PuerAI at SemEval-2025 Task 9: Research on Food Safety Data Classification Using ModernBERT

Jiaxu Dao, Zhuoying Li, Xiuzhong Tang,
Youbang Su, Qingsong Zhou, Weida He, Xiaoli Lan
School of Technology
Pu'er University
Contact:{daojiaxu, lizhuoying}@peu.edu.cn

Abstract

This paper presents our research in the SemEval-2025 Task 9: Food Hazard Detection Challenge, focusing on the application of ModernBERT for food safety data classification. Our system achieved 12th place in the official evaluation of subtask ST1, attaining a validation score of 0.7952 and a final test score of 0.7729. Through comparative experiments with various deep learning architectures, we demonstrate that ModernBERT exhibits superior performance in handling domain-specific semantics and long-tail distributions. These results validate the potential of ModernBERT for real-world food safety monitoring systems. The code is available at: https://github.com/daojiaxu/semeval_2025_Task-9.

1 Introduction

The rapid development of artificial intelligence (AI) has led to its widespread use across various sectors, with significant impacts on society (Ertel, 2024). In food safety, which directly affects public health, AI technologies such as big data analytics and machine learning offer innovative solutions to enhance food safety measures (Chhetri, 2024). Foodborne illnesses remain a global concern, and these advancements present new opportunities to address this issue. This study, part of SemEval 2025 Task 9: The Food Hazard Detection Challenge, explores the potential of pre-trained models for detecting and classifying food hazards (Randl et al., 2025).

The primary objective of this study is to classify food products into hazard and product categories based on safety-related attributes. We explore the application of pre-trained models to categorize food hazards into 10 types and products into 22 types. By leveraging AI, we aim to create a more efficient, accurate, and automated approach to managing food safety data.

We selected several state-of-the-art models, including BERT, RoBERTa, Qwen, and ModernBERT, as candidate models for the classification tasks. Our experiments indicate that ModernBERT consistently outperforms other models, demonstrating its effectiveness in food safety applications on both the validation and test sets.

By comparing the performance of these models, we seek to identify the most effective pre-trained models based method for managing food safety information. These findings not only contribute to advancing theoretical research but also provide practical insights for real-world food safety management, with the potential to enhance public health by improving food safety and preventing foodborne diseases.

2 Related Work

The advent of artificial intelligence has profoundly impacted various fields, particularly food safety research, with many scholars making significant contributions. Leonieke's systematic reviews evaluated multiple machine learning algorithms and combinations, with the hybrid Naive Bayes-Support Vector Machine (NB-SVM) model reducing expert workload and improving review accuracy (van den Bulk et al., 2022). Sina integrated multi-criteria decision analysis (MCDA) into an AI-driven database system for automated food incident report classification, verified through field tests (Röhrs et al., 2024).

With the rise of Large Language Models (LLMs) (Zhao et al., 2024), the research landscape has shifted. Zhao's 2024 survey emphasized LLMs' transformative potential across fields. Hassani demonstrated that BERT and GPT architectures excelled in regulatory text classification, with the optimized GPT-4o model outperforming traditional methods (Hassani et al., 2025). Randl introduced an LLM-in-the-loop framework, enhancing classifier performance while reducing energy consumption. Their analysis showed that logistic regres-

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 180–185

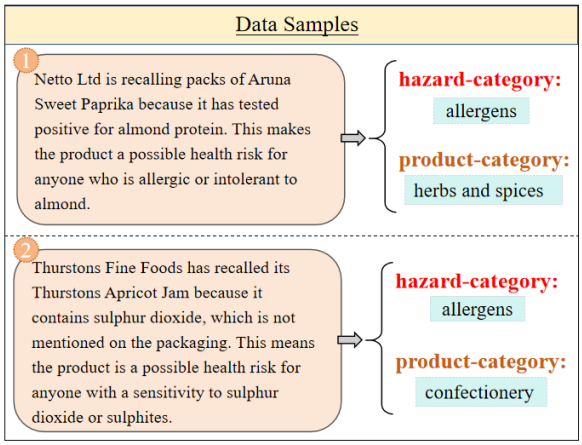

Figure 1: Data Samples

sion models with TF-IDF features outperformed advanced models in certain food recall categories (Randl et al., 2024).

Neris also contributed significantly. Neris used LLMs for zero-shot chemical hazard extraction, proving their effectiveness in environmental monitoring (Özen et al., 2025). Ma showcased LLM applications in decision support, driving progress in food science (Ma et al., 2024). Zhang Dan's ICL2FID framework improved annotation accuracy in food-poisoning event labeling on social media, offering cost-effective advantages over traditional methods (Zhang et al., 2024).

