PAI at SemEval-2026 Task 3: An LLM and Data Redistribution Adaptation-Based Predictive Strategy for Valence-Arousal Scores

Zhihao Ruan*, Kaifeng Yang*, Cheng Chen,

{archfool.ruan, yangkaifeng1985}@gmail.com, cchen237-c@my.cityu.edu.hk
Wenwen Dai, Wenjia Mao
{jk123124dww, nora.davis.wj695}@gmail.com
Ping An Life Insurance Company of China

Abstract

To address the valence and arousal score prediction task in Dimensional Aspect-Based Sentiment Analysis (DimABSA), we propose a two-stage strategy. In the first stage, we conduct post-training on a Large Language Model (LLM) via a Supervised Fine-Tuning (SFT) scheme, followed by generating initial predictions for valence and arousal scores. In the second stage, we perform distribution adaptation on the initial results by leveraging the training set distribution through various techniques, including Gaussian distribution modeling, quantile mapping, and the Sinkhorn algorithm. Our system achieved first place on eight leaderboards across multiple tracks in Task 3.

1 Introduction

SemEval-2026 Task 3(Yu et al., 2026; Lee et al., 2026; Becker et al., 2026) comprises multiple tracks encompassing various tasks, such as score prediction, entity extraction, and classification. Notably, the prediction of valence and arousal (VA) scores(Russell, 1980) is a common component across all tracks. This paper focuses specifically on the VA prediction task, where valence and arousal serve as the two primary dimensions for emotion assessment in Dimensional Aspect-Based Sentiment Analysis (DimABSA). Within this framework, Dimensional Aspect Sentiment Triplet Extraction (DimASTE) jointly identifies aspect-opinion pairs and their corresponding VA scores, whereas Dimensional Aspect Sentiment Quadruplet Prediction (DimASQP) extends this extraction objective by additionally classifying the domain-specific aspect categories. Furthermore, Dimensional Stance Analysis (DimStance) broadens the applicability of this continuous representation by treating stance targets as aspects, thereby reformulating traditional stance detection as a multidimensional VA regression task(Yu et al., 2026; Lee et al., 2026; Becker et al., 2026).

We identify two primary challenges in VA score prediction. First, cross-lingual disparities exist in the perception and determination of VA intensity for various emotional lexicons (e.g., happy, relaxed, tired) across different languages, such as English, Chinese, and Russian. Second, VA labeling is inherently subjective, leading to inter-annotator variability. Furthermore, the task organizers often apply heuristic corrections to the gold standard; our empirical analysis also reveals a significant correlation between valence and arousal dimensions.

To address the first challenge, we employ a post-training approach to enable Large Language Models (LLMs) to capture language-specific nuances in emotional dimensions. Specifically, we utilize the LoRA (Low-Rank Adaptation) technique within a Supervised Fine-Tuning (SFT) framework. To tackle the second challenge, we conduct a statistical analysis of the training data distribution. We then implement distributional adaptation strategies—including Gaussian modeling, quantile mapping, and the Sinkhorn algorithm—to align the predicted two-dimensional VA scores with the training set distribution. This process also serves to reinforce the inherent correlation between the two dimensions.

Given that Subtask 1 is a pure VA score prediction task and our unified system for all subtasks is trained on its data, the subsequent experiments, data analysis, and discussions in this paper will be centered on Subtask 1.

2 Related Work

2.1 Large Language Models in Affective Computing

Recent advancements in Large Language Models (LLMs) have significantly catalyzed the development of affective computing. In this study, we

*Equal contributions

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1489–1494

ST	lang	domain	score	baseline
1	eng	lap	1.4394	2.1893
1	eng	res	1.2141	2.1461
1	jpn	fin	0.7584	1.6396
1	jpn	hot	0.6508	1.7553
1	rus	res	1.219	1.7768
1	tat	res	1.5294	1.9380
1	ukr	res	1.1888	1.7805
1	zho	fin	0.5977	1.9652
1	zho	lap	0.68	1.6440
1	zho	res	0.9766	1.8959
2	eng	lap	0.6169	0.4424
2	eng	res	0.6903	0.4920
2	jpn	hot	0.5682	0.3464
2	rus	res	0.5793	0.4242
2	tat	res	0.4908	0.3577
2	ukr	res	0.5787	0.4220
2	zho	lap	0.5306	0.2494
2	zho	res	0.5638	0.3529
3	eng	lap	0.3758	0.2795
3	rus	res	0.5599	0.2963
3	tat	res	0.4523	0.2380
3	ukr	res	0.5437	0.2971
3	zho	lap	0.4316	0.1900
3	zho	res	0.536	0.2859

