國立成功大學資訊工程學系
專題研究
Department of Computer Science and Information Engineering
College of Electrical Engineering and Computer Science
National Cheng Kung University
Capstone Project

小模型，大貢獻：準確且高效的中文新聞摘要模型訓練研究

Small Model, Big Impact: A Study on Training for Accurate and Efficient Chinese News Summarization Models

王郁豪
Yu-Hao Wang

指導教授: 陳響亮博士
Advisor: Shang-Liang Chen, Ph.D.

中華民國 114 年 6 月
June, 2025

Abstract

In the modern era of information overload, people often lack time to read full news articles and may be influenced by sensational headlines, leading to biased understanding. This thesis develops a compact Chinese news summarization model capable of running on resource-limited devices, such as smartphones, while maintaining high comprehension and summary quality. Combining knowledge distillation and curriculum training, we compress model parameters while enhancing generalization and language understanding. A five-stage curriculum is designed to gradually teach the model traditional Chinese conversion, essential aspect extraction, reasoning construction, and summary generation. Various data generation strategies and training approaches are evaluated, including learning rate adjustments, parameter freezing, and LoRA fine-tuning. Experimental results demonstrate that a 0.5B-parameter student model achieves comparable summary quality to 3B-parameter models, outperforming similar small-scale models and generating outputs with high fluency and content coverage. This work demonstrates the feasibility of lightweight yet high-quality Chinese news summarization models and provides a practical framework for mobile deployment.

Contents

1. Methodology

1.1 Data Collection

To construct the training corpus, a web crawler was implemented to collect news articles from *United Daily News*. In addition, the YeungNLP/firefly-pretrain-dataset was incorporated as supplementary pretraining data. This combination provided both domain-specific content and a broader linguistic foundation for model development.

1.2 Model Selection

Two models from the Qwen2.5 family were adopted. The student model was Qwen2.5-0.5B-Instruct, a lightweight 0.5-billion-parameter model intended for efficient training and deployment. The teacher model was Qwen2.5-32B-Instruct, which was used not only to generate synthetic training data but also to provide preprocessing guidance and evaluation signals during experimentation.

1.3 Curriculum Training

The training process was organized into five progressive stages (S1–S5) designed to gradually increase task complexity. First, traditional Chinese text was standardized using OpenCC (S1). Next, essential aspects were extracted from each article (S2), followed by the construction of reasoning triples that captured semantic relationships among these aspects (S3). Building on this foundation, the model then generated summaries based on the content together with the extracted aspects and triples (S4). Finally, the most challenging stage required the model to generate summaries directly from the raw news content (S5).

1.4 Data Generation Strategies

Different strategies were explored to generate training data. The **V1 strategy** relied on a single step to jointly produce aspects, triples, and summaries. **V2** followed a sequential approach in which aspects were first generated, then expanded into triples, and finally used to construct summaries. **V3** reversed the order by generating the summary first and subsequently deriving aspects and triples from it. **V4** extended V3 with manual corrections to reduce noise and improve training quality. The final dataset was divided into 23,848 training articles, 5,960 validation articles, and 6,052 test articles, as summarized in Table 1.1.

Table 1.1: Dataset Statistics

Dataset	Training	Validation	Testing
Number of articles	23848	5960	6052

1.5 Training Strategies

Several fine-tuning strategies were evaluated. The default configuration applied a decaying learning rate. A variant, denoted `lr_adj`, maintained a constant learning rate in Stage 5 to prevent premature convergence. Other experiments selectively froze model components, either keeping only attention layers (`only_attn`) or only MLP layers (`only_mlp`) trainable, to measure their relative contributions. LoRA-based adaptation was also tested, applying low-rank factorization to MLP (rank 160) and attention (rank 32) layers. Finally, combinations of these methods were explored to assess possible synergies.

1.6 Evaluation Metrics

Model outputs were evaluated using both automatic and human-aligned metrics. ROUGE-1, ROUGE-2, and ROUGE-L were applied to measure n-gram overlap with reference summaries, while BERTScore was used to capture semantic similarity. In addition, the teacher model (32B) was employed as a judge to assess naturalness and information coverage, providing a complementary qualitative evaluation beyond automatic metrics.