In conclusion, previous studies have explored food safety from various perspectives with different methods and models, laying a foundation for our research, which aims to further expand and deepen the field.

3 Experiment Setup

3.1 Dataset

The Food Recall Incidents dataset (Randl et al., 2025) contains 6644 short texts of English food recall notices annotated by two food science experts (character range 5-277, mean 88), sourced from official agencies such as the FDA.

The training dataset suffers from class imbalance. In the hazard-category classification task, the most frequent category is biological, with 2,018 samples, while the least frequent category is migration, with only 13 samples. A similar class imbalance is observed in the product-category classification task. For instance, the meat, egg and dairy products category has 1,686 samples, while the sugar and syrups category has just 5 samples. This significant discrepancy in sample distribution between categories can influence model training.

In comparison to the 5,433 samples in the training set, the validation set consists of only 565 samples. Within the validation set, the allergens category has the highest number of hazard-category samples, totaling 207. However, the migration hazard category has 0 samples. Regarding product-category classification, the meat, egg and dairy products category also has the largest number of samples, totaling 146. Meanwhile, some categories have very few samples, such as pet feed and feed materials, which have only 1 sample, and the sugar and syrups, honey and royal jelly, and food contact materials categories, which have 0 samples. Detailed data statistics can be found in Figure 2.

3.2 Pre-trained Models

In this study, we selected BERT, RoBERTa, Qwen, and ModernBERT as candidate models due to their proven effectiveness in natural language processing (NLP) tasks.

The Bidirectional Encoder Representations from Transformers (BERT) is renowned for its simplicity and efficiency, requiring only an extra output layer for fine-tuning, adaptable to a wide range of NLP tasks. Its key strength lies in handling diverse tasks without major architectural changes, making it efficient and flexible. BERT has set new benchmarks in NLP, achieving SOTA results in 11 benchmarks (Devlin et al., 2019). This performance has proven the value of pre-training deep bidirectional representations, a concept that BERT has popularized across the NLP community. BERT pre-trained model has set off a revolution in the field of natural language processing, and is gradually established as a new industry benchmark with its excellent accuracy in a number of automatic text processing tasks (Koroteev, 2021). BERT performs well in language comprehension tests, and experiments have shown that it can capture language structures, from low-level phrase information to rich linguistic levels in the middle, and then to combining information in a tree structure. It is particularly adept at handling long-distance dependency information.

Online public opinion helps reduce the impact of food safety, and experiments have shown that the BERT-BLSTM-CRF model has a higher accuracy in extracting entity relationships in the food safety public opinion dataset than other models by 3.29% to 23.25% (Zhang et al., 2022).

RoBERTa's enhancements make it an excellent candidate for tasks where language understanding

