

Instructions for *ACL Proceedings

Anonymous ACL submission

Abstract

theoretical significance, but also will provide useful reference for the practical application in the field of food safety.

1 Introduction

With the rapid development of artificial intelligence technology, its wide application scenarios and far-reaching social impact are gradually emerging. Especially in the key field of food safety related to public health, with the continuous progress of big data and machine learning technology, challenges and opportunities coexist in the field of food safety. In this context, this study aims to explore the application potential of Large Language Models (LLMs) in the field of food safety, especially for the classification task of food hazard category and product category. Specifically, our basic task is to subdivide hazard category into 10 categories and product category into 22 categories, in order to achieve accurate classification and efficient management of food safety information through AI technology.

On this basis, this study selected a variety of large language models including BERT, RoBERTa, Qwen and ModernBERT as candidate models, and conducted experiments on the classification tasks of hazard category and product category. The final experimental results show that the ModernBERT model shows excellent performance in the validation set, with a score of 0.7952, which surpasses other models involved in the comparison. On the final test data set, the score of ModernBERT model is reached 0.7729, which not only verifies the application potential of large language model in the field of food safety, but also provides strong support for our current research.

By comparing and analyzing the performance of different models, we hope to find an efficient and accurate LLMs method to achieve accurate classification and efficient management of food safety information. This research not only has important

2 Related Work

With the progress and development of artificial intelligence era, AI technology has been applied in various fields. Leonieke carefully designed experiments for two systematic reviews on food safety (chemical hazards in grains and safety problems of green leafy vegetables), and comprehensively tested eight machine learning algorithms and all possible combination and integration models. The experimental results show that the integration model of Naive Bayes and Support Vector Machine shows the best performance. This model not only significantly reduced the number of documents to be reviewed by experts by 32.8% on average, but also greatly reduced the proportion of irrelevant documents, up to 57.8%, and ensured that 95% of relevant documents were retained. This achievement has brought substantial improvement and efficiency improvement to the literature screening work of food safety system review (van den Bulk et al., 2022). Sina has developed a risk assessment scheme based on multi criteria decision analysis (MCDA) for food safety problems, which is integrated into the AI database to realize the independent classification of food incident reports. Experiments show that this scheme is time-saving and efficient, and can quickly screen relevant reports, which verifies its practicability in AI environment (Röhre et al., 2024).

In recent years, the development of Large Language Model (LLM) technology has had an important impact on the academic and industrial as well as the entire AI technology community (Zhao et al., 2023). Hassani's research revealed that in the task of automatic classification of regulatory provisions driven by the concept of food safety, the

two language model families of BERT and GPT highlighted their extraordinary potential. In particular, the GPT-4o model, which has been deeply optimized, with an average accuracy of 89% and an average recall rate of 87%, strongly proves the excellent accuracy and extensive coverage of large language models (LLMS) in the task of classification of food safety regulations. Furthermore, when using a small number of samples learning strategy, the recall rate of GPT-4o was significantly increased to 97%. Although the accuracy was slightly reduced, this discovery not only revealed the delicate balance between model fine-tuning and a small number of samples learning, but also highlighted the good generalization performance of LLMS in dealing with regulations in different jurisdictions. In general, compared with the traditional baseline method based on long-term and short-term memory (LSTM) network and automatic keyword extraction, the large language models (LLMS) shows significant advantages in the task of automatic classification of food safety regulations, and opens up a new path for the automatic review of regulatory documents (Hassani et al., 2025). Randl’s research used machine learning and natural language processing technology to focus on the detection of food risk, and released a data set of 7546 short texts of food recall announcements. Through comparative analysis, the study reveals that in a specific category, the logistic regression model based on TF IDF features shows better performance than RoBERTa and XLM-R. In addition, the research also innovatively proposed a LLM in the loop framework based on conventional prediction, which not only enhanced the performance of the classifier, but also effectively reduced the energy consumption (Randl et al., 2024). Neris’ research met the challenge of the proliferation of food safety literature, using large-scale language model to automatically extract chemical hazard information without additional training or large-scale calculation. Through the test, it is found that the tip strategy of subdividing tasks has the best effect, with an accuracy rate of 93%. The chemical pollutants in a variety of monitoring are successfully identified, which proves the value of large-scale language model in literature information extraction (Özen et al., 2025). The review article co authored by Ma P et al. Elaborated the wide application and far-reaching impact of large-scale language models (LLMS) in the field of food science. This paper deeply analyzes the breakthrough applications of

