PATeam at SemEval-2025 Task 9: LLM-Augmented Fusion for AI-Driven Food Safety Hazard Detection

Xue Wan **Fengping Su** **Ling Sun** **Yuyang Lin** **Pengfei Chen**
Ping An Life Insurance Company of China, Ltd.
wx18707735705@163.com fengpings@outlook.com
sunling583@163.com lyy476629663@gmail.com 1012673739@qq.com

Abstract

This paper introduces the approach we adopted for the SemEval-2025 "Food Hazard Detection" task, which aims to predict coarse-grained categories (such as "product category" and "hazard category") and fine-grained vectors (such as specific products like "ice cream" or hazards like "salmonella") from noisy, long-tailed text data. To address the issues of dirty data, as well as the severe long-tail distribution of text labels and length in the data, we proposed a pipeline system. This system combines data cleaning, LLM-based enhancement, label resampling, and ensemble learning to tackle data sparsity and label imbalance problems. The two subtasks have strong semantic relatedness. By integrating them into a unified multiturn dialogue framework, we fine-tuned five models using a bagging approach. Ultimately, we achieved notable results on both subtasks, with F1 scores of 80.17% (ranked 4th) for Subtask 1 (ST1) and 52.66% (ranked 3rd) for Subtask 2 (ST2).

1 Introduction

Food safety incidents pose significant risks to public health and economic stability, necessitating rapid detection and transparent decision-making systems. The SemEval 2025 Task on Food Hazard Detection addresses this challenge by evaluating systems that classify food incident reports from web resources into predefined categories and specific vectors for "product" and "hazard." This task focuses on English-language reports, aiming to automate the discovery of food-related risks from social media and news platforms, where timely and interpretable predictions are critical for mitigating economic and health impacts. The task requires dual subtasks: ST1 for predicting hazard and product categories (e.g., "meat, eggs, and dairy" or "pathogenic bacteria") and ST2 for identifying exact hazard and product entities (e.g., "Salmonella" or "ice cream"). With 1,142 unique products and 128 hazards distributed across imbalanced categories, the task demands robustness against long-tail distributions and noisy text, reflecting real-world complexities in food safety monitoring (Randl et al., 2025, 2024).

Our system integrates data augmentation, label resampling, and ensemble learning to tackle these challenges. Inspired by advances in NLP for low-resource scenarios, (Wei and Zou, 2019) proposed some traditional data augmentation methods: Synonym Replacement, Random Insertion, Random Swap, and Random Deletion. In addition to these, advanced strategies like metadata-aware data augmentation (Zhang et al., 2021) (e.g., substituting similar products from a food ontology) and prototypical networks (Snell et al., 2017) show promise but remain untested in multi-task food safety contexts. Against this backdrop, we employed large language models (LLMs) to generate synthetic summaries of raw incident reports. This approach not only enhanced the diversity of the dataset but also preserved its semantic integrity.

To address the severe class imbalance in our dataset, we employed a combination of oversampling techniques for minority classes (Chawla et al., 2002) and a bagging ensemble (Breiman, 1996) comprising five fine-tuned models with Low-Rank Adaptation (LoRA) (Hu et al., 2022). This approach effectively mitigated the imbalance and enhanced model performance. MTLN (Multi-dimensional Type-slot label interaction Network) (Wan et al., 2023) is a neural network-based MTL framework designed to handle multiple natural language processing tasks through a unified architecture. Compared with single-task learning, multi-task learning (MTL) demonstrates enhanced generalization capabilities by leveraging task correlations and complementarity, which has been theoretically validated. Given that ST1 and ST2 in our challenge are both focused on food hazard detection and are highly related, we integrated Large

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1912–1918
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

Language Models (LLMs) to combine ST1 and ST2 into a multi-turn dialogue framework. This framework enables the model to effectively utilize data from multiple tasks, leading to improved generalization and adaptability, while also mitigating issues such as underfitting or overfitting (Guo et al., 2018).

