# SCUNLP at ROCLING-2025 Shared Task: Systematic Guideline Refinement for Continuous Value Prediction with Outlier-Driven LLM Feedback

**Hong-Rui Pan, Jheng-Long Wu**
Department of Data Science
Soochow University
Taiwan, Taipei
Contact: rui0000525@gmail.com

## Abstract

Regression-based prediction is widely applied to continuous outputs, such as emotion dimension estimation. However, traditional methods struggle to handle unclear annotation standards and ambiguous cases. To address this challenge, we propose a dual-layer agent-executor framework, where the agent is responsible for constructing and refining guidelines, while the executor applies these guidelines to annotate large-scale data. Notably, we introduce a novel refinement mechanism that can detect outlier instances and provide feedback to the agent for guideline revision, thereby achieving iterative improvement. We applied this method to the ROCLING 2025 shared task (Lee et al., 2025) for predicting valence-arousal (VA) values in medical self-reflection texts. Compared to the unmodified version, the outlier-driven configuration effectively reduced MAE for both V/A, with A-MAE significantly decreased by 7.7%. The final valence-MAE was 0.51 and arousal-MAE was 0.87, ranking fourth.

***Keywords:*** LLM Prediction, Dimensional Sentiment Analysis, Prompt Optimization

## 1 Introduction

Dimensional emotion analysis have highlighted the importance of continuous valence-arousal (VA) prediction for understanding emotional states in text (Russell, 1980; Buechel & Hahn, 2017). These models have demonstrated remarkable capabilities in capturing the nuanced nature of human emotions across various domains, from social media analysis to clinical applications (Mohammad, 2018; Park et al., 2021; Mitsios et al., 2024). However, despite their impressive performance, current approaches

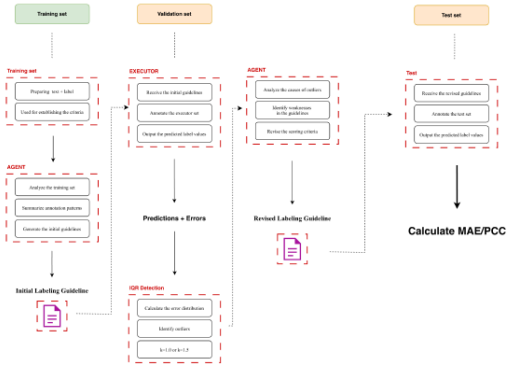


**Figure 1: Overview of the dual-layer Agent-Executor framework. The Agent constructs and revises guidelines, while the Executor performs annotations under three configurations, with IQR-based outlier detection driving the feedback loop.**

struggle with fundamental challenges in specialized domains such as medical self-reflection texts, where emotional expressions are often subtle, context-dependent, and require domain-specific understanding (Teodorescu et al., 2023; Alvarez-Gonzalez et al., 2021). This highlights the need for more adaptive and interpretable frameworks that can systematically improve annotation quality through iterative refinement.

Many current methods for emotion prediction rely on static annotation guidelines and traditional regression approaches that treat valence and arousal dimensions independently (Park et al., 2021; Bobicev & Sokolova, 2018). These approaches typically depend on large-scale manually annotated datasets with consistent labeling criteria. However, there are two main challenges with these methods. Firstly, creating high-quality annotations for dimensional emotion analysis is expensive and time-consuming, particularly in specialized domains like healthcare where expert knowledge is required (Wei et al., 2021; Giachelle et al., 2021). For instance, Wei et al. (2021) demonstrated that "manual annotation



*Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, pages 418–426
November 20-22, 2025 

by clinical experts is both time consuming and expensive," while Giachelle et al. (2021) noted that "manual annotation of large datasets is an expensive and time-consuming task requiring plenty of expert annotators with extensive experience in biomedical contents." Secondly, static guidelines cannot adapt to the diverse and ambiguous emotional expressions encountered in real-world texts, limiting their effectiveness for edge cases and domain-specific nuances (Alvarez-Gonzalez et al., 2021).

