System Report for CCL25-Eval Task 4: Factivity Inference Based on Dynamic Few-Shot Learning

Sunyan Gu, Taoyu Lu, Siqi Liu, Kan Guo, Yan Shao
China Mobile (Hangzhou) Information Technology Co., Ltd., Hangzhou, China
{gusunyan,lutaoyu,liusiqi,guokan,shaoyan}@cmhi.chinamobile.com

Abstract

This paper presents the implementation approach we employ in the First Chinese Factivity Inference Evaluation 2025 (FIE2025). Factivity inference (FI) is a semantic understanding task related to judging the truth value of events, based on the use of semantic verbal elements, such as "believe", "falsely claim", "realize". We approach factivity inference as a large language model (LLM) based task. We aim to enhance LLM's discriminative capability by adequately integrating the task-specific information via prompts, as well as constructing dynamic few-shot datasets for fine-tuning. Additionally, we incorporate data augmentation and ensemble strategies to further boost the performance. Our approach achieves a score of 93.41% in the official evaluation of the shared task, ranking second in the leaderboard.

Keywords: Factivity Inference , Dynamic Few-shot Learning

1 Introduction

The FIE2025 dataset serves as a valuable resource for factivity inference, featuring a diverse array of content sources such as news reports, literary works as well as artificially synthesized materials in Chinese, as illustrated in Table 1. The dataset is categorized into two distinct groups, namely artifactual samples (Art) and natural samples (Nat). The Art is created manually and reviewed by experts, the Nat is manually selected, organized, and reviewed by experts from the Peking University CCL corpus.

The *Predicate* column refers to the predicates in the sentences. Most of these predicates are verbs, while a small portion are adjectives. The *Type* column indicates the factuality type of the predicates, and it is only available in Art. The *Text* column contains the background context for reasoning. The models utilize the context as the knowledge base to determine the truth of the statements, denoted by the *Hypothesis* column. The *Answer* column provides the response, with three possible values *T*, *F*, and *U*, corresponding to the *Hypothesis* being true, false, and undetermined respectively.

We conduct a statistical analysis of the factual types of predicates in the FIE2025 dataset. It reveals that the majority of predicates are associated only with a singular factuality type, factive, counter-factive or non-factive. Only a limited subset of predicates display multiple factuality attributes. For instances, the predicate "以为" ("think" or "believe") exhibits both factive and counter-factive characteristics, while "假装" ("pretend") demonstrates both factive and counter-factive attributes (Yuan, 2020).

2 Method

Factivity inference can be modeled as mapping linguistic symbols to objective facts, focusing on inferring the factuality implied in the object clauses through factuality predicates (Chen and Jiang, 2018).

In recent years, large language models (LLMs) (Vaswani et al., 2017; Ouyang et al., 2022), represented by the GPT series (Radford et al., 2019; Brown et al., 2020), demonstrate great capabilities in text generation and understanding. We therefore assume that LLMs can be leveraged to comprehend the meaning of *predicates* in the given background context and determine the truth of *hypothesis* sentences

Predicate	Type	Text	Hypothesis	Answer
看见(see)	正叙实 (factive)	小张看见小李去上课了。(Xiao Zhang saw Xiao Li going to class.)	小李去上课了。(Xiao Li went to class.)	T
幻想(fantasize)	反叙实 (counterfactive)	小张幻想自己是富翁。(Xiao Zhang fantasized about being a millionaire.)	小张是富翁。(Xiao Zhang is a millionaire.)	F
估计(estimate)	非叙实 (non-factive)	小张估计小李生病了。(Xiao Zhang estimated that Xiao Li was sick.)	小李生病了。(Xiao Li was sick.)	U
暴露出(expose)	Null	它暴露出一些重大工程建设项目存在着资金管理等方面的混乱和漏洞。(It exposed some major construction projects having issues with fund management and loopholes.)	确实有一些重大工程建设项目存在着资金管理等方面的混乱和漏洞。(Indeed, some major construction projects had issues with fund management and loopholes.)	T
猜(guess)	Null	布罗克竖起大拇指对同事说：“我猜傅也是党员。”(Brock gave a thumbs-up to his colleague and said, “I guess Fu is also a Party member.”)	傅确实也是党员。(Fu was indeed also a Party member.)	U
诬陷(frame)	Null	庞涓“偷梁换柱”，诬陷孙膑带走楚王的珠宝。(Pang Juan “switched the beams” and framed Sun Bin for stealing the King of Chu’s jewels.)	孙膑确实带走了楚王的珠宝。(Sun Bin did indeed steal the King of Chu’s jewels.)	F

Table 1: Samples from the FIE2025 dataset.

by their reasoning capabilities. Additionally, further improvement is expected by using dynamic few-shot learning to construct training data for fine-tuning the LLM.

