TeleAI at SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection with Prompt Engineering and Data Augmentation

Shiquan Wang, Mengxiang Li, Shengxiong Peng, Ruiyu Fang, Zhongjiang He*, Shuangyong Song*, Yongxiang Li

Institute of Artificial Intelligence (TeleAI), China Telecom Corp Ltd

wangsq23, hezj, songshy@chinatelecom.cn

Abstract

This paper presents the approach we employed in SemEval-2025 Task 11: "Bridging the Gap in Text-Based Emotion Detection." The core objective of this shared task is emotion perception, focusing on determining the emotion the speaker is likely expressing when uttering a sentence or short text fragment, as perceived by the majority. In this task, we applied a prompt optimization strategy based on in-context learning, combined with data augmentation and ensemble voting techniques, to significantly enhance the model's performance. Through these optimizations, the model demonstrated improved accuracy and stability in emotion detection. Ultimately, in both Track A (Multi-label Emotion Detection) and Track B (Emotion Intensity Prediction), our approach achieved top-3 rankings across multiple languages, showcasing the effectiveness and cross-lingual adaptability of our method.

1 Introduction

Emotion recognition is one of the core tasks in the field of Natural Language Processing (NLP), aiming to identify and understand human emotional states from texts, dialogues, and other forms of data. With the rapid growth of data sources such as social media, online reviews, and customer feedback, sentiment analysis has become an indispensable tool across various industries, particularly in fields such as marketing, brand monitoring, public opinion analysis, and mental health(Saffar et al., 2023; Mohammad et al., 2018). Despite significant progress in sentiment classification and prediction tasks(Dadebayev et al., 2022; Zhang et al., 2024; Liu et al., 2024), the subjective and complex nature of emotions makes emotional expression more challenging due to factors such as individual differences, cultural background, and context. For instance, people may have vastly different emotional reactions to the same event, necessitating that sentiment recognition systems possess enhanced adaptability and flexibility to handle the complex and varied expressions of emotions across diverse contexts.

To address these challenges and bridge existing gaps, SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection introduces a large-scale emotion recognition dataset covering multiple languages(Muhammad et al., 2025a; Belay et al., 2025), aimed at advancing emotion detection technologies. This task consists of three sub-tasks: Track A, Multi-label Emotion Detection; Track B, Emotion Intensity; and Track C, Cross-lingual Emotion Detection(Muhammad et al., 2025b). It presents new challenges and opportunities for researchers in the field of emotion recognition, particularly in handling cross-lingual and multi-label sentiment tasks.

In this paper, we employed a prompt optimization strategy based on in-context learning, combined with data augmentation and ensemble voting techniques, to significantly enhance the model's performance. Specifically, we dynamically adjusted the prompt designs to help the model better understand and capture the subtle nuances of emotional expressions. The data augmentation techniques expanded the training set by generating synthetic data, particularly for categories with fewer emotion intensity samples, effectively addressing the data imbalance issue. Furthermore, the ensemble voting strategy, which combining predictions from multiple models, further improved the accuracy and stability of emotion detection.

During the testing phase, we selected the optimal model combination based on the results from the validation set for submission. Our approach achieved second place for Chinese in Track A, second place for Chinese, and third place for English in Track B.

*Corresponding authors.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 102–108

2 Relate Work

2.1 In-context Learning

In-context Learning (ICL) is an emerging machine learning paradigm that enables models to learn and infer without explicit training, by leveraging contextual information(Rubin et al., 2022; Dong et al., 2022). The core of ICL lies in the model's ability to dynamically adapt to the given context, analyzing examples or instructions within it to generate appropriate outputs(Giray, 2023; Marvin et al., 2023). This learning approach has shown great potential in the field of Natural Language Processing (NLP), especially in few-shot learning scenarios, where models can understand task patterns through a small number of examples(Li et al., 2024). The working principle of ICL can be broken down into two parts: the learning algorithm computes a task vector from the context, and then the task vector is used to modulate the model to generate outputs.