LLMS in the core links of recipe innovation and development, accurate nutrition analysis, strict food safety supervision and efficient supply chain management. Through a series of application examples, this paper shows how LLMS can provide powerful decision support, accurate prediction and analysis, and excellent natural language processing ability for the field of food science, which greatly promotes the vigorous development of this field (Ma et al., 2024). Zhang Dan Proposed the ICL2FID framework, which uses the large language model technology to realize the hierarchical labeling of food poisoning events on social media with only a few examples. The framework improves the annotation accuracy through cross level information fusion, effectively coping with model illusion and circular dependence, and has lower cost and higher efficiency than traditional supervised learning and other LLM methods. Icl2fid is suitable for scenarios with limited resources but large data sets (Zhang et al., 2024).

3 Experiment Setup

3.1 Datasets

The Food Recall Incidents dataset contains 7546 short and refined texts, which are used as the titles of food recall announcements and are carefully crawled from 24 authoritative websites of the public food safety administration. These titles are written in six languages, of which English dominates with 6644 titles and German contains 888 titles. Most of these texts were written after 2010, describing in detail the recalled food products due to specific risks. Professional experts have carefully and manually divided these texts into four categories to accurately describe hazards and products, including 261 detailed hazard descriptions, 10 hazard categories, 1256 detailed product descriptions and 22 product categories (Zenodo). The data sample display is shown in Figure 1.

The data set has the problem of class imbalance. In the hazard-category classification task, the category with the largest number of samples is biological, which has reached 2018, while the category with the smallest number of samples is migration, which has only 13 samples (Figure 2). In the product-category classification task, this kind of class imbalance is also very obvious. For example, there are up to 1686 samples in the eat, egg and dairy products category, while there are only 5 samples in the sugar and syrups category (Figure

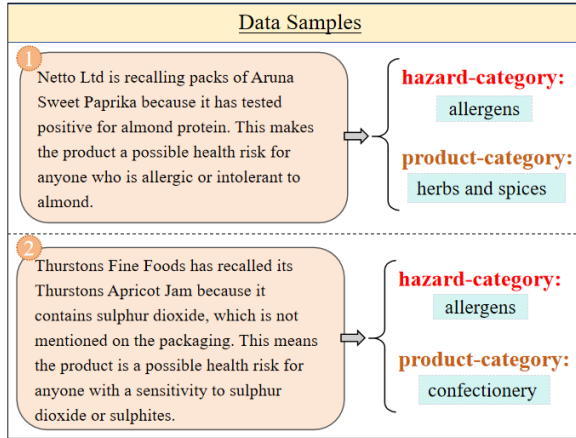


Figure 1: Data Samples

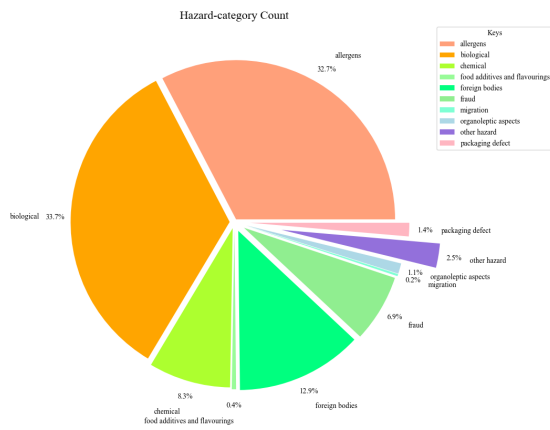


Figure 2: Hazard-category Statistics

3). This huge difference in the number of samples between categories will have an impact on model training.

3.2 LLMs

We chose BERT, RoBERTa, Qwen and ModernBERT were selected as candidate models.