2. Experiments

2.1 Data Generation Strategy Comparison

The first set of experiments focused on evaluating different strategies for generating training data, with the goal of identifying which approach most effectively facilitates downstream summarization. Direct reasoning (V1), in which the teacher model was required to infer key elements and triples before producing the summary, consistently produced incomplete and incoherent outputs. Although some degree of logical consistency was maintained, the generated summaries often lacked sufficient coverage of important content, limiting their utility as high-quality training data.

An alternative stage-by-stage strategy (V2) was designed to mitigate these issues by splitting the generation process into multiple steps. Contrary to expectation, however, this approach worsened coherence and frequently led to "out-of-focus" summaries in which the generated key elements diverged substantially from the final summaries. This indicates that decomposing the generation into smaller stages can actually amplify inconsistencies, making it difficult for the student model to learn meaningful reasoning patterns.

In contrast, the summary-first strategy (V3) demonstrated clear advantages. By first producing a coherent summary and then deriving supporting elements and triples, this method yielded the highest performance across all automatic and human-aligned evaluation metrics, including ROUGE, BERTScore, and Judge scores. These results suggest that ensuring coherence at the summary level provides a stable foundation for subsequent reasoning tasks, which in turn improves both training efficiency and output quality. Table 2.1 presents the quantitative comparison of the three strategies.

Table 2.1: Performance Comparison by Data Generation Strategy

Model	R-1	B-F1	Judge	R-2	R-L
4stg_v3	**45.5**	**77.9**	**70.3**	**24.3**	**37.6**
4stg_v1	43.8	76.8	64.0	22.1	35.5
4stg_v2	37.6	69.4	65.1	17.5	23.4

2.2 Training Strategy Comparison

The second set of experiments explored alternative training strategies, with particular attention to learning rate schedules and parameter-freezing techniques. A key finding was that adopting a non-decaying learning rate significantly improved model performance. This suggests that, unlike larger-scale models where decay is often beneficial to avoid overfitting, smaller student models benefit from maintaining a steady learning signal throughout training. Another observation concerns the role of different parameter groups. Freezing only the attention layers or only the MLP layers still resulted in relatively strong performance, indicating that both components of the network contribute meaningfully to summarization ability. For example, the “only_mlp” variant retained a high BERTScore (78.6%), suggesting that attention-driven representations alone can sustain semantic fidelity.

By contrast, the application of LoRA, though successful in other low-resource fine-tuning scenarios, failed to provide improvements in this summarization setting. In fact, performance stagnated or slightly regressed. This result not only underscores the task-specific nature of parameter-efficient methods but also highlights a potential conflict between LoRA’s parameter expansion and the principle of maintaining a lightweight student model. The results of these comparisons are reported in Table 2.2.

Table 2.2: Performance Comparison by Training Strategy

Model	R-1	B-F1	Judge	R-2	R-L
4stg_v3-lr_adj	**48.4**	**79.3**	72.8	**25.7**	**40.1**
4stg_v3-lr_adj-only_mlp	46.6	78.6	71.5	24.2	38.4
4stg_v3-lr_adj_lora	45.6	78.0	**73.6**	23.3	37.4
4stg_v3	45.5	77.9	70.3	24.3	37.6
4stg_v3-lr_adj-only_attn	45.2	77.8	69.1	23.0	37.1

2.3 Curriculum Stage Comparison

The final set of experiments investigated the influence of curriculum learning, particularly the number of curriculum stages, on model performance. The findings reveal a nuanced picture. When trained without a decaying learning rate, models with a single stage occasionally achieved marginally higher ROUGE-1 scores compared to those trained with multiple stages. However, multi-stage setups (four or five stages) consistently yielded more accurate Traditional Chinese generation and greater stability in output quality. This suggests that while curriculum depth may not always maximize surface-level evaluation scores, it enhances the stylistic fidelity and robustness of the generated summaries. Tables 2.3–2.4 summarize these results, and Figure 2.1 illustrates the proportion of Traditional Chinese in outputs among different models, especially in terms of improvements.