2.2 Prompt Engineering

Prompt Engineering refers to the process of designing and optimizing text prompts that are fed into large language models (LLMs)(Sahoo et al., 2024; Wang et al., 2024; He et al., 2024). By carefully crafting prompts with clear instructions, relevant context, specific examples, and accurate inputs, it guides LLMs to generate high-quality outputs that meet expectations. Prompt Engineering has a wide range of applications in text generation, data augmentation, and question-answering systems, significantly enhancing the performance and practicality of models across diverse application scenarios(Chen et al., 2024; Shao and Li, 2025).

2.3 Data Augmentation

Data Augmentation is the process of generating new training data to expand the dataset, thereby improving the generalization performance of models. In the field of natural language processing, traditional data augmentation methods often rely on techniques such as synonym substitution, sentence reconstruction, and context insertion(Hedderich et al., 2021; Feng et al., 2021; Liu et al., 2023). However, these methods are limited by the understanding of language, leading to lower-quality synthetic data. With the widespread use of large language models (LLMs), data augmentation techniques have undergone significant advancements. Leveraging the few-shot learning capabilities of LLMs, large amounts of synthetic data can be generated for low-resource tasks(Chintagunta et al., 2021; Møller et al.; Li, 2022), and utilizing the language understanding abilities of LLMs, vast amounts of unlabeled data can be annotated for cross-lingual tasks(Zhang et al., 2023; Meoni et al., 2023).

2.4 Supervised Fine-tuning

Supervised Fine-Tuning (SFT)(Wei et al.) is the process of further training a pre-trained model using a labeled dataset for a specific task. By guiding the model to make predictions and inferences based on labeled data, the model's weights are adjusted to match the data distribution of the specific task(Honovich et al., 2023). SFT can significantly improve the model's performance on particular tasks but requires high-quality labeled data and sufficient computational resources(Liu et al., 2022).

3 Methods

- **Profile**: You are an expert in sentiment analysis with extensive experience in identifying and categorizing emotions embedded in text.
- **Goals**: To accurately identify and classify emotions contained in the text. The candidate list of emotions is [anger, fear, joy, sadness, surprise].
- **Workflow**:
  1. Read and comprehend the given text.
  2. Detect the emotions present in the text; if no specified emotion is detected, output "no emotions".
  3. Return the prediction in the format provided in the examples.
- **Examples**:
  - Example 1: Input: "But not very happy." Output: joy,sadness
  - Example 2: Input: "Still had sex with her, though." Output:joy
  - Example 3: Input: "I still cannot explain this." Output: fear,surprise
- **Input**:
  [input_text]
- **Output**:

Figure 1: Prompt example for multi-label emotion detection

3.1 Track A: Multi-label Emotion Detection

In the multi-label emotion detection task, we propose a method that combines prompt design, data augmentation, and model fine-tuning with ensemble voting to enhance model performance.

Prompt Design: As shown in Figure 1, to guide the model in understanding the task and improving sentiment detection accuracy, we design diversified prompts and, based on In-context learning, provide rich example data within the prompts to help the model capture more contextual information. During the optimization process, we employ a dynamic prompt optimization procedure. Specifically, we

test various prompt designs, including variations in prompt construction, changes in the examples, and emphasis on specific emotions. These prompts are iteratively adjusted based on the model's feedback. For instance, if the model encounters difficulty in detecting subtle emotional nuances, we optimize the prompts by incorporating stronger emotional cues or context that helps clarify the sentiment. In selecting examples, we also compare the impact of different example selection methods on the final results. Through this iterative process, we ensure that the model receives the most effective prompts, thereby enhancing sentiment detection accuracy.

Data Augmentation: We first leveraged a large language model (LLM) to create synthetic data that aligns with the emotional characteristics of the original training set. The objective was to enhance data diversity while maintaining label consistency, ensuring no biased samples were introduced. Subsequently, we initialized a pre-trained model using the original training set and filtered the synthetic data based on the model's predictions. Only samples with labels matching the original dataset were retained, ensuring that the augmented data preserved accurate emotion classifications without introducing noise.

Model Fine-Tuning and Ensemble Voting: During the model fine-tuning phase, we further fine-tune the model using both the augmented data and the original training set. Finally, we employ an ensemble voting strategy to combine the predictions of multiple models, thereby achieving more stable and accurate sentiment classification results.