Bidirectional encoder representations from transformers (BERT) model only needs an additional output layer to easily fine tune and build cutting-edge models for many tasks such as question answering and language reasoning. This process does not need to make cumbersome and large-scale adjustments to the architecture of specific tasks. BERT is not only simple and clear in concept, but also shows extraordinary strength in empirical research. It has successfully set a new record in 11 natural language processing tasks (Kenton and Toutanova, 2019).

Liu pointed out that BERT training is insufficient, and its performance can be improved by longer training, larger batches, more data, remov-

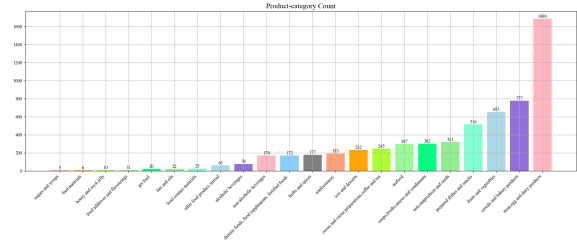


Figure 3: Product-category Statistics

ing the next sentence prediction target, training longer sequences and dynamic mask mode. The improved model RoBERTa achieved the best results in the benchmark tests of glue, race and squiad (Liu, 2019).

Qwen2.5 is Alibaba’s big language and multi-modal model, which is pre trained and fine tuned based on large-scale data. It provides a multi-scale language model with 18 trillion tokens of pre training data. It is good at command following, long text and structured data processing, and supports diversified prompts and multi languages, including 29 kinds of Chinese. (Qwin2.5)

Benjamin recommended the ModernBERT model as a modern optimization variant of the encoder only transformer model (such as BERT). ModernBERT stands out from many evaluation tests by virtue of its training on 2trillion token data and the length of native 8192 sequences, including diversified classification tasks and one-way and multi vector retrieval across different fields (covering codes), and has achieved top performance. It is particularly worth mentioning that ModernBERT is not only a leader in downstream application performance, but also in speed and memory efficiency. It is carefully designed to adapt to the reasoning requirements on the general GPU (Warner et al., 2024).

3.3 Methods

In this experiment (Figure 2), we used seven pre-trained models to classify products and hazard categories. First, we preprocess the data set and divide it into training set, validation set and testing set. The training set is used for model training, the validation set is used for model tuning, and the testing set is used for final evaluation of model performance. We loaded the pre-trained model and adjusted the output layer of the model according to the number of labels of product and hazard classification tasks. Next, we define the optimizer and set the super parameters such as learning rate.

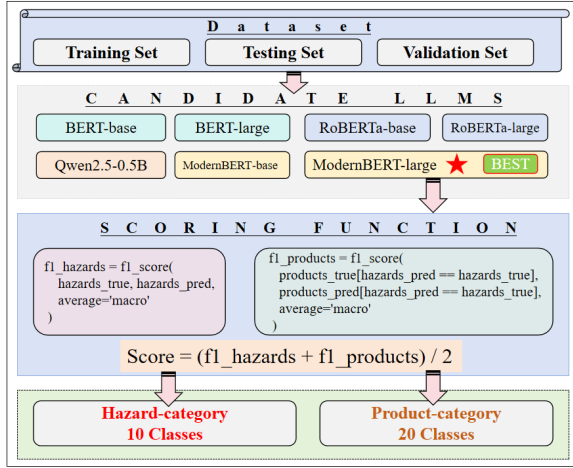


Figure 4: Model Selection

In the training process, we use the data loader to read data in batches and train the model through multiple iterations. In each iteration, we calculate the loss function and update the model parameters through back propagation. At the same time, we also recorded the loss value during the training process, and used the validation set to evaluate the macro F1 score of the model.

4 Results

We selected seven LLMs in the table to carry out the experiment (Table 1). The experimental results showed that the score of the BERT-base model is 0.7409, the BERT-large model is 0.7423, the RoBERTa-base model is 0.7778, the RoBERTa-large model is 0.7679, the Qwen2.5-0.5B model is 0.743, the ModernBERT-base model is 0.7915, the ModernBERT-large model is 0.7952. We can make it clear that the ModernBERT-large model is the best choice. The model showed excellent performance in the validation set, with a score of 0.7952, surpassing other models involved in the comparison, including different variants of BERT series and RoBERTa series. Although the score of ModernBERT-large model (0.7729) in the final test data set is slightly lower than its performance in the validation set, it is still enough to prove its strong generalization ability and dominant position in related tasks.