Table 2.3: Performance of Different Stage Numbers with Decaying Learning Rate

Model	R-1	B-F1	Judge	R-2	R-L
Qwen2.5-0.5B_4stg_v3	45.5	**77.9**	**70.3**	**24.3**	37.6
Qwen2.5-0.5B_1stg_v3	**45.9**	77.9	68.8	23.7	**37.7**

Table 2.4: Performance of Different Stage Numbers without Decaying Learning Rate

Model	R-1	B-F1	Judge	R-2	R-L
Qwen2.5-0.5B_1stg_v4-lr_adj	**47.0**	**78.8**	**75.9**	**24.5**	**38.9**
Qwen2.5-0.5B_4stg_v4-lr_adj	46.8	78.6	75.2	24.2	38.6
Qwen2.5-0.5B_5stg_v4-lr_adj	46.5	78.6	75.1	23.9	38.3

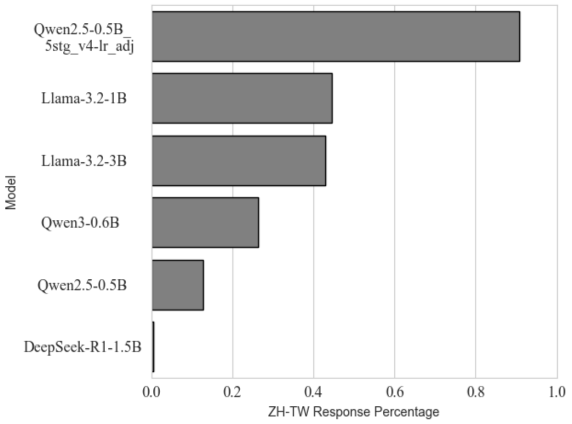

Figure 2.1: Proportion of Traditional Chinese responses generated by each model. The proposed 0.5B model with 5-stage curriculum achieves the highest ratio, surpassing even 3B models.

Furthermore, the benefits of curriculum learning appeared to depend heavily on the scale of pretraining. When large-scale pretraining was employed, curriculum learning only slightly improved summary quality. In contrast, with smaller pretraining datasets, curriculum learning provided substantial performance boosts, indicating that staged training helps compensate for weaker initial representations by guiding the model through progressively more complex tasks. As shown in Table 2.5, the 4-stage curriculum model (Custom-pretrained-4stg_v3-lr_adj) consistently outperforms the single-stage counterpart across most metrics, particularly in ROUGE-1 and Judge scores, highlighting the effectiveness of gradual task progression in low-resource pretraining scenarios.

Table 2.5: Performance of Custom Pretrained Models with Different Curriculum Stages

Model	R-1	B-F1	Judge	R-2	R-L
Custom-pretrained-4stg_v3-lr_adj	**13.3**	**50.6**	**0.9**	**1.8**	4.0
Custom-pretrained-1stg_v3-lr_adj	11.8	50.5	0.6	1.5	**5.0**

3. Results

3.1 Final Model Performance

The final 0.5B model achieves performance approaching that of larger 3B models in ROUGE-1 and BERTScore, demonstrating substantial improvements in summary generation quality. It significantly outperforms comparable small models, such as the Gemma series, with ROUGE-1 gains of approximately 5–7%, reaching a practically usable level. Compared with its pre-trained version before post-training and curriculum stages, the model’ s

Judge score increased by over 0.16, indicating that the training strategy effectively enhanced the naturalness and informativeness of generated summaries.

Table 3.1 shows a comparison of the final 0.5B model with other models of similar parameter scale and larger teacher models. The results suggest that careful curriculum training allows smaller models to approach the performance of significantly larger ones, closing the gap in both automatic evaluation metrics and human-judged quality.