Our experiments yielded competitive results: 80.17% F1-score (4th rank) for ST1 and 52.66% F1-score (3rd rank) for ST2. Quantitative analysis demonstrated that, compared to single-model predictions, employing bagging voting with five models boosted performance by 1.09% for Subtask 1 and 3.14% for Subtask 2. This indicates the effectiveness of the bagging voting approach, especially in significantly enhancing the model's generalization ability when dealing with long-tailed label distributions. However, the system encountered difficulties in handling ambiguous hazard descriptions (for example, distinguishing between "listeria monocytogenes" and "listeria spp"). This reflects the limitations of fine-grained entity recognition observed in the food safety literature. Qualitative errors further highlighted the need for context-aware disambiguation, particularly for overlapping hazard categories such as "sulphur dioxide and sulphites" versus "sulphates/sulphites." These findings align with the broader challenges in interpretable AI for food risk assessment (Ribeiro et al., 2016), emphasizing the trade-off between model complexity and explainability.

This paper demonstrates the following contributions:

- Through text summarization based on prompt engineering, we enhanced the diversity of the data, which helps to improve the model's generalization ability.
- Given the correlation between the two subtasks, we constructed them into a multi-turn dialogue format, which improved the model's performance.
- On the validation set leaderboard, ST1 and ST2 achieved scores of 86.41% (ranked 1th) and 54.32% (ranked 4th), respectively. On the test set leaderboard, we were ranked 4th for ST1 and 3rd for ST2.

2 System Overview

As illustrated in Figure 1, our experimental workflow begins with a dataset sourced from web scraping, which contains noisy data and suffers from severe label imbalance as well as a pronounced long-tail distribution of text lengths. To mitigate these issues, we first applied regular expressions to clean the data by removing elements such as hyperlinks, HTML formatting, and email addresses. Following this, we utilized a large language model with prompt engineering to generate textual summaries of the cleaned data, aiming to augment our dataset and enhance its diversity. This was achieved by concatenating the original texts with their generated summaries to form an enriched dataset.

To address the label imbalance problem within this enhanced dataset, we implemented label resampling techniques. Subsequently, using a Bagging approach, we sampled five subsets of data with replacement. Considering the strong interrelation between the two subtasks (ST1: food hazard prediction; ST2: precise vector detection), we combined them into a unified framework through a multi-turn dialogue format. Specifically, we performed Supervised Fine-Tuning (SFT) using the LoRA method across these five subsets. This process resulted in the training of five distinct models, whose outputs were aggregated through voting to determine the final predictions, thereby achieving improved performance and robustness.

2.1 Supervised Fine-Tuning (SFT)

The main approach employed across all two subtasks was Supervised Fine-Tuning (SFT) (Ouyang et al., 2022). In this training phase, model parameters are optimized through a supervised learning objective designed to enhance predictive performance on annotated datasets.

The standard formulation of the SFT loss function can be expressed as:

$$\mathcal{L}_{SFT} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{cross-entropy}}(y_i, \hat{y}_i) \qquad (1)$$

where $\mathcal{L}_{\text{crossentropy}}$ is the cross-entropy loss between the true label y_i and the predicted label $\hat{y}_i$, and N is the number of training samples.

2.2 LoRA

LoRA was implemented for adjusting large pretrained models. This methodology deploys trainable low-rank decomposition matrices A and B to approximate parameter adjustments, thereby minimizing trainable parameters while preserving the