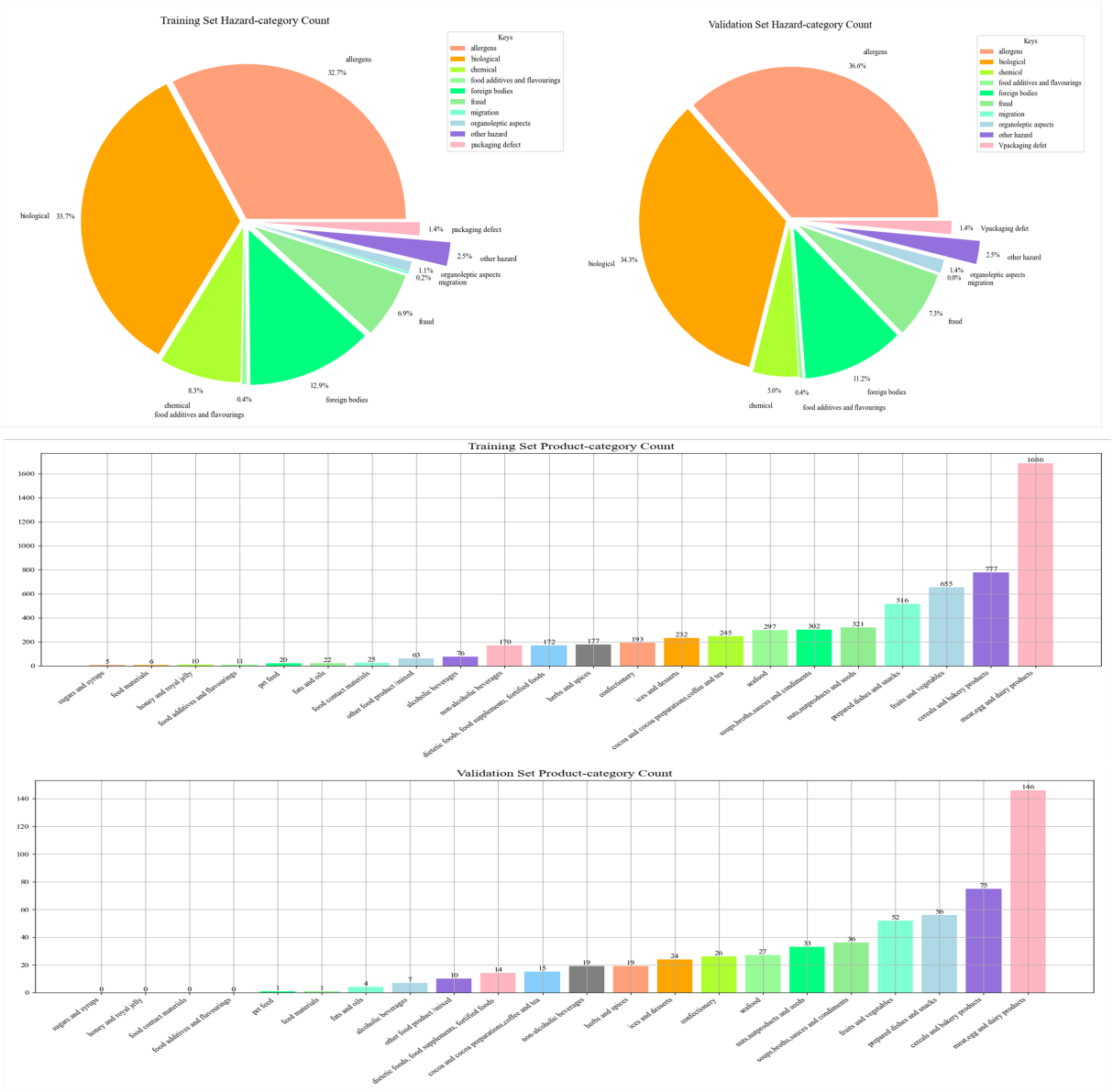

Figure 2: Count of Hazard and Product - Categories in Training Set and Validation Set

and fine-tuned performance are critical. In the research on the squiad V2 dataset, both Bert and Roberta showed strong text question answering ability. Bert stood out with its profound language understanding ability, while Roberta further improved its performance through optimized training strategies (Chopra et al., 2024). Liao proposed a multi-task sentiment analysis model based on RoBERTa, utilizing deep bi-Transformer for feature extraction and cross-attention for feature focus, outperforming other models experimentally (Liao et al., 2021). Briskilal proposed a predictive ensemble model based on BERT and RoBERTa for the classification of idioms and literal meanings. Tested on a newly created internal dataset, the model performed better than the baseline, with an accuracy improvement of 2% (Briskilal and Subalalitha, 2022).

Qwen 2.5, by Alibaba, is a pre-trained LM and multimodal model fine-tuned for tasks. With 18T tokens, it excels in commands, long texts, structured data. It is flexible in language and tasks, adaptable to various apps needing mixed data. (Yang et al., 2024).

ModernBERT, optimized by Benjamin, is an advanced encoder-only transformer, trained on 2 trillion tokens and handling up to 8192 tokens. It excels in diverse tasks, including retrieval and code-related applications. Its design ensures high speed, memory efficiency, and superior performance in downstream apps and real-time reasoning on GPUs (Warner et al., 2024). Given its advantages in

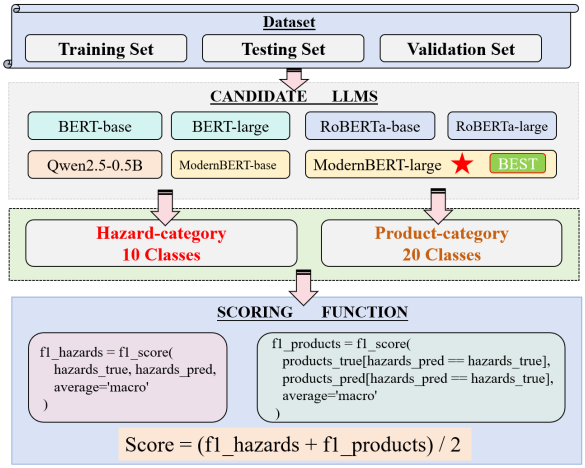

Figure 3: Model Selection and Evaluation Process

both performance and efficiency, ModernBERT has emerged as one of the leading models in NLP tasks. ModernBERT uses masked language model (MLM) for generative classification, showing excellent zero-shot learning ability (Clavié et al., 2025).