2.1 Dynamic Few-shot Learning

We formulate factivity inference as a text generation task with LLMs. To enhance model performance, we augment the input prompt with carefully selected exemplars. These examples serve two key purposes: (1) providing task-specific context, and (2) guiding the model toward desired response patterns. Through exposure to such demonstrations, the model better captures the expected output structure, thereby improving its factivity inference accuracy. When these exemplars remain static throughout the dataset, we classify this approach as fixed few-shot learning (Wertheimer et al., 2021; Wang et al., 2020).

While fixed few-shot learning leverages training data exemplars to enhance model predictions, it faces inherent scalability limitations. The approach is constrained by: (1) the context window size of LLMs, which restricts the number of demonstrative examples that can be provided, and (2) the computational overhead associated with processing lengthy prompts, even if context length limitations were hypothetically removed. To address these challenges, dynamic few-shot learning emerges as an effective solution, where examples are randomly sampled from the training set for prompt construction (Weber et al., 2024). Our experiments demonstrate that models fine-tuned with the dynamic approach achieve superior performance on factivity inference tasks compared to their fixed few-shot counterparts.

2.2 Data Augmentation

Data augmentation is commonly employed to address limited training samples or imbalanced label distributions. By generating synthetic data that mimics the original distribution, it enhances both dataset diversity and model robustness. We identify 14 novel predicate verbs only present in the test set. To improve generalization, we leverage Qwen2.5 72B (Bai et al., 2023) to generate additional training data for these new predicates. We generate 238 additional data samples with the prompt in Table 2.

2.3 Model Ensemble

Model ensemble is a widely adopted technique for aggregating predictions from multiple models. To leverage in this approach, we employ dynamic few-shot learning to generate three distinct training datasets, each used to train separate models. Each model independently produces a prediction, and the fi-

【你是一名人工智能数据生成和标注工程师，当前有一项文本数据生成任务,请你思考后务必返回准确的结果。请根据给出的汉语谓词含义，围绕它的使用场景，生成假设依据文本以及假设。其中，假设的内容一般为假设依据文本中谓词后的文本或其简单变种，且判断结果需在真、假、不确定3种结果范围内。例如，谓词：哀叹。假设文本：凯尔特人队的远投使这支东部联盟卫冕冠军队难以招架,29投19中的高效率三分球,只能让76人队哀叹自己的军火库弹尽粮绝。假设：自己的军火库确实是弹尽粮绝【真】；自己的军火库确实还没有弹尽粮绝【假】。现在，请根据谓词'prd'，生成不同假设结果的数据，按照格式：谓词；假设文本；假设，进行输出。】
(You are an AI data generation and annotation engineer. Your current task involves generating textual data. After careful consideration, you must return accurate results. Based on the given Chinese predicate's meaning and its usage context, generate hypothesis premise texts and corresponding hypotheses. The hypothesis should generally be the text following the predicate in the premise (or a simplified variant), with judgment results strictly limited to True (T), False (F), or Uncertain (U). For example, Predicate: 哀叹 (lament); Premise: "The Celtics' long-range shots overwhelmed the Eastern Conference defending champions. With 19 out of 29 three-pointers made at high efficiency, the 76ers could only lament their depleted arsenal." Hypotheses: "Their arsenal was indeed depleted" [T] "Their arsenal was not yet depleted" [F] Now, for the given predicate prd, generate data with varying hypothesis results, formatted as: Predicate; Premise Text; Hypothesis.)

Table 2: Prompt for additional data generation.

nal output is determined through majority voting. This ensemble strategy effectively mitigates overfitting risks while enhancing model robustness through collective decision-making.