Table 3.1: Performance Comparison of Final 0.5 Model with Other Models of Similar Parameter Scale

Model	R-1	B-F1	Judge	R-2	R-L
Qwen2.5-32B	**55.6**	**82.2**	81.7	**34.9**	**48.2**
Qwen2.5-3B	49.5	79.8	**83.8**	26.2	40.1
Qwen2.5-0.5B_5stg_v4-lr_adj	46.5	78.6	75.1	23.9	38.3
Llama-3.2-3B	43.3	76.8	74.3	20.7	34.7
Gemma-3-1B	41.8	76.0	72.6	18.3	31.2
Qwen2.5-0.5B	39.7	74.8	58.7	18.5	30.9
Gemma-2-2B	39.3	71.9	83.4	18.3	26.2
Llama-3.2-1B	38.7	74.1	63.4	16.9	29.5
DeepSeek-R1-1.5B	32.2	71.3	65.8	12.0	21.4

3.2 Traditional Chinese Generation

In addition to overall performance, the model demonstrates marked improvements in linguistic handling. After post-training, the proportion of Traditional Chinese output rises significantly, even exceeding that of the teacher model. The percentage of responses completely free of Simplified Chinese characters is also higher than in other models, highlighting the effectiveness of the training data and curriculum strategy in producing culturally and linguistically appropriate text.

3.3 Mitigation of Teacher Model Errors

Beyond linguistic improvements, the model also mitigates issues inherited from the teacher model. Quantitative analysis shows that abnormal or unexpected endings are fully corrected: among 30,296 samples, 2,871 instances contained irregular endings (e.g., "`\n hổ \n`") in the teacher model, whereas after post-training, none of these anomalies appeared. This demonstrates that the staged curriculum learning and high-quality training dataset effectively addressed imperfections present in the original teacher outputs.

4. Conclusion

This study systematically explored strategies for building lightweight Chinese summarization models, focusing on data generation, training configurations, and curriculum learning. The results provide both methodological insights and practical guidance for the development of efficient yet high-quality summarization systems.

In terms of **data generation**, the experiments reveal that direct reasoning and stage-by-stage decomposition tend to produce incomplete or unfocused outputs, limiting their usefulness as training data. By contrast, a **summary-first**

approach consistently delivers coherent and comprehensive summaries, establishing a stable foundation for downstream reasoning and enabling more effective student–teacher distillation.

Regarding **training strategies**, a **non-decaying learning rate combined with full parameter fine-tuning** outperforms methods such as LoRA and partial freezing, indicating that full fine-tuning remains the most effective choice. Moreover, contrary to expectations, LoRA and partial freezing provide little benefit in terms of training time or VRAM savings, further reinforcing the practicality of full parameter fine-tuning.

The investigation of **curriculum learning** further highlights its importance in low-resource conditions. While multi-stage curricula contribute modest gains when large-scale pretraining is available, they prove highly effective for smaller, custom-pretrained models, substantially improving summarization accuracy, stylistic fidelity, and robustness. This staged progression also enables the student model to mitigate errors inherited from the teacher, such as abnormal token sequences and inconsistent linguistic output.

Taken together, these strategies enabled the development of a **0.5B-parameter model** that approaches the performance of much larger 3B models, while outperforming comparable small-scale baselines. Notably, the model demonstrates strong handling of Traditional Chinese, surpassing even its teacher in linguistic appropriateness.

In conclusion, this work demonstrates that through **summary-first data generation, full fine-tuning with steady learning signals, and curriculum-based task progression**, it is possible to achieve a balance of **efficiency, robustness, and high-quality summarization** within a compact architecture. The proposed framework not only advances the practical deployment of lightweight summarization models, including mobile applications, but also provides a replicable pathway for future research on resource-efficient natural language processing.

Acknowledgement

I would like to express my sincere gratitude to my advisor, Professor Shang-Liang Chen, for his invaluable guidance and insightful advice throughout this research. I also thank my classmates for their thoughtful discussions and feedback, the senior lab members for their assistance in setting up the experimental environment, and an anonymous friend for providing high-end GPUs, which allowed me to perform tests and training without waiting in turn. Finally, I am deeply grateful to my family and friends for their unwavering support and en-

couragement throughout the course of my research.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. (2017a). Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

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

[3] Pengcheng Jiang, Cao Xiao, Zifeng Wang, Parminder Bhatia, Jimeng Sun, Jiawei Han. (2024). TriSum: Learning Summarization Ability from Large Language Models with Structured Rationale. arXiv preprint arXiv:2403.10351.

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

A Prompt Templates and Examples

This appendix provides the prompt templates and examples used in this study, including the complete prompt designs for different versions of essential aspect extraction, triple generation, summary generation, and model evaluation.