Table 1: Task 3 Track A Test Dataset Results

ST	lang	domain	score	baseline
1	deu	pol	1.511	1.5914
1	eng	env	1.6768	1.6431
1	pcm	pol	1.1399	1.7392
1	swa	pol	2.2519	2.2992
1	zho	env	0.6269	0.7403

Table 2: Task 3 Track B Test Dataset Results

leverage both proprietary and open-source LLMs to capture the complex semantics of emotional dimensions.

Specifically, we employ the Gemini-3(Google DeepMind, 2025) and Grok-4(xAI, 2025) architectures via their respective high-performance APIs.

In parallel, to further specialize the models for the nuances of valence and arousal dimensions, we utilize the Qwen3(Yang et al., 2025) series as our open-source backbone. We perform domain-specific post-training—specifically via Supervised Fine-Tuning (SFT)—across three scaling variants: Qwen3-8B, Qwen3-14B, and Qwen3-32B. This allows for a granular comparison between zero-shot API-based inference and parameter-efficient fine-tuning on specialized sentiment datasets.

2.2 Statistical Modeling and Distributional Alignment

The prediction of continuous emotional scores often suffers from distributional shifts between training and inference phases. To mitigate this, we incorporate classical statistical frameworks, namely Gaussian (Normal) distribution and Quantile-based mapping, to model the empirical characteristics of the valence-arousal space. These methods serve as a foundational step for capturing the central tendency and dispersion of annotator subjectivity.

Furthermore, to achieve a more rigorous alignment between the predicted results and the ground-truth training distribution, we introduce Optimal Transport (OT) theory. Traditional OT solvers are often computationally prohibitive for real-time inference; however, the Sinkhorn algorithm(Cuturi, 2013) circumvents this by introducing entropic regularization. This transformation allows for an efficient iterative solution to the Monge-Kantorovich problem, yielding an optimal transport plan with significantly reduced complexity. In our framework, the Sinkhorn algorithm is specifically employed to adapt the predicted valence-arousal distributions, ensuring that the model outputs strictly conform to the prior statistical properties of the human-annotated training set.

3 System Overview

The proposed system adopts a sequential dual-stage framework designed to accurately estimate valence and arousal scores. The first stage leverages the linguistic power of Large Language Models (LLMs) for initial score generation, while the second stage

applies distributional adaptation to calibrate predictions and preserve the inter-dimensional correlation between valence and arousal.

3.1 Stage I: LLM-based Valence-Arousal Estimation

In the primary stage, we formulate the Valence-Arousal (VA) prediction as a conditional regression task. For each instance, the model receives an input tuple consisting of the Text and the target Aspect, and is tasked with generating the corresponding V and A scores. To evaluate the efficacy of different architectures, we compare two paradigms: proprietary API-based inference and open-source fine-tuning.

Proprietary Models via In-Context Learning We utilize four models: gemini-3-flash-preview, grok-4-fast-reasoning, grok-4-fast, and grok-4. To enable these models to internalize the domain-specific annotation standards, we implement a few-shot prompting strategy. Specifically, each prompt is augmented with 10–20 curated conversation turns, where each turn provides a gold-standard example from the training set.

Open-source Models via Parameter-Efficient Fine-Tuning We employ the Qwen3 series (8B, 14B, and 32B) as our open-source backbones. Unlike the proprietary models, these are adapted via Supervised Fine-Tuning (SFT) focusing on Low-Rank Adaptation (LoRA) modules. Notably, during the inference phase, these models perform zero-shot prediction without additional few-shot exemplars, as the task-specific knowledge is already embedded in the fine-tuned parameters.