Large Language Models (LLMs), on the other hand, have shown remarkable ability to understand complex instructions and provide detailed feedback, indicating potential for more sophisticated annotation and refinement methods (Brown et al., 2020; Ouyang et al., 2022). Recent work in prompt optimization has demonstrated that iterative refinement can significantly improve model performance through systematic feedback incorporation (Wan et al., 2023). Madaan et al. (2023) showed that LLMs can iteratively improve their outputs through self-generated feedback, achieving approximately 20% improvement across various tasks without additional training.

Considering the advantages and disadvantages mentioned above, We propose a dual-layer Agent-Executor framework (Figure 1) for iterative guideline refinement, addressing annotation quality challenges in limited-data settings. In this approach, a high-capacity Agent formulates guidelines by analyzing domain knowledge and annotation complexities, while an efficient Executor applies these guidelines at scale to annotate data. A key feature is an outlier-driven feedback loop that decrease deviations in Executor predictions and feeds them back to the Agent for guideline revision. Evaluated on the ROCLING 2025 shared task for predicting valence-arousal values in Chinese medical self-reflection texts, our framework achieved fourth place. These findings highlight the effectiveness of adaptive guideline refinement and outlier feedback in enhancing annotation consistency and performance in specialized medical emotion prediction tasks.

The main contributions of this work are: (1) A novel dual-layer Agent-Executor framework that separates high-level guideline construction from efficient large-scale annotation; (2) An outlier-driven feedback mechanism that enables systematic identification and correction of problematic predictions; (3) Empirical validation on medical self-reflection texts showing the effectiveness of iterative refinement for dimensional emotion analysis.

## 2 Related Work

Recent advances in dimensional emotion analysis, prompt optimization, and agent-based NLP frameworks have converged to address the challenges of continuous emotion prediction in specialized domains (Buechel & Hahn, 2017; Madaan et al., 2023; Zhao et al., 2024). In this section, we review three key research areas that inform our approach: dimensional emotion analysis for valence-arousal prediction, iterative prompt refinement methods, and hierarchical agent frameworks for NLP tasks.

### 2.1 Dimensional Emotion Analysis

Dimensional models of emotion, particularly the valence-arousal framework, have emerged as robust representations for capturing the continuous nature of emotional states in text (Russell, 1980). Buechel and Hahn (2017) established EmoBank, a foundational corpus of 10,000 sentences annotated with Valence-Arousal-Dominance dimensions, demonstrating the superiority of reader's perspective over writer's perspective in terms of inter-annotator agreement. This bi-perspectival approach highlighted the inherent challenges in dimensional emotion annotation, where different viewpoints can lead to substantially different emotional interpretations.

Recent advances have addressed the gap between categorical and dimensional emotion representations. Park et al. (2021) presented a novel approach for predicting fine-grained VAD dimensions from categorical emotion annotations using Earth Mover's Distance loss, showing that traditional regression approaches treating dimensions independently suffer from significant limitations. Mitsios et al. (2024) further advanced the field by introducing ordinal classification techniques for two-dimensional emotion spaces, addressing perceptual similarities among emotional classes and achieving substantial improvements in prediction accuracy.

However, significant challenges persist in dimensional emotion analysis. Bagdon et al. (2024) noted that "humans perform worse when tasked to choose values from a rating scale," highlighting fundamental annotation reliability issues that affect

model training. These challenges are compounded in specialized domains such as medical texts, where emotional expressions are often subtle and context-dependent.

### 2.2 Prompt optimization

The limitations of static prompting approaches have driven significant research into adaptive and iterative prompt optimization methods. Ye et al. (2024) introduced the PE2 framework, which addresses static prompt limitations through meta-prompt components that enable iterative refinement and targeted prompt editing. This approach demonstrated the ability to rectify erroneous prompts and adapt to domain-specific requirements through systematic feedback incorporation.