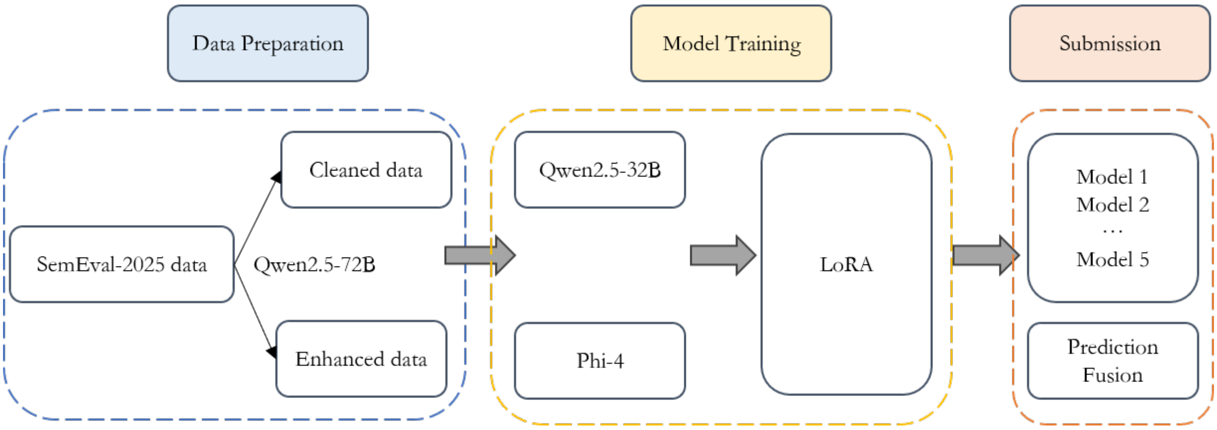


Figure 1: Task experimental progress

base model's capabilities. The adaptation process for a given weight matrix W operates through additive low-rank projections:

In addition to the techniques above, LoRA was implemented for adjusting large pre-trained models. This methodology deploys trainable low-rank decomposition matrices A and B to approximate parameter adjustments, thereby minimizing trainable parameters while preserving the base model's capabilities. The adaptation process for a given weight matrix W operates through additive low-rank projections:

$$W_{\text{new}} = W + \Delta W = W + AB^T \tag{2}$$

where A and B are low-rank matrices that are learned during fine-tuning. This approach allows the model to adapt to new tasks with fewer trainable parameters, making it computationally efficient.

The LoRA loss function is typically added to the standard SFT loss:

$$\mathcal{L}_{\text{LoRA}} = \mathcal{L}_{SFT} + \lambda \left\|A\right\|_F^2 + \lambda \left\|B\right\|_F^2 \tag{3}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and λ is a regularization parameter that controls the strength of the low-rank adaptation.

2.3 Data Preprocessing

The data provided for this task is sourced from web pages. Through data analysis, we identified the presence of unwanted elements such as hyperlinks, HTML formatting, and email addresses. To address this, we applied regular expression preprocessing to remove these components.

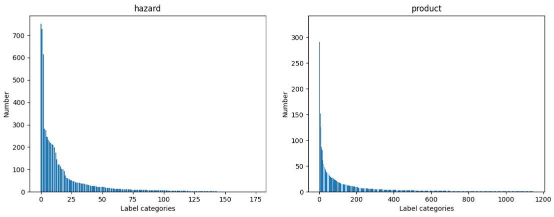


Figure 2: Distribution of labels for hazard and product in Subtask 2.

2.4 Data Augmentation

In our approach, we address the challenges posed by noisy data and imbalanced label distributions through robust data augmentation and resampling strategies. Initially, we generate concise summaries of cleaned texts using a large language model based on prompt-based summarization. These summaries are then concatenated with their original texts to form an augmented dataset. This enhancement strategy serves multiple purposes:

- Summarization aids in standardizing text representation by distilling essential information, thereby reducing noise and improving model interpretability.

- It boosts data diversity by introducing alternative phrasings and sentence structures, which helps mitigate the risk of overfitting.

- It accentuates crucial contextual elements, particularly beneficial for tackling label imbalance. By explicitly highlighting key information related to product categories and hazard classes, the model receives clearer classification signals.