About the modernbert large model in the task of predicting food hazard category (task is divided into 10 categories), with the increase of training steps, the overall loss value shows a downward trend, which shows that the model is gradually learning the characteristics of the data, and the pre-

Model	Score
BERT-base	0.7409
BERT-large	0.7423
RoBERTa-base	0.7778
RoBERTa-large	0.7679
Qwen2.5-0.5B	0.743
ModernBERT-base	0.7915
ModernBERT-large	0.7952

Table 1: Model Score

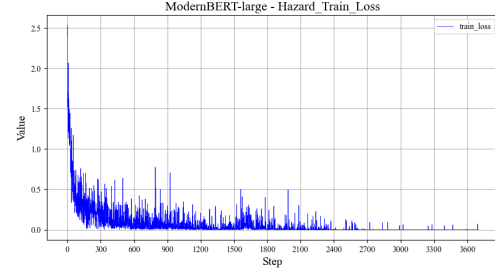


Figure 5: Harzard Train Loss

diction ability is constantly improving (Figure 3). In the task of predicting product category (tasks are divided into 22 categories), similarly, with the increase of training steps, the overall loss value also shows a downward trend. Because the task of product classification is more complex and the number of categories is more, the training loss may be higher than that of hazard prediction task, and the convergence speed may be relatively slow (Figure 4).

5 Conclusion

Acknowledgments

References

- Shabnam Hassani, Mehrdad Sabetzadeh, and Daniel Amyot. 2025. An empirical study on llm-based classification of requirements-related provisions in food-safety regulations. *arXiv preprint arXiv:2501.14683*.
- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacl-HLT*, volume 1. Minneapolis, Minnesota.
- Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.
- Peihua Ma, Shawn Tsai, Yiyang He, Xiaoxue Jia, Dongyang Zhen, Ning Yu, Qin Wang, Jaspreet KC Ahuja, and Cheng-I Wei. 2024. Large language models in food science: Innovations, applications, and

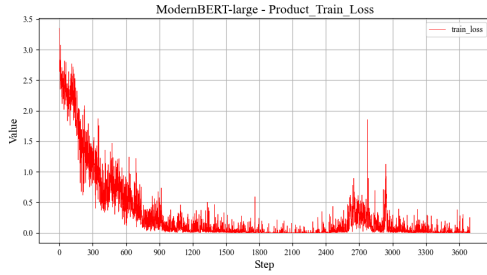


Figure 6: Product Train Loss

future. *Trends in Food Science & Technology*, page 104488.

Neris Özen, Wenjuan Mu, Esther D van Asselt, and Leonieke M van den Bulk. 2025. Extracting chemical food safety hazards from the scientific literature automatically using large language models. *Applied Food Research*, 5(1):100679.

Korbinian Randl, John Pavlopoulos, Aron Henriksson, and Tony Lindgren. 2024. Cicle: Conformal in-context learning for largescale multi-class food risk classification. *arXiv preprint arXiv:2403.11904*.

Sina Röhrs, Sascha Rohn, and Yvonne Pfeifer. 2024. Risk classification of food incidents using a risk evaluation matrix for use in artificial intelligence-supported risk identification. *Foods*, 13(22):3675.

Leonieke M van den Bulk, Yamine Bouzembrak, Anand Gavai, Ningjing Liu, Lukas J van den Heuvel, and Hans JP Marvin. 2022. Automatic classification of literature in systematic reviews on food safety using machine learning. *Current Research in Food Science*, 5:84–95.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Dongyu Zhang, Ruofan Hu, Dandan Tao, Hao Feng, and Elke Rundensteiner. 2024. Llm-based hierarchical label annotation for foodborne illness detection on social media. In *2024 IEEE International Conference on Big Data (BigData)*, pages 7272–7281. IEEE.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

343