Self-adaptive prompting has emerged as a key paradigm for dynamic prompt optimization. Wan et al. (2023a) proposed Consistency-based Self-adaptive Prompting (COSP), which dynamically selects examples based on consistency measures, achieving 15% improvement over static baselines. Their subsequent work (Wan et al., 2023b) extended this approach to Universal Self-Adaptive Prompting, automatically selecting suitable queries and responses as pseudo-demonstrations across diverse task types.

Madaan et al. (2023) introduced Self-Refine, demonstrating that large language models can iteratively improve their outputs through self-generated feedback without additional training. Their approach achieved approximately 20% absolute improvement on average across various tasks, establishing the viability of iterative refinement for quality enhancement. This work is particularly relevant to our outlier-driven approach, as it shows how models can identify and correct problematic aspects of their outputs through systematic feedback loops.

### 2.3 Hierarchical Frameworks

Hierarchical and multi-agent frameworks have shown substantial promise for complex NLP tasks requiring coordinated reasoning and execution. Zhao et al. (2024) presented EPO, a hierarchical LLM agent framework with separate components for subgoal prediction and action generation, achieving first place on the ALFRED leaderboard through effective dual-layer architecture design. This work demonstrates the power of role specialization in agent frameworks.

Wang et al. (2024) explored executable code actions in agent frameworks, showing that dual-component architectures with structured agent-executor separation can achieve 20% higher success rates than monolithic approaches. Their work highlights the importance of clear separation between high-level reasoning and low-level execution components.

Recent work has also addressed the specific challenges of outlier detection and iterative improvement in NLP systems. Zhang et al. (2024) further advanced this area by decomposing LLM confidence into uncertainty and fidelity components, providing the foundation for systematic identification of problematic examples. Hu et al. (2024) demonstrated Self-Refinement Tuning using model-generated feedback for iterative improvement, showing how outlier-driven learning can enhance model performance through systematic identification and correction of problematic outputs.

Our work builds upon these foundations by combining dimensional emotion analysis challenges with iterative prompt refinement techniques within a specialized agent-executor framework, specifically designed for prediction tasks where both accuracy and interpretability are crucial.

## 3 Method

This section presents our dual-layer Agent-Executor framework for iterative guideline refinement in valence-arousal prediction. We first introduce the overall framework architecture, then detail the Agent and Executor components, followed by our outlier-driven feedback mechanism for systematic guideline improvement.

### 3.1 Guideline Formulation

The Agent component serves as the rule-making authority responsible for understanding the complexities of dimensional emotion analysis and constructing comprehensive annotation guidelines. The Agent's primary functions encompass theoretical knowledge synthesis, systematic guideline construction, and iterative refinement based on feedback analysis.

**Theoretical Foundation Integration**: The Agent synthesizes established theoretical frameworks from dimensional emotion literature with empirical patterns observed in annotated text data. It

incorporates understanding of the circumplex model of affect while adapting to the nuanced emotional expressions characteristic of specialized text domains.

**Initial Guideline Construction Process** The Agent employs an open-ended instruction framework designed to enable flexible and comprehensive guideline development. Rather than imposing rigid structural constraints that might limit the model's reasoning capabilities, the initial prompt template encourages creative and thorough guideline formulation, the initial prompt template as shown in Figure 2.

> I will give you a text and its corresponding VA values. Based on this information, please draft a set of guidelines for scoring VA. The guidelines must include separate scoring rubrics (tables) for V and for A. I will use this prompt as the standard for prediction, so please write it carefully.

**Figure 2: The template for initial annotation guideline**

This unconstrained prompt design allows the Agent to autonomously determine the most appropriate organizational structure and content depth for the annotation guidelines. We found that We found that even without specific formatting requirements, the enables the agent to leverage framework its inherent reasoning capabilities to identify key evaluation dimensions, establish a logical hierarchy, and establish multifaceted criteria. These criteria emerge naturally from the data patterns instead of being restricted by prescriptive predefined rules.

### 3.2 Annotation Execution

The Executor component is responsible for applying the Agent's guidelines to perform large-scale valence-arousal prediction with rigorous quality control and standardized output formatting.