3.2 Track B: Emotion Intensity

In the emotion intensity task, we focus on a multi-class classification approach for each emotion. Our method involves predicting the intensity of a single emotion at a time, avoiding the interference of multiple emotions, and improving accuracy. We employ a carefully designed prompt system to guide the model's understanding and classification of emotion intensity, supplemented by data augmentation techniques to balance underrepresented categories. Finally, we employ the same ensemble voting strategy as in Track A to combine predictions from multiple models, further improving the stability and accuracy of the emotion intensity classification.

Prompt Design: To enhance the model's understanding of the task and improve the accuracy of sentiment intensity detection, we designed diverse prompts based on contextual learning. As shown in Figure 2, each prompt is designed to predict the intensity level of a specific perceived emotion in a given text. The intensity levels are categorized as follows: 0 (No emotion), 1 (Low intensity), 2 (Moderate intensity), and 3 (High intensity). The prompts were carefully structured to guide the model in identifying the intensity of a given emotion by considering both the content and context of the text. The model is instructed to first read and comprehend the input text, then focus on the specified emotion, and finally determine its intensity level.

We also ensure the diversity of examples included in the prompts by incorporating various sentence structures, vocabulary choices, and emotional expressions to represent different intensity levels of emotions. This provides the model with a diverse set of examples, enabling it to adapt to different emotional expressions and contexts. For instance, when the perceived emotion is anger, the examples range from mild irritation (level 1) to intense rage (level 3). Through this approach, the

- **Profile**: You are an expert in emotion intensity analysis with extensive experience in evaluating the strength of emotions in text.
- **Goals**: Your task is to predict the intensity level of a specific perceived emotion within the given text.
  - Intensity levels are classified as follows:
    - 0: No emotion
    - 1: Low degree of emotion
    - 2: Moderate degree of emotion
    - 3: High degree of emotion
- **Workflow**:
  1. Carefully read the input text to understand its content and context.
  2. Focus on the specified perceived emotion from the input.
  3. Determine the intensity level of the emotion based on the text.
  4. If the emotion is absent, assign an intensity level of 0.
  5. Return the prediction in the specified format.
- **Examples**:
  - Example 1:
    Input: Text: Colorado, middle of nowhere. | Perceived Emotion: anger
    Output: anger:0
  - Example 2:
    Input: Text: You know what happens when I get one of these stupid ideas in my head. | Perceived Emotion: anger
    Output: anger:1
  - Example 3:
    Input: Text: And then we have the ultimately retarded `` Spanish Lesson '' ( which I kind of like because it's so entertainingly bad ) and `` Incredible, '' which just flat-out gets on my nerves. | Perceived Emotion: anger
    Output: anger:2
  - Example 4:
    Input: Text: I got lie after lie. | Perceived Emotion: anger
    Output: anger:3
- **Input**:
  Text: **[input text]** | Perceived Emotion: anger
- **Output**:

Figure 2: Prompt example for emotion intensity

model learns the subtle distinctions between emotional intensities and becomes more proficient in predicting them accurately.

Task Formulation: We innovatively reformulated the task as a multi-class classification problem, where the model predicts the intensity level of a single emotion at a time. This approach ensures that the model focuses on one emotional intensity per prediction, minimizing potential interference from simultaneously processing multiple emotions. By simplifying the task in this manner, the model can concentrate on a single emotion and make more precise intensity assessments. For each input, the model determines the intensity of the specified emotion, categorizing it into one of four predefined intensity levels.

Data Augmentation: To address the challenge of data imbalance, particularly in cases where certain emotion intensity categories have fewer samples, we employed data augmentation techniques. Although we initially explored the use of a large language model (LLM) to generate synthetic data to expand the training set, the performance of the LLM-generated data on this task was relatively suboptimal. As a result, we adopted a more effective over-sampling strategy to supplement the underrepresented categories. This approach allowed the model to be exposed to a greater number of examples from the less-represented emotion intensity categories during training, thus improving the model's generalization ability and the accuracy of emotion intensity classification for these categories. By appropriately resampling the samples, we not only increased the number of instances in the underrepresented categories but also ensured the diversity and balance of the training set across different emotion intensity levels. This enhanced the model's robustness and accuracy in predicting emotion intensity, ensuring more reliable and stable performance across all intensity categories.