Prompt for generating V1 summary with essential aspects and triples

```
System prompt:
給定一份文章，完成以下任務：
(1) 提取新聞的關鍵要素，關鍵要素應為關鍵短句、名詞或事實。
(2) 對於每個關鍵要素，檢索詳細的三元組，格式為 [實體 1 | 關係 | 實體 2]，這些三元組用於構成摘要。
(3) 使用檢索到的三元組撰寫一份摘要。
核心要素、三元組和撰寫的摘要應該在同一份回應中，並以換行符分隔。所有三元組 [實體 1 | 關係 | 實體 2] 的長度必須為 3（以”|” 分隔）。
範例：
================ 範例 =================
提示：
新聞：
{新聞}

更新：
核心要素：
[關鍵要素 1]、[關鍵要素 2]、[關鍵要素 3]、...

三元組：
[實體 1_1 | 關係 _1 | 實體 1_2]
[實體 2_1 | 關係 _2 | 實體 2_2]
[實體 3_1 | 關係 _3 | 實體 3_2]
...

生成摘要：
{摘要}

=====================================================================================
User prompt:
新聞：
{news}
```

Prompt for generating V2 essential aspects

```
System prompt:
請根據以下新聞內容，提取新聞的關鍵要素，關鍵要素應為關鍵短句、名詞或事實，請用中文回答，並且不要使用任何標點符號。請將每個關鍵要素用 [] 與、分隔。例如：
關鍵要素：
[關鍵要素 1]、[關鍵要素 2]、[關鍵要素 3]
=====================================================================================
User prompt:
新聞：
{news}
```

Prompt for generating V2 triples

System prompt:
請根據以下新聞內容與關鍵要素，檢索詳細的三元組，格式為 [實體 1 | 關係 | 實體 2]，這些三元組用於構成摘要，請用中文回答，並且不要使用任何標點符號。所有三元組用 [] 與、分隔，且長度必須為 3。
例如：
三元組：
[實體 1_1 | 關係 _1 | 實體 1_2]、[實體 2_1 | 關係 _2 | 實體 2_2]、...

==

User prompt:
新聞：
{news}

關鍵要素：
{essential_aspects}

Prompt for generating V2 summary

System prompt:
請根據以下新聞內容與檢索到的關鍵要素以及三元組，為新聞生成一份摘要，請用繁體中文回答。
例如：
生成摘要：

==

User prompt:
新聞：
{news}

關鍵要素：
{essential_aspects}

三元組：
{triples}

Prompt for generating V3 summary

System prompt:
請根據以下新聞內容，為新聞生成一份 100 字內精簡的摘要，請用繁體中文回答。
例如：
生成摘要：

User prompt:
新聞：
{news}

Prompt for generating V3 essential aspects and triples

System prompt:
請根據以下新聞內容以及摘要，提取新聞的關鍵要素與三元組，關鍵要素應為關鍵短句、名詞或事實，三元組應為 [實體 1 | 關係 | 實體 2] 的格式，這些三元組用於構成摘要，請用繁體中文回答。請將每個關鍵要素與三元組用 [] 與、分隔。例如：
關鍵要素：
[關鍵要素 1]、[關鍵要素 2]、[關鍵要素 3]、...

三元組：
[實體 1_1 | 關係 _1 | 實體 1_2]、[實體 2_1 | 關係 _2 | 實體 2_2]、...

User prompt:
新聞：
{news}

摘要：
{summary}

Example news, essential aspects, triples, and summary (V3)

News:
母愛不分物種。動保組織接獲民眾通報發現草叢有一窩胖嘟嘟奶汪，沒想到牠們的母親為了照顧這些孩子把自己餓成皮包骨，還有嚴重營養不良跟脫水狀況，對比起小狗們都肥碩健康，狗媽媽更是令人心疼。

根據 The Dodo 報導，美國密蘇里州 (Missouri) 聖路易斯流浪動物救援組織 (Stray Rescue of St. Louis,SRSL) 日前接獲民眾通報，說草叢裡面發現一窩肥胖的奶汪，但卻找不到狗媽媽，希望他們可以派人來協助一下。動物救援組織工作人員湯姆森 (Natalie Thomson) 表示，當他們趕往民眾通報的現場，的確真的看到一窩被照顧得好好的奶汪，很像是被人飼養後遺棄在附近。