Once receiving the comprehensive annotation guidelines from the Agent, the executor systematically applies them to the unlabeled text dataset. To ensure high-quality annotations, the component implements strict adherence protocols that require explicit reference to guideline criteria during the prediction process. Each input text undergoes systematic evaluation against the established rubrics, with the Executor required to demonstrate clear reasoning chains linking textual features to scoring decisions.

### 3.3 Outlier-Driven Feedback Mechanism

The core innovation of our framework lies in the systematic identification and utilization of outlier predictions for guideline improvement. Following initial template-based prediction on the validation set, we employ the Interquartile Range (IQR) method to identify outlier instances based on prediction errors relative to ground truth values.

**IQR-Based Outlier Identification** We utilize the IQR method for outlier. Unlike methods that rely on standard deviation, IQR provides several key advantages: (1) Distributional Robustness - IQR remains stable even when error distributions are non-normal or contain extreme outliers, making it particularly suitable for emotion prediction tasks where errors may not follow Gaussian distributions; (2) Percentile-Based Thresholds - By defining outliers as values beyond Q1 - k×IQR or Q3 + k×IQR, the method provides interpretable thresholds that correspond to natural data quartiles; (3) Insensitivity to Outliers - Since IQR is calculated using only the 25th and 75th percentiles, it is not influenced by extreme values, preventing the masking effect where true outliers make other outliers appear normal.

For each dimension (valence and arousal), we calculate the absolute prediction error as:

$$Error_i = |y_{pred,i} - y_{true,i}| \tag{1}$$

where $y_{pred,i}$ and $y_{true,i}$ represent the predicted and ground truth values for instance $i$, respectively. The IQR is then computed as:

$$IQR = Q_3 - Q_1 \tag{2}$$

where $Q_1$ and $Q_3$ are the first and third quartiles of the error distribution. Outliers are identified as instances where:

$$Error_i \in (Q_3 + k \times IQR, +\infty) \tag{3}$$

where $k$ is the threshold multiplier. Given that prediction errors are non-negative (absolute values), we primarily focus on the upper bound criterion.

## 4 Experiment

### 4.1 Dataset

Our experimental evaluation employs datasets provided by the ROCLING 2025 shared task organizers (Lee et al., 2025).

We utilize The Chinese EmoBank (Lee et al., 2022) for initial guideline construction. Due to our focus on sentence-level emotion recognition and hardware constraints, we extracted 600 instances from the sentence-level (CVAS) and text-level (CVAT) components through stratified sampling to maintain distributional consistency with the complete dataset.

The validation set contains 994 doctors' self-reflection texts for system development, while the test set provides 1,541 doctors' self-reflection texts for final performance evaluation.

### 4.2 Outlier Detection Configurations

To evaluate the effectiveness of our outlier-driven feedback mechanism, we implement three experimental configurations that systematically assess the impact of iterative guideline refinement:

- **Baseline:** Employing the original prompt template to produce static guidelines that remain unchanged throughout the evaluation process, providing a reference point for measuring improvement.
- **Conservative Detection:** Implementing outlier-driven feedback mechanism with a conservative threshold setting ($k$ = 1.5). Outliers are identified when prediction errors exceed threshold, focusing on the most significant prediction failures to drive targeted guideline improvements while maintaining stability in the refinement process.
- **Aggressive Detection:** This configuration employs a more aggressive threshold setting ($k$ = 1.0) to identify a broader range of prediction anomalies. By lowering the outlier detection threshold to $Q_3 + 1.0 \times IQR$, this approach captures more instances for feedback analysis, enabling comprehensive guideline refinement at the potential cost of including less critical prediction errors.

### 4.3 Evaluation Metrics

Following the official evaluation protocol established by the ROCLING 2025 shared task organizers, we employ two primary metrics for assessing dimensional emotion prediction performance:

**Mean Absolute Error (MAE):** The MAE measures the average magnitude of prediction errors without considering their direction:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_{pred,i} - y_{true,i}| \quad (4)$$

where $n$ represents the total number of instances. MAE provides several advantages for dimensional emotion analysis: (1) Interpretability - MAE values directly correspond to the average prediction error in the original scale, making results easily interpretable; (2) Robustness to Outliers - Unlike squared error metrics, MAE is less sensitive to extreme prediction errors, providing a more stable assessment of typical model performance; (3) Linear Penalty - MAE assigns equal weight to all errors regardless of magnitude, avoiding the disproportionate influence of large errors that can skew evaluation in squared metrics.

