task 4: Factuality reasoning in large language models via factual consistency classification and contextual features. In *the 24th China National Conference on Computational Linguistics (CCL)*, Jinan, Shandong, August 11-14.

Xinhua Zhang. 2020. *A study on the Madarin Factive Predicates (Hanyu Xushi Weici Yanjiu)*. Fudan University Press, Shanghai.

Peixiang Zhao, Mingzhu Li, Liya Mei, Fang Wang, Nianxin Gao, and Lang Zhao. 2025. System report for ccl25-eval task4: The pragmatic reasoning mechanism of factivity inference of modern chinese verbs. In *the 24th China National Conference on Computational Linguistics (CCL)*, Jinan, Shandong, August 11-14.

Xiang Zhou, Yixin Nie, and Mohit Bansal. 2021. Distributed nli: Learning to predict human opinion distributions for language reasoning. *arXiv preprint arXiv:2104.08676*.

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.

Appendices

A Why do we skip out *jiazhuang* and *zhuangzuo (pretend)*?

Jiazhuang or *zhuangzuo (pretend)* are at a special position in the study of factivity, as it relates to biological behaviors of deception or camouflage. In the case of humans, the object of pretending can be either an action or a state. When the subject pretends to be in a certain state, the proposition embedded in the complement clause is generally inferred to be false, and there is little disagreement on this point. While the complication arises when the subject pretends to perform an action. In such cases, scholars haven't come to a consensus on whether the action described in the embedded clause has actually been performed.

We believe that the root of this disagreement lies in the nature of human pretense, which involves two components: the intent to deceive and the execution of an action or display of a state. When someone pretends to do something, they perform a real action driven by a false intention. We hold the view that the act of pretending inherently involves the subject performing some sort of deceptive behavior, even though researchers may differ in their assessment of the quality or extent of the action's realization.

Therefore, to avoid divergences in answers caused by scholarly disagreement, we decided to retain all questions involving *pretend* (so that the academic community can further study this issue), but to exclude them from the final scoring regardless of the model's output. At the same time, in order to maintain the overall size of the evaluation set, we selected additional 38 items from other verbs to refill the gap.

B Prompting engineering methods

No.	Methods on Prompt Engineering	Introduction	Reported Teams
1	Chain-of-Thought (CoT)	This method allows non-reasoning models to simulate a reasoning process by demonstrating the thought steps one by one.	#2, #3, #6, #7, #12, #16, #25, #28, #29, #30, #35, #37, #41, #42, #44, #49
2	Few-shot	This method incorporates a variable number of question-answer examples in the prompt. These examples may be fixed or dynamically generated through some computational approach.	#2, #3, #6, #12, #25, #28, #37, #44, #48, #51
3	Role assignment	This method starts the prompt with a sentence assigning the model a role as an expert suited to solving the task.	#2, #7, #12, #15, #16, #25, #28, #35, #41, #43, #44, #48
4	Factive types indicating	This method includes factivity classification labels of the predicates in the prompt. Some teams used the labels provided by us, while others created their own or modified ours.	#3, #6, #9, #12, #15, #25, #28, #29, #30, #35, #41, #42, #43, #44, #48, #49, #51
5	Contextual components indicating	This method directs the model's attention to contextual elements beyond the predicate that may influence inference results, such as modal words, discourse adverbs, negative components, and/or event types of clauses.	#3, #6, #12, #15, #25, #30, #41, #42, #43, #44, #48, #49
6	Model-summarized rules	This method first uses the model to summarize a task-specific set of answering rules, which are then included in the prompt.	#29, #37, #42, #49
7	Linguistic knowledge providing	This method enriches the prompt with dictionary definitions, custom definitions, or relevant factuality-related linguistic knowledge from academic literature.	#3, #9, #12, #16, #28, #29, #30, #42, #48, #49
8	Data enhancement	This method generates new, self-constructed data using the model to mimic the given examples.	#2, #9, #12, #48

Table 5: This table shows the 8 main prompting engineering methods and the teams who reported them.

C Structured conduction strategies

No.	Structured Conduction Strategies	Introduction	Reported Teams
1	Repeated questions voting	This method seeks the consistency of a single model's answers under one prompt by using internal voting across multiple repeated queries.	#25, #44
2	Multiple models voting	This method sends the same prompt to multiple models to obtain multiple responses and determines the final answer through a voting mechanism.	#2, #12, #25, #41, #48, #49
3	Multiple prompts voting	This method sends several parallel prompts with the same objective to a single model to collect multiple responses, and then determines the final answer via voting.	#2, #7, #12, #15, #16, #25, #28, #35, #41, #43, #44, #48
4	Texts revision	This method first modifies the given test set data, obtains model responses based on the revised data, and then maps the answers back to the original data, aiming to reduce distributional bias in the original inputs.	#9, #43
5	Distinctive questions	This method designs customized prompts based on the different factivity types of predicates.	#3, #6, #9, #12, #15, #29, #30, #35, #37, #42, #43, #44
6	Word vector extraction	This method uses external models to extract word embedding vectors from the test set and calculates semantic similarity to provide response examples based on semantically similar predicates.	#29, #51
7	Syntactic parsing by knowledge graph	This method constructs knowledge graph triples from the given dataset to extract key syntactic information about event structures, enabling further automated classification.	#42
8	Quantified labels	This method discards the nominal truth-value labels and instead adopts a continuous labeling scheme along the spectrum of {Counter-factive; Non-factive; Factive}.	#7

Table 6: This table shows the 8 main structured conduction strategies and the teams who reported them.