Figure 2 illustrates the label distribution for hazard and product classes, revealing a severe long-tail distribution. To further mitigate this inherent issue, we employed resampling techniques on the augmented dataset. Specifically, for underrepresented categories (e.g., 74.2% of product categories have fewer than five samples), we performed resampling to ensure each category had at least five samples. Through multiple experimental validations, we found that augmenting small sample categories to a minimum of five entries yielded the best performance on the validation set. This balanced representation enabled our models to learn more robust and generalizable features, significantly improving overall performance.

2.5 Ensemble

After undergoing data preprocessing, data augmentation, and label resampling, the original dataset was transformed into a more balanced, diverse, and clean augmented dataset. Based on this enhanced dataset, we employed a bagging approach to perform bootstrapping, generating five subsets of data for training five Phi-4 models. For each model, we selected the weights that achieved the best performance on the validation set to evaluate the models on the official test set provided by the competition organizers. Finally, we integrated the predictions from all five models using an ensemble voting mechanism to produce our final submission for the test leaderboard.

2.6 Metrics

This task evaluates the joint marco-F1 for hazard and product. The composite evaluation metric is defined as:

$$\text{Composite-F1} = \frac{1}{2}\left(\text{F1}_\text{h} + \text{F1}_{\text{p}|\text{h}}\right) \tag{4}$$

Where the component metrics are calculated as:

$$\text{F1}_\text{h} = \frac{2 \cdot \text{Precision}_\text{h} \cdot \text{Recall}_\text{h}}{\text{Precision}_\text{h} + \text{Recall}_\text{h}} \tag{5}$$

$$\text{F1}_{\text{p}|\text{h}} = \frac{2 \cdot \text{Precision}_{\text{p}|C} \cdot \text{Recall}_{\text{p}|C}}{\text{Precision}_{\text{p}|C} + \text{Recall}_{\text{p}|C}} \tag{6}$$

The conditional set C is formally defined as:

$$C = \{\, i \mid \hat{y}_{\text{h},i} = y_{\text{h},i} \,\} \tag{7}$$

Where the subscripts h and p represent hazard and product, respectively.

3 Experimental Setup

This section describes various experiments conducted on model fine-tuning and inference, aimed at exploring the impact of different approaches on the model's F1-score. We applied the LoRA method to fine-tune the Phi-4 models, using a rank of 4, an alpha of 8, and targeting all layers. The Phi-4 model parameters were frozen, with only the low-rank adapter parameters being trained.

During the LoRA-based domain fine-tuning, we trained large models using the Llama-Factory (Zheng et al., 2024) framework and the Adam (Kingma and Ba, 2014) optimization algorithm. The training included a warm-up step of 10% and a learning rate of 5e-5. Additionally, we performed distributed training using Deepspeed Zero-3 (Rajbhandari et al., 2020) on two NVIDIA A100 GPUs (80GB), with a batch size of 1 per GPU and gradient accumulation of 12, for a total of 5 epochs.

In this study, we employed various techniques and used bagging to sample five subsets of data, training five distinct models. We then evaluated their performance on the SemEval-2025 official dataset by aggregating the predictions from the five LoRA-fine-tuned LLMs. The best checkpoints were selected based on the highest F1-score on the validation set, and a voting mechanism was used to make final predictions.

4 Results

4.1 Main Quantitative Findings

As can be seen from Table 1, our system performed robustly on both subtasks of the SemEval 2025 Food Hazard Detection Challenge. For ST1, which involves the prediction of food hazard categories, our model achieved an F1 score of 80.17%, ranking 4th among all participants. For ST2, which involves the prediction of the exact product and hazard vectors, our model achieved an F1 score of 52.66%, also ranking 3rd. These results highlight the effectiveness of our approach in tackling the challenges of class imbalance and long-tail distribution in the dataset.

Task name	**F1(%)**	**Rank**
ST1	80.17	4
ST2	52.66	3

Table 1: Results of Phi-4 on the test leaderboards.

4.2 Ablation Analysis

We conducted an ablation study to evaluate the contributions of various components in our system, as shown in the table 2. The experiments were conducted on the official test split.