**Pearson Correlation Coefficient (PCC):** The PCC measures the linear correlation between predicted and true values:

$$PCC = \frac{\sum_{i=1}^{n}(y_{pred,i} - \bar{y}_{pred})(y_{true,i} - \bar{y}_{true})}{\sqrt{\sum_{i=1}^{n}(y_{pred,i} - \bar{y}_{pred})^2 \sum_{i=1}^{n}(y_{true,i} - \bar{y}_{true})^2}} \quad (5)$$

where $\bar{y}$ represents the mean values. PCC offers complementary evaluation benefits: (1) Scale Invariance - PCC is unaffected by linear transformations, focusing on the relationship structure rather than absolute values; (2) Ranking Preservation - High PCC values indicate that the model maintains the relative ordering of emotional intensities across instances; (3) Distributional Alignment - PCC captures how well the predicted distribution matches the true distribution pattern, essential for dimensional emotion modeling.

The combination of MAE and PCC provides comprehensive evaluation coverage, with MAE assessing absolute prediction accuracy and PCC evaluating the preservation of emotional intensity relationships across the continuous valence-arousal space.

### 4.4 Implementation Details

Our dual-layer Agent-Executor framework leverages different model configurations optimized for their respective roles in the annotation pipeline.

**Agent**: The agent employs GPT-o3 as the underlying language model. This high-capacity model provides the advanced reasoning capabilities necessary for analyzing complex annotation patterns, identifying systematic errors in outlier feedback, and formulating comprehensive guideline refinements that address domain-specific challenges in medical text emotion analysis.

**Executor**: The Executor utilizes GPT-4o-mini as the base model. This configuration balances annotation quality with computational efficiency, enabling cost-effective processing of the extensive validation and test datasets while maintaining consistent application of the Agent's refined guidelines across all instances.

## 5 Results

| Config | V-MAE | V-PCC | A-MAE | A-PCC |
|---|---|---|---|---|
| Baseline | 0.5095 | **0.7676** | 0.9381 | 0.5745 |
| Conservative | **0.5105** | 0.7625 | **0.8661** | **0.5860** |
| Aggresive | 0.5470 | 0.5105 | 0.9964 | 0.5461 |

**Table 1:Results under different configurations. These configurations all use the same training, validation, and test data sizes.validation, and test data sizes.**

Table 1 presents the comparative performance of our dual-layer Agent-Executor framework across three experimental configurations on the ROCLING 2025 shared task. The results demonstrate the effectiveness of our outlier-driven feedback mechanism for improving dimensional emotion prediction in medical self-reflection texts.

**Arousal Prediction Benefits More from Iterative Refinement** Conservative configuration achieves the best overall valence prediction results, with V-MAE of 0.5105 and V-PCC of 0.7625, representing improvements over the Baseline While the MAE shows marginal improvement, the slight decrease in PCC suggests that the conservative outlier detection may not capture sufficient feedback for substantial correlation enhancement. The Aggressive configuration shows degraded performance, indicating that overly broad outlier detection may introduce noise that compromises guideline quality.

The outlier-driven feedback mechanism demonstrates more pronounced improvements in arousal prediction. The Conservative configuration achieves substantial improvements with A-MAE of 0.8661and A-PCC of 0.5860. This represents a reduction in prediction error by approximately 7.7% and an improvement in correlation by 2.0%. The Aggressive configuration shows mixed results with A-MAE of 0.9964 (worse than baseline) but maintains comparable correlation performance

**Aggressive Feedback Threshold Leads to Performance Degradation** Conservative approach consistently outperforms both Baseline and Aggressive configurations across most metrics, particularly for arousal prediction. This suggests that targeted identification of the most significant prediction failures provides optimal feedback for guideline refinement without introducing excessive noise. The Aggressive approach appears to suffer from over-correction, where the inclusion of marginal outliers leads to guideline instability and reduced prediction accuracy.