4 Experiment

In our experiments, we selected Qwen2.5-72B-Instruct(Yang et al., 2024) as the base model and fine-tuned it using LoRA methodology. The batch size was set to 32, the learning rate was set to 1.0e-4, and the model was trained for a total of 5 epochs.

4.1 Track A: Multi-label Emotion Detection

The experimental results on the Track A development set are shown in Table 1. The term "+Fine-tuning" refers to the fine-tuning of the base model using In-context Learning strategy, "+Data Augmentatio" indicates the incorporation of LLM-generated synthetic data during training to enhance data diversity, and "+Vote" denotes the use of an ensemble voting strategy during inference to combine predictions from multiple models. The experimental results demonstrate that the base model achieved a score of 0.6090 for English and 0.4826 for Chinese. After applying fine-tuning , the model's performance improved significantly, with scores of 0.8120 for English and 0.6892 for Chinese. Further, by introducing data augmentation , the scores for English and Chinese increased to 0.8164 and 0.6958, respectively, showing that the synthetic data generated by LLM notably enhanced the model's generalization ability. Finally, employing the ensemble voting strategy further improved the model's performance in both languages, with final scores of 0.8473 for English and 0.7412 for Chinese. We observed that fine-tuning and the ensemble voting strategy significantly improved the model's performance on the validation set. Additionally, we noticed that the performance across different emotion categories varied substantially across different step models, which could be attributed to the influence of the data quantity and label distribution in the validation set.

Method	English	Chinese
Base Model	0.6090	0.4826
+ Finetuning	0.8120	0.6892
+ Data Augmentation	0.8164	0.6958
+ Voted	0.8473	0.7412

Table 1: Our dev set results on the track a.(Only use Chinese and English data for solution exploration.)

Code	Language	Score	Rank
chn	Chinese	0.6817	2
eng	English	0.8064	4

Table 2: Our test set results on the track a. (Only the top 5 results are displayed.)

The experimental results on the Track A test set are shown in Table 2. Testing on both the Chinese and English datasets, our model demonstrated a certain level of performance in emotion detection. Specifically, the model achieved a score of 0.6817 on the Chinese dataset, ranking 2rd, indicating the

model's effectiveness in handling emotion detection for Chinese. For the English dataset, the score was 0.8064, ranking 4th. Although it did not place in the top three, the model still exhibited strong emotion detection capabilities.

4.2 Track B: Emotion Intensity

Method	English	Chinese
Base Model	0.6493	0.5290
+ Finetuning	0.8384	0.7668
+ Data Augmentation	0.8466	0.7704
+ Voted	0.8593	0.7833

Table 3: Our dev set results on the track b. (Only use Chinese and English data for solution exploration.)

The experimental results on the Track B development set are shown in Table 3. Compared to the results in Table 1, the "+Data Augmentation" here refers to the use of oversampling for data augmentation. The experimental results indicate that the base model achieved scores of 0.6493 for English and 0.5290 for Chinese. After fine-tuning the model with a carefully designed prompt and contextual learning strategy, the scores improved to 0.8384 for English and 0.7668 for Chinese. By applying the oversampling strategy to augment the training data, the scores increased to 0.8466 for English and 0.7704 for Chinese. Finally, using the ensemble voting strategy, the scores reached 0.8593 for English and 0.7833 for Chinese, achieving relative improvements of 32.34% and 48.07%, respectively, compared to the base model.

Code	Language	Score	Rank
chn	Chinese	0.7077	2
eng	English	0.8321	3
deu	German	0.7425	2
esp	Spanish	0.7861	4
ptbr	Portuguese	0.6896	2
ron	Romanian	0.7044	4
rus	Russian	0.9185	2

Table 4: Our test set results on the track b. (Only the top 5 results are displayed.)