3.3 Methods

Experiment (Figure 3) used 7 pre-trained models to classify product and hazard categories. Data preprocessing was followed by splitting into train, validation and test sets for model training, tuning and evaluation, respectively.

Once the dataset is prepared, we load the pre-trained models and adjust their output layers according to the number of labels specific to the product and hazard classification tasks. The next step is to define the optimizer and set key hyperparameters, such as the learning rate. During the training process, we employ a data loader to read data in batches, enabling efficient model training over multiple iterations. In each iteration, we compute the loss function, and update the model's parameters using backpropagation.

Throughout the training process, we continuously monitor the loss values and use the validation set to evaluate the model's performance, specifically calculating the macro F1 score to assess its classification accuracy.

3.4 Evaluation Metric

The evaluation metric employs a conditional macro-averaged F1-score framework to align with the operational priorities of food safety detection. For hazard classification, the macro-F1 score is computed across all samples to ensure balanced evaluation of all hazard categories, regardless of their frequency in the dataset. This is mathematically defined as:

$$F1_{\text{hazards}} = \frac{1}{C_h} \sum_{c=1}^{C_h} \frac{2 \cdot P_c \cdot R_c}{P_c + R_c} \tag{1}$$

where C_h denotes the total number of hazard categories, P_c represents precision, and R_c denotes recall for class c.

The product classification evaluation is performed only on samples where hazard predictions match the ground truth, reflecting real-world constraints where incorrect hazard identification invalidates subsequent product categorization. The product macro-F1 score is calculated as:

$$F1_{\text{products}|H=H^*} = \begin{cases} \frac{1}{C_p} \sum_{k=1}^{C_p} F1_k \\ \quad \text{where } H^{(i)}_{\text{pred}} = H^{(i)}_{\text{true}}, \\ 0 \\ \quad \text{(invalid hazard prediction)} \end{cases} \tag{2}$$

where C_p represents the total product categories, and $F1_k$ is the F1-score for the k-th product class.

The final composite score is derived by averaging the two components:

$$\text{Score} = \frac{1}{2} \left(F1_{\text{hazards}} + F1_{\text{products}|H=H^*} \right) \tag{3}$$

4 Results

We selected seven pre-trained models from the table to conduct the experiment (Table 1). The experimental results indicate that the BERT-base model scored 0.7409, the BERT-large model scored 0.7423, the RoBERTa-base model scored 0.7778, the RoBERTa-large model scored 0.7679, the Qwen2.5-0.5B model scored 0.743, the ModernBERT-base model scored 0.7915, and the ModernBERT-large model scored 0.7952. It is clear that the ModernBERT-large model is the best choice. This model demonstrated excellent performance on the validation set, achieving a score of 0.7952, surpassing all other models, including different variants of both the BERT and RoBERTa series. Although the performance of the ModernBERT-large model on the final test set (0.7729) is slightly lower than its performance on the validation set, this still sufficiently demonstrates its strong generalization ability and its dominant position in related tasks.

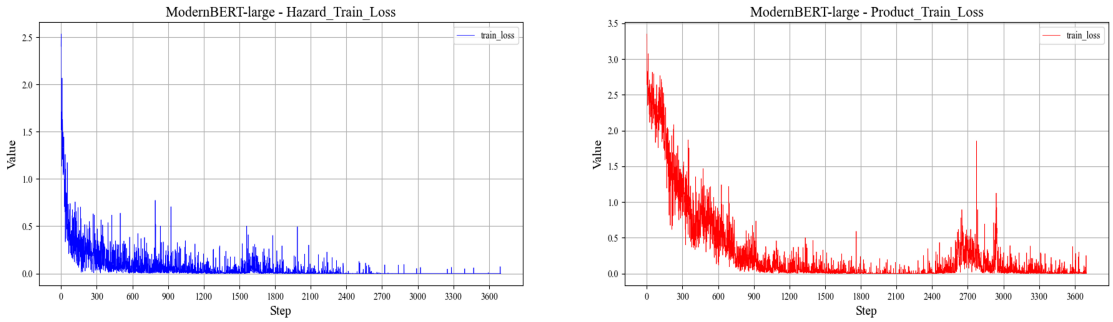

Figure 4: Loss Curves of ModernBERT-large on the Test Set (Food Hazard Classification and Product Classification)