**Context Complexity Makes Arousal Assessment More Challenging Than Valence** Results reveal interesting asymmetries between valence and arousal prediction improvements. The outlier-driven mechanism shows greater effectiveness for arousal prediction, possibly indicating that arousal-related annotation guidelines benefit more from iterative refinement compared to valence guidelines. This may reflect the inherent complexity of arousal assessment in medical contexts, where emotional intensity can be more ambiguous than emotional polarity.

**Overall** The results validate our hypothesis that systematic outlier identification and feedback can enhance dimensional emotion prediction, with the Conservative configuration representing the optimal balance between comprehensive feedback and guideline stability.

## 6 Conclusion

This paper presented a novel dual-layer Agent-Executor framework for iterative guideline refinement in dimensional emotion analysis, specifically designed to address the challenges of

valence-arousal prediction in medical self-reflection texts. Our approach systematically combines high-level reasoning capabilities with efficient execution through a hierarchical architecture that enables cost-effective scaling while maintaining annotation quality.

The key innovation lies in our outlier-driven feedback mechanism, which transforms prediction errors from isolated failures into systematic learning opportunities. By employing IQR-based outlier detection, we identified problematic predictions and fed them back to the Agent component for targeted guideline improvements.

This iterative refinement process enables continuous adaptation to domain-specific challenges without requiring extensive manual annotation efforts.

Our experimental evaluation on the ROCLING 2025 shared task demonstrated the effectiveness of this approach. The Conservative outlier detection configuration achieved optimal performance balance, with particularly notable improvements in arousal prediction. The results reveal important insights about dimensional emotion analysis in medical contexts: arousal assessment benefits more from iterative refinement than valence prediction, suggesting that emotional intensity evaluation presents greater annotation challenges than emotional polarity in healthcare narratives.

The framework's practical contributions extend beyond performance improvements. The dual-layer architecture provides a cost-effective solution that leverages expensive high-capacity models only for guideline construction while using efficient models for large-scale annotation. The outlier-driven feedback mechanism offers interpretability through explicit identification of systematic weaknesses, enabling targeted improvements rather than global parameter adjustments.

**Limitations and Future Work** While our approach shows promising results, several limitations warrant acknowledgment. The framework's effectiveness depends on the quality of initial guidelines, and extremely poor starting points may require multiple refinement iterations. The IQR-based outlier detection, while robust, may not capture all forms of systematic errors, particularly those involving subtle contextual nuances. Future work should explore more sophisticated outlier detection methods that incorporate semantic similarity and domain-specific error patterns.

Additionally, our evaluation focused on a single domain (medical self-reflection texts) and language (Chinese). Extending the framework to multilingual settings and diverse text domains would strengthen its generalizability claims. Investigation of different Agent-Executor model combinations and the integration of human-in-the-loop refinement mechanisms represent promising research directions.

The proposed dual-layer Agent-Executor framework with outlier-driven feedback provides a principled approach to iterative guideline improvement in dimensional emotion analysis, offering both practical benefits for annotation quality and theoretical insights into systematic error correction in specialized domains.

## References


Nurudin Alvarez-Gonzalez, Andreas Kaltenbrunner, and Vicenç Gómez. 2021. Uncovering the Limits of Text-based Emotion Detection. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2560–2583, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christopher Bagdon, Prathamesh Karmalkar, Harsha Gurulingappa, and Roman Klinger. 2024. "You are an expert annotator": Automatic Best–Worst-Scaling Annotations for Emotion Intensity Modeling. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7924–7936, Mexico City, Mexico. Association for Computational Linguistics.

Victoria Bobicev and Marina Sokolova. 2018. Thumbs Up and Down: Sentiment Analysis of Medical Online Forums. In Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task, pages 22–26, Brussels, Belgium. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and

Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020), pages 1877–1901.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Fabio Giachelle, Ornella Irrera, and Gianmaria Silvello. 2021. MedTAG: a portable and customizable annotation tool for biomedical documents. BMC Medical Informatics and Decision Making, 21(1):1–12.