As illustrated in Table 3, the fine-tuned open-source models consistently outperform the few-shot proprietary APIs. Furthermore, within the Qwen3 family, we observed that increasing model scale from 8B to 32B parameters yielded marginal performance gains, suggesting that moderate-sized models may be sufficient for capturing the nuances of this specific task.

model	lang-domain	Score
gemini-3-flash-p	rus-res	1.3749
grok-4-fast	rus-res	1.3924
grok-4-fast-r	rus-res	1.5341
grok-4	rus-res	1.471
Qwen3-8B	rus-res	1.0628
Qwen3-14B	rus-res	1.0398
Qwen3-32B	rus-res	1.1599
gemini-3-flash-p	tat-res	1.432
grok-4-fast	tat-res	1.559
grok-4-fast-r	tat-res	1.5521
grok-4	tat-res	1.5912
Qwen3-8B	rus-res	1.4359
Qwen3-14B	rus-res	1.2918
Qwen3-32B	rus-res	1.4309
gemini-3-flash-p	ukr-res	1.388
grok-4-fast	ukr-res	1.4289
grok-4-fast-r	ukr-res	1.5263
grok-4	ukr-res	1.3963
Qwen3-8B	rus-res	1.1949
Qwen3-14B	rus-res	1.1412
Qwen3-32B	rus-res	1.1744

Table 3: Task 3 Track A Subtask 1 Dev Dataset Results in language of Russian, Tatar and Ukrainian in Stage I

3.2 Stage II: Data Redistribution Adaptation

In the second stage, we introduce a distributional adaptation strategy to rectify systematic prediction biases inherent in the LLM outputs. As depicted in Figure 1 and Figure 2, a significant correlation exists between valence and arousal scores. Therefore, our adaptation process not only aligns the marginal distributions but also calibrates the output based on this inter-dimensional dependency. This stage takes the raw VA predictions from Stage I as input and produces the final calibrated scores.

We investigated three distinct mapping strategies:

Gaussian (Normal) Distribution Mapping Aligning the mean and variance of predictions to match the training set.

Quantile Distribution Mapping Matching the cumulative distribution functions (CDF) to ensure percentile consistency.

Sinkhorn Algorithm Applying Optimal Transport (OT) theory to align the joint distribution of valence and arousal.

While Gaussian and Quantile mapping are univariate approaches that process each dimension independently, they fail to account for the joint distribution of V and A. In contrast, the Sinkhorn algorithm treats the VA scores as a bivariate distribution. As reported in Table 4, although all three strategies improve performance, the Sinkhorn algorithm demonstrates superior accuracy and greater stability, effectively bridging the distributional gap

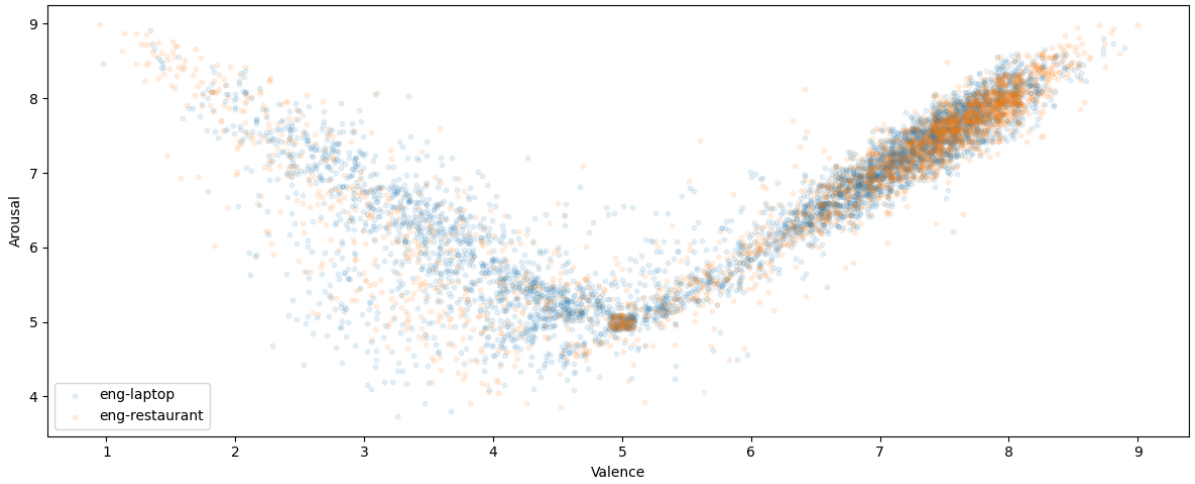

Figure 1: The data distribution of valence-arousal scores across multiple English domains in the training set.