Method	ST1 F1 (%)	ST2 F1(%)
Single-turn	61.52	35.43
multi-turn	63.97	37.52
multi-turn + Data Augmentation	79.08	49.52
multi-turn + Data Augmentation + Bagging	80.17	52.66

Table 2: Ablation results on Phi-4.

Single-turn: The model is tasked with completing both ST1 and ST2 predictions simultaneously. The baseline approach, which did not incorporate multi-turn dialogue or data augmentation, achieved an F1 score of 61.52% on ST1 and 35.43% on ST2.

Multi-turn: First, the model predicts the hazard-category and product-category in ST1. Then, based on these initial predictions, it proceeds to predict the results for ST2. Detailed prompt information can be found in Appendix A.1. By incorporating multi-turn dialogue, the system showed improvement with F1 scores of 63.97% (ST1) and 37.52% (ST2), demonstrating that the use of contextual dialogue helps in better capturing task-specific information.

Multi-turn + Data Augmentation: Adding data augmentation through text summarization further boosted the system's performance, with F1 reaching 79.08% (ST1) and F1 increasing to 49.52% (ST2). This indicates that data augmentation effectively enhanced model generalization by introducing more diverse training examples.

Multi-turn + Data Augmentation + Bagging: The final system, which included data augmentation and Bagging for model ensembling, showed the highest performance with F1 of 80.17% (ST1) and F1 of 52.66% (ST2). This demonstrates the benefits of combining multiple models to improve robustness and reduce variance in predictions.

These results underscore the effectiveness of our approach, where multi-turn dialogue, data augmentation, and ensemble learning via Bagging were key contributors to the performance improvements.

4.3 Error Analysis

To gain insights into the types of errors made by our system, we analyzed a sample of the predictions. While the system performed well overall, it tended to struggle with highly imbalanced classes, particularly in ST2 where the task requires predicting specific product and hazard vectors. In some cases, the model incorrectly predicted the exact hazard or product due to the complexity of distinguishing between similar categories in long-tail distributions. Further investigation and manual tagging of errors revealed that the most common mistakes were due to ambiguous or noisy text data, which is a challenge inherent in web-scraped datasets.

5 Conclusion

In this work, we proposed a multi-turn dialogue modeling approach combined with data cleaning, prompt-based data augmentation, label resampling, and a bagging strategy to tackle the SemEval 2025 food hazard detection challenge. Our final system achieved F1 scores of 80.17% and 52.66% on Subtask 1 and Subtask 2, respectively, securing 4th place in ST1 and 3rd place in ST2. From the ablation experiments, we observed that combining multi-turn modeling with data augmentation and ensemble methods can effectively mitigate the long-tail distribution and noise issues in real-world datasets. For future work, we plan to explore more advanced model interpretability techniques, domain-specific knowledge incorporation, and automated sampling strategies to further improve both the robustness and explainability of food hazard detection systems.

6 Limitations

Despite the promising results achieved by our multi-turn approach with data augmentation and bagging, several limitations remain. First, our reliance on large language models for text summarization and augmentation introduces potential biases in the generated data. Since these models are trained on broad corpora, they may inadvertently produce content that is contextually inconsistent or irrelevant for specific food hazard scenarios, thereby influencing both model training and evaluation outcomes.

Second, although label resampling and bagging helped address class imbalance, rare classes remain challenging. In real-world applications, novel or extremely infrequent hazards and products may not be adequately represented, leading to degraded performance when encountering such cases. Furthermore, the final system relies on multiple model ensembles and LoRA-based fine-tuning, which can be computationally expensive, making the approach less feasible for teams with limited resources.

Finally, while we integrated ST1 (hazard-category, product-category) and ST2 (hazard, product) within a multi-turn framework, our current interpretability methods are still somewhat simplistic. Generating "vector" explanations offers initial transparency, yet deeper domain-specific insights—such as causal chains or uncertainty estimates—are not thoroughly explored. Future work could incorporate more advanced explanation mechanisms to provide richer, more reliable interpretability in real-world food safety applications.