但令動保人員意外的一幕出現了，他們過沒多久在附近的草叢找到了狗媽媽，可是這隻渾身骨瘦如柴、幾乎可以用皮包骨形容的黃狗顯然非常營養不良，跟牠一窩肥壯幼崽形成強烈對比，這隻狗媽媽顯得有些害怕人類，但牠並沒有逃跑或是圖保護孩子，反而是將目光投射在小狗身上，希望眼前的人類不要傷害牠的孩子。

每一隻狗寶寶都相當健康可愛甚至還有點肥。(圖/取自 Stray Rescue of St. Louis 官網)

後來這一窩共 10 隻奶汪跟牠們孱弱的母親都被聖路易斯流浪動物救援組織帶回收容所，經過健康檢查後反而讓獸醫跟動保人員更難過了，因為這 10 隻小奶狗除了有點寄生蟲問題之外，沒有任何營養不良的狀況，甚至還有些過重，但狗媽媽卻嚴重營養不良還脫水，可以說為了照顧孩子鞠躬盡瘁。

目前這一窩小奶汪都受到良好的照顧，不日將可開放認養，而狗媽媽則因為身體虛弱還需要靜養一段時間才可以考慮出養。湯姆森說，這隻狗媽媽其實是很溫柔的，只是個性比較慢熟，所以需要有愛心跟耐心的飼主陪伴，牠就會慢慢敞開心房願意相信人類。

Summary:
美國密蘇裡州一窩小狗被發現肥胖健康，但牠們的母親卻因照顧孩子而嚴重營養不良、脫水。這 11 隻狗已被救出，小狗們即將開放認養，母親則需繼續靜養。救援組織尋找有耐心的飼主，以陪伴這位溫柔但慢熟的母親。

Essential aspects:
[美國密蘇裡州]、[一窩小狗]、[肥胖健康]、[母親嚴重營養不良]、[脫水]、[已被救出]、[小狗們即將開放認養]、[母親需繼續靜養]、[救援組織尋找有耐心的飼主]、[溫柔但慢熟的母親]

Triples：
[美國密蘇裡州 | 發現 | 一窩小狗], [小狗 | 肥胖健康 | 狗媽媽], [狗媽媽 | 照顧孩子 | 嚴重營養不良], [狗媽媽 | 照顧孩子 | 脫水], [一窩小狗 | 被 | 救出], [小狗們 | 即將 | 開放認養], [狗媽媽 | 需 | 繼續靜養], [救援組織 | 尋找 | 有耐心的飼主], [狗媽媽 | 是 | 溫柔但慢熟的]

Model scoring prompt

System prompt:
你是一位語言評估專家。你的任務是根據文章與標準摘要，評估模型生成的摘要品質。
請根據以下評分標準，從 0 到 20 為其打分：
0：格式不正確或無意義的文字。
1：完全無關，與文章毫不相干。
2：虛構內容，語意不明。
3：嚴重誤解，包含重大錯誤。
4：幾乎無法反映原文，非常不完整。
5：文法錯誤，缺乏連貫性與相關性。
6：內容不完整且部分離題。
7：遺漏關鍵要點，有輕微虛構。
8：摘要過於模糊，缺乏具體性。
9：簡潔，涵蓋大部分重點。
10：可理解但可能遺漏細節。
11：忠實但略有遺漏。
12：大致正確但稍顯冗餘。
13：準確、結構良好，但有輕微風格問題。
14：涵蓋完整、清晰，語氣尚可改進。
15：清楚、忠實且具風格。
16：簡潔優雅，涵蓋所有重點。
17：非常接近理想摘要，僅有些微瑕疵。
18：優秀的摘要，易讀且內容完整。
19：幾近完美，僅可做細微風格潤飾。
20：完美——清楚、忠實、完整且優雅。
請回傳”分數：”加一個整數分數（0–20），接著是一句簡短的理由（例如：「分數：17 ——非常接近理想摘要，僅有些微瑕疵」）。

==

User prompt:
文章：
{article}

標準摘要：
{ground_truth}

模型生成摘要：
{response}