At the final submission stage, we used the model that performed best on the validation set for prediction and ensemble voting. The experimental results are shown in Table 4. The model achieved a score of 0.7707 for the Chinese dataset, ranking 2nd, and a score of 0.8321 for the English dataset, ranking 3th. Due to time and resource constraints, for other languages, we only fine-tuned the model using carefully designed prompts, without applying data augmentation or ensemble voting strategies. Nevertheless, we still achieved top 5 rankings in five additional languages, further validating the effectiveness and generalizability of our approach. This demonstrates that, through carefully designed prompts and fine-tuning strategies, our method not only performs well in English and Chinese, but also adapts to other languages, showcasing strong cross-lingual generalization ability. In the future, with further investment in resources and optimization of strategies, the model's performance is expected to improve even further across more languages.

5 Conclusion

In this study, we have proposed an effective approach for emotion intensity prediction and multi-label emotion detection. By leveraging techniques such as carefully designed prompts, data augmentation through LLM-generated synthetic data, and dynamic optimization, we significantly improved model performance. The introduction of ensemble voting further stabilized and enhanced the model's classification accuracy. The experimental results on both Track A and Track B validate the effectiveness of our method, demonstrating its strong performance in both English and Chinese, and its generalizability to other languages. Future work could focus on extending the application to more languages, refining the model's ability to handle nuanced emotional expressions, and improving the scalability of the data augmentation strategies.

Limitations

While our approach achieved strong performance in English and Chinese, its effectiveness in other languages was limited due to time and resource constraints. These languages only underwent prompt fine-tuning without data augmentation or ensemble voting, leading to suboptimal results and highlighting the need for further optimization. Additionally, although LLM-generated synthetic data improved performance, its varying quality may have affected generalization. Future work should focus on refining data quality control and developing more robust language-specific strategies to enhance cross-lingual adaptability.

References

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3523–3540, Abu Dhabi, UAE. Association for Computational Linguistics.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2024. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *Preprint*, arXiv:2310.14735.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 66–76.

Didar Dadebayev, Wei Wei Goh, and Ee Xion Tan. 2022. Eeg-based emotion recognition: Review of commercial eeg devices and machine learning techniques. *Journal of King Saud University-Computer and Information Sciences*, 34(7):4385–4401.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988.

Louie Giray. 2023. Prompt engineering with chatgpt: a guide for academic writers. *Annals of biomedical engineering*, 51(12):2629–2633.

Zhongjiang He, Zihan Wang, Xinzhang Liu, Shixuan Liu, Yitong Yao, Yuyao Huang, Xuelong Li, Yongxiang Li, Zhonghao Che, Zhaoxi Zhang, et al. 2024. Telechat technical report. *arXiv preprint arXiv:2401.03804*.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Xiang Li, Yiqun Yao, Xin Jiang, Xuezhi Fang, Chao Wang, Xinzhang Liu, Zihan Wang, Yu Zhao, Xin Wang, Yuyao Huang, et al. 2024. Tele-flm technical report. *CoRR*.

Xuelong Li. 2022. Positive-incentive noise. *IEEE Transactions on Neural Networks and Learning Systems*.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Shixuan Liu, Chen Peng, Chao Wang, Xiangyan Chen, and Shuangyong Song. 2023. icsberts: Optimizing pre-trained language models in intelligent customer service. *Procedia Computer Science*, 222:127–136.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487–5496.

Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. 2023. Prompt engineering in large language models. In *International conference on data intelligence and cognitive informatics*, pages 387–402. Springer.

Simon Meoni, Eric De La Clergerie, and Théo Ryffel. 2023. Large language models as instructors: A study on multilingual clinical entity extraction. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 178–190.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

AG Møller, JA Dalsgaard, A Pera, and LM Aiello. Is a prompt and a few samples all you need? using gpt-4 for data augmentation in low-resource classification tasks. arxiv 2023. *arXiv preprint arXiv:2304.13861*.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya,

Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *Preprint*, arXiv:2502.11926.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671.

Alieh Hajizadeh Saffar, Tiffany Katharine Mann, and Bahadorreza Ofoghi. 2023. Textual emotion detection in health: Advances and applications. *Journal of Biomedical Informatics*, 137:104258.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *Preprint*, arXiv:2402.07927.

Jiawei Shao and Xuelong Li. 2025. Ai flow at the network edge. *IEEE Network*.

Zihan Wang, Yitong Yao, Li Mengxiang, Zhongjiang He, Chao Wang, Shuangyong Song, et al. 2024. Telechat: An open-source billingual large language model. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 10–20.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906.