Chi Hu, Yimin Hu, Hang Cao, Tong Xiao, and JingBo Zhu. 2024. Teaching Language Models to Self-Improve by Learning from Language Feedback. In Findings of the Association for Computational Linguistics: ACL 2024, pages 6090–6101, Bangkok, Thailand. Association for Computational Linguistics.

Lung-Hao Lee, Tzu-Mi Lin, Hsiu-Min Shih, Kuo-Kai Shyu, Anna S. Hsu, and Peih-Ying Lu. 2025. ROCLING-2025 Shared Task: Chinese Dimensional Sentiment Analysis for Medical Self-Reflection Texts. In Proceedings of the 37th Conference on Computational Linguistics and Speech Processing.

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese EmoBank: Building Valence-Arousal Resources for Dimensional Sentiment Analysis. ACM Transactions on Asian and Low-Resource Language Information Processing, 21(4): Article 65, 1-18.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023), pages 46534–46594.

Michail Mitsios, Georgios Vamvoukakis, Georgia Maniati, Nikolaos Ellinas, Georgios Dimitriou, Konstantinos Markopoulos, Panos Kakoulidis, Alexandra Vioni, Myrsini Christidou, Junkwang Oh, Gunu Jho, Inchul Hwang, Georgios Vardaxoglou, Aimilios Chalamandaris, Pirros Tsiakoulis, and Spyros Raptis. 2024. Improved Text Emotion Prediction Using Combined Valence and Arousal Ordinal Classification. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 808–813, Mexico City, Mexico. Association for Computational Linguistics.

Saif Mohammad. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems 35 (NeurIPS 2022), pages 27730–27744.

Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, and Alice Oh. 2021. Dimensional Emotion Detection from Categorical Emotion. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4367–4380, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

James A. Russell. 1980. A circumplex model of affect. Journal of Personality and Social Psychology, 39(6):1161–1178.

Daniela Teodorescu, Tiffany Cheng, Alona Fyshe, and Saif Mohammad. 2023. Language and Mental Health: Measures of Emotion Dynamics from Text as Linguistic Biosocial Markers. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 3117–3133, Singapore. Association for Computational Linguistics.

Xingchen Wan, Ruoxi Sun, Hanjun Dai, Sercan Arik, and Tomas Pfister. 2023a. Better Zero-Shot Reasoning with Self-Adaptive Prompting. In Findings of the Association for Computational Linguistics: ACL 2023, pages 3493–3514, Toronto, Canada. Association for Computational Linguistics.

Xingchen Wan, Ruoxi Sun, Hootan Nakhost, Hanjun Dai, Julian Martin Eisenschlos, Sercan Arik, and Tomas Pfister. 2023b. Universal self-adaptive prompting. Computing Research Repository, arXiv:2305.14926.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024. Executable code actions elicit better LLM agents. In

Proceedings of the 41st International Conference on Machine Learning, pages 50682–50695.

Qiang Wei, Amy Franklin, Trevor Cohen, and Hua Xu. 2021. Clinical text annotation: What factors are associated with the cost of time? In AMIA Annual Symposium Proceedings, volume 2018, pages 1552–1560.

Qinyuan Ye, Mohamed Ahmed, Reid Pryzant, and Fereshte Khani. 2024. Prompt Engineering a Prompt Engineer. In Findings of the Association for Computational Linguistics: ACL 2024, pages 355–385, Bangkok, Thailand. Association for Computational Linguistics.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. 2024. Calibrating the Confidence of Large Language Models by Eliciting Fidelity. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 2959–2979, Miami, Florida, USA. Association for Computational Linguistics.

Qi Zhao, Haotian Fu, Chen Sun, and George Konidaris. 2024. EPO: Hierarchical LLM agents with environment preference optimization. Computing Research Repository, arXiv:2408.16090.