7 Acknowledgments

This task has been accomplished with the funding of *Ping An Life Insurance Company of China, Ltd*. All research presented in this paper was conducted during the Semeval-2025 competition. The opinions and conclusions expressed herein are solely those of the authors. We would also like to thank the task organizers and anonymous reviewers for their valuable feedback and constructive comments.

References

Leo Breiman. 1996. Bagging predictors. *Machine learning*, 24:123–140.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. *arXiv preprint arXiv:1805.11004*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

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

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Korbinian Randl, John Pavlopoulos, Aron Henriksson, and Tony Lindgren. 2024. CICLe: Conformal in-context learning for largescale multi-class food risk classification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7695–7715, Bangkok, Thailand. Association for Computational Linguistics.

Korbinian Randl, John Pavlopoulos, Aron Henriksson, Tony Lindgren, and Juli Bakagianni. 2025. SemEval-2025 task 9: The food hazard detection challenge. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Xue Wan, Wensheng Zhang, Mengxing Huang, Siling Feng, and Yuanyuan Wu. 2023. A unified approach to nested and non-nested slots for spoken language understanding. *Electronics*, 12(7):1748.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Conference on Empirical Methods in Natural Language Processing*.

Yu Zhang, Xiusi Chen, Yu Meng, and Jiawei Han. 2021. Hierarchical metadata-aware document categorization under weak supervision. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 770–778.

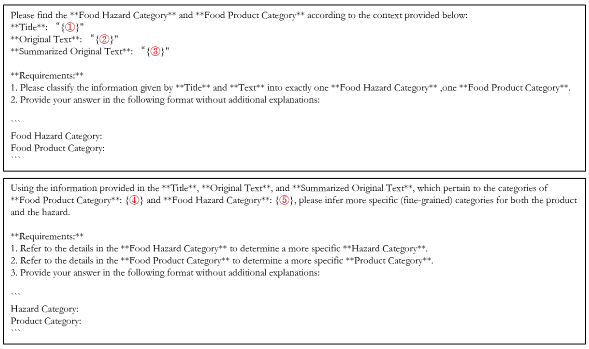
Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

A.1 Fine-tuning Prompt

In Figure 3, the upper part shows the prompt for the first dialogue round and the lower part shows the prompt for the second dialogue round. The red numbers indicate the required input information:① represents the input title;② represents the input text after cleaning it using regular expressions;③ represents the summary of the input text generated by LLM;④ represents the product-category predicted in the previous round;⑤ represents the hazard-category predicted in the previous round.

By constructing two-round dialogue fine-tuning prompts in this manner, the model can focus on both coarse-grained food and hazard categories, as well as the relationships between finer-grained food and hazard details. This approach enhances the model's performance by allowing it to better capture the nuances between different levels of categorization.



~~~~
Please find the **Food Hazard Category** and **Food Product Category** according to the context provided below:
**Title**:  "{①}"
**Original Text**:  "{②}"
**Summarized Original Text**:  "{③}"

**Requirements:**
1. Please classify the information given by **Title** and **Text** into exactly one **Food Hazard Category** ,one **Food Product Category**.
2. Provide your answer in the following format without additional explanations:

```
Food Hazard Category:
Food Product Category:
```
~~~~

~~~~
Using the information provided in the **Title**, **Original Text**, and **Summarized Original Text**, which pertain to the categories of
**Food Product Category**: {④} and **Food Hazard Category**: {⑤}, please infer more specific (fine-grained) categories for both the product
and the hazard.

**Requirements:**
1. Refer to the details in the **Food Hazard Category** to determine a more specific **Hazard Category**.
2. Refer to the details in the **Food Product Category** to determine a more specific **Product Category**.
3. Provide your answer in the following format without additional explanations:

```
Hazard Category:
Product Category:
```
~~~~

Figure 3: An Example of Prompt Engineering for Multi-turn Dialogue Based on LoRA Fine-tuning.