3 Experiment

We use the Qwen2.5 series (Bai et al., 2023) as the backbone models. This series provides open source models in multiple versions and scales for different purposes. Our computing resource is a 4 * 80G Nvidia A800 GPU server.In model training, our compute type is bf16, cutoff length is 2048, learning rate is 5e-5 and epoch is 5.The tempture of our inference model is 0.1, max length is 512.

The competition dataset comprises 1,070 samples, which we partitioned into 1,000 training instances and 70 development instances through random selection. The general prediction framework follows three key points:

(1) **Task Familiarization**: We develop foundational prompts to formally define the factivity inference task and facilitate LLMs' understanding of the objective.

(2) **Context Enhancement**: Through dynamic few-shot learning, we augment the base prompts with demonstrative examples to provide richer contextual information.

(3) **Data Completion**: The final training data samples are generated by incorporating target questions into the enhanced prompt structure.

The prompt for factivity prediction is describe in Table 3.

【你是一个文本叙实性推理专家，能够从某些动词性语言成分（如"相信""谎称""意识到"等）的使用推知其他语言成分所描述的相关事件的真实性。 现在请根据输入的文本和其对应的问题进行叙实性判断，有四个可以返回的结果。 分别为真、假、不能确定、其他。真表示hypothesis问题事实性符合输入的文本，假表示hypothesis问题事实性不符合输入的文本，不能确定表示事实性根据当前线索还无法确定，其他表示出现其他问题。 返回结果请以json返回，返回你的判断结果answer。 以下是一些参考用例：{few_shot_samples}】
(You are a factivity inference expert capable of deducing the veracity of described events based on verbal components (such as "believe", "falsely claim", "realize", etc.) in textual input. Your task is to perform factivity judgment on given input text and its corresponding hypothesis question, with four possible return values: True (indicating the hypothesis factually aligns with the input text), False (indicating factual contradiction), Uncertain (when veracity cannot be determined from available evidence), or Other (for exceptional cases). Results should be returned in JSON format with your judgment under the "answer" key. Below are some reference examples: {few_shot_samples})

Table 3: Base prompt for factivity prediction.

3.1 Evaluation Metrics

We use overall accuracy as the evaluation metric:

$$total\_acc = \frac{correct\_art + correct\_nat}{total\_art + total\_nat} \quad (1)$$

"total_acc" signifies the overall accuracy rate. "correct_art" represents the count of correct answers provided by our models on the artifactual samples(Art), while "correct_nat" indicates the count of correct answers given by the models on the natural samples(Nat). Furthermore, "total_art" denotes the total number of Art dataset and "total_nat" is the total number of Nat dataset.

3.2 Training Data Construction Methods with Various Few-shot Setups

Prompt	Art ACC	Nat ACC	Score
Fixed Few-shot 7B	92.83%	87.3%	88.81%
Fixed Few-shot 14B	93.19%	88.58%	89.84%
Dynamics Few-shot 7B	93.37%	89.8%	90.78%
Dynamics Few-shot 14B	95.34%	89.46%	91.07%

Table 4: Accuracy of different training data construction methods.

Two forms of prompts, namely fixed few-shot prompts and dynamic few-shot prompts, are evaluated respectively. As indicated in Table 4, when using training data constructed through the dynamic few-shot approach under the same parameter scale, the model's performance outperforms that of the fixed few-shot approach by 1.5% to 2%.

It should be noted that we have also experimented with other approaches in data construction, such as utilizing Qwen2.5 72B to generate inner thoughts to assist the model in understanding, and incorporating external data. Unfortunately, no further improvements are obtained and we will leave them for further exploration.

3.3 Different Fine-tuning Techniques

We conduct a comparative analysis between two fine-tuning approaches: full parameter fine-tuning and Low-Rank Adaptation (LoRA) tuning (Hu et al., 2021). As demonstrated in Table 5, our evaluation reveals that full parameter fine-tuning consistently achieves superior performance compared to the LoRA-based method. Notably, the 14B model with full parameter fine-tuning attains 1.18% higher accuracy than the 72B model using LoRA fine-tuning, despite the substantial difference in model size.

Fine-tuning Setups	Art ACC	Nat ACC	Score
Full Fine-tuning 7B	93.37%	89.8%	90.78%
Full Fine-tuning 14B	95.34%	89.46%	91.07%
LoRA Fine-tuning 32B	91.32%	87.32%	88.41%
LoRA Fine-tuning 72B	92.29%	88.99%	89.89%

Table 5: Accuracy of different fine-tuning methods.

3.4 Data Augmentation and Model Ensemble

Data augmentation and model ensemble are effective methods for improving model accuracy, and we have successfully increased the model accuracy to 93.41% using these methods.The detailed methods can be found in Appendix.

4 Conclusion

The FIE2025 evaluation in this paper demonstrates that LLMs can effectively perform Chinese factivity inference when enhanced with dynamic few-shot learning and full parameter fine-tuning. Our experiments reveal three key insights: (1) dynamic few-shot prompts outperform fixed prompts by introducing controlled variability; (2) full fine-tuning of 14B models surpasses LoRA-tuned 72B models despite size differences; (3) homogeneous model ensembles further improve performance by 2.3%, yielding a final score of 93.41% that finally ranks second in the shared task. While data augmentation for novel

predicates showed limited gains, these results establish that meticulous prompt engineering and complete parameter optimization are crucial for semantic tasks. The evaluation sets a strong baseline for future work in Chinese semantic understanding, particularly for applications requiring fine-grained truth of judgments.

References

Yulin Yuan. 2020. *The Factive Shift of "jide" ("Remember") and Its Conceptual Structural Basis*. *Language Teaching and Linguistic Studies*, number 1, pages 36–47.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. *Attention is all you need*. *Advances in neural information processing systems*, volume 30.

Zhenyu Chen and Yining Jiang. 2018. *Factivity and Fictivity: Towards the Opacity and Transparency of the Declarative World*. *Journal of Linguistic Research*, number 1, pages 15–37.

Davis Wertheimer, Luming Tang, and Bharath Hariharan. 2021. *Few-shot classification with feature map reconstruction networks*. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8012–8021.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and others. 2023. *Qwen technical report*. *arXiv preprint arXiv:2309.16609*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and others. 2022. *Training language models to follow instructions with human feedback*. In *Advances in neural information processing systems*, volume 35, pages 27730–27744.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. *Generalizing from a few examples: A survey on few-shot learning*. In *ACM computing surveys (csur)*, volume 53, number 3, pages 1–34. Published by ACM New York, NY, USA.

Philipp Weber, Christian Uhle, Meinard Müller, and Matthias Lang. 2024. *Real-Time Automatic Drum Transcription Using Dynamic Few-Shot Learning*. In *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, pages 1–8. Organized by IEEE.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. *Language Models are Unsupervised Multitask Learners*. In *OpenAI Technical Report*, volume 1, pages 1–24. Published by OpenAI, San Francisco, CA, USA.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. 2020. *Language Models are Few-Shot Learners*. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901. Published by Curran Associates, Inc., Red Hook, NY, USA.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv preprint arXiv:2106.09685.

A Appendix A. Data Augmentation

There are 14 predicate verbs in the test set that are absent from the training set. To tackle this issue, we generate 278 additional training samples based on these predicate verbs. However, the improvement is very marginal. The detailed evaluation results can be found in Table 6.

Fine-tuning Method	Art ACC	Nat ACC	Score
Full fine-tuning 14B	95.34%	89.46%	91.07%
Full fine-tuning with Data Augmentation 14B	95.52%	89.46%	91.12%

Table 6: Evaluation of data augmentation.

B Appendix B. Model Ensemble

Model ensemble is evaluated on the 14B model using enhanced data. Model ensemble generally achieves 2.3% additional improvement, resulting the final evaluation score of 93.41%. The result also shows that ensembles composed of models with identical architectures consistently achieve superior performance compared to heterogeneous architecture ensembles, indicating that maintaining architectural compatibility is crucial for effective ensemble voting. The detailed evaluation results can be found in Table 7.

voting method	Art acc	Nat acc	score
Ensemble with Different Structured Models	95.70%	90.68%	92.05%
Ensemble with Identical Structured Models	97.06%	92.03%	93.41%

Table 7: Evaluation of model ensemble.