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 Scores on the Validation Set

In the task of predicting food hazard categories (with the task divided into 10 categories), as the number of training steps increases, the overall loss value shows a downward trend, indicating that the model is gradually learning the characteristics of the data, and its prediction ability is continuously improving (Figure 4). Similarly, in the task of predicting product categories (with the task divided into 22 categories), the overall loss value also shows a downward trend with the increase in training steps. Since the product classification task is more complex with a larger number of categories, the training loss may be higher than that of the hazard prediction task, and the convergence speed may be relatively slower (Figure 4).

5 Conclusion

This study proposes a ModernBERT-based framework for food safety data classification in SemEval 2025 Task 9 Subtask ST1. Through systematic comparisons with pre-trained models including BERT, RoBERTa, and Qwen, we demonstrate that ModernBERT achieves superior performance in food hazard detection. Experimental results show that the framework obtains macro F1-scores of 0.7952 and 0.7729 on the validation and final test sets respectively, ranking 12th in the official evaluation of SemEval-2025 Task 9. This work establishes an effective technical pathway for applying language models to food safety management systems.

6 Limitations

While ModernBERT demonstrates superior performance in food safety classification tasks, this study has several limitations. First, the experimental data primarily focuses on structured text, leaving the model's generalizability to unstructured or diverse text sources insufficiently validated. Second, the model's efficiency in capturing semantic relationships within long textual sequences remains suboptimal, particularly when handling complex contextual dependencies. Additionally, classification performance on minority classes still requires improvement, necessitating further exploration of strategies to mitigate class imbalance effects. Finally, the current approach predominantly relies on an end-to-end supervised learning framework, with its adaptability to zero-shot or few-shot scenarios yet to be thoroughly assessed.

Acknowledgments

This work has been supported by the Special Basic Cooperative Research Programs of Yunnan Provincial Undergraduate Universities Association (grant NO. 202401BA070001-049), the 2024 Science and Technology Special Project of Pu'er University (grant NO. PYKJZX202401), and the Research Project on Education and Teaching Reform in Yunnan Province in 2023 (grant NO. JG2023217). The authors would like to thank the anonymous reviewers for their constructive comments.

References

J Briskilal and CN Subalalitha. 2022. An ensemble model for classifying idioms and literal texts using

bert and roberta. *Information Processing & Management*, 59(1):102756.

Krishna Bahadur Chhetri. 2024. Applications of artificial intelligence and machine learning in food quality control and safety assessment. *Food Engineering Reviews*, 16(1):1–21.

Sonali Chopra, Parul Agarwal, Jawed Ahmed, Siddhartha Sankar Biswas, and Ahmed J Obaid. 2024. Roberta and bert: Revolutionizing mental healthcare through natural language. *SN Computer Science*, 5(7):889.

Benjamin Clavié, Nathan Cooper, and Benjamin Warner. 2025. It's all in the [mask]: Simple instruction-tuning enables bert-like masked language models as generative classifiers. *arXiv preprint arXiv:2502.03793*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Zhe Dong, RuoQi Shao, YuLiang Chen, and JiaWei Chen. 2021. Named entity recognition in the food field based on bert and adversarial training. In *2021 33rd Chinese Control and Decision Conference (CCDC)*, pages 2219–2226. IEEE.

Wolfgang Ertel. 2024. *Introduction to artificial intelligence*. Springer Nature.

Shabnam Hassani, Mehrdad Sabetzadeh, and Daniel Amyot. 2025. An empirical study on llm-based classification of requirements-related provisions in food-safety regulations. *Empirical Software Engineering*, 30(3):72.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Mikhail V Koroteev. 2021. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.

Wenxiong Liao, Bi Zeng, Xiuwen Yin, and Pengfei Wei. 2021. An improved aspect-category sentiment analysis model for text sentiment analysis based on roberta. *Applied Intelligence*, 51:3522–3533.

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 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. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7695–7715.

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.

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*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *CoRR*.

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.

Qingchuan Zhang, Menghan Li, Wei Dong, Min Zuo, Siwei Wei, Shaoyi Song, and Dongmei Ai. 2022. [retracted] an entity relationship extraction model based on bert-blstm-crf for food safety domain. *Computational Intelligence and Neuroscience*, 2022(1):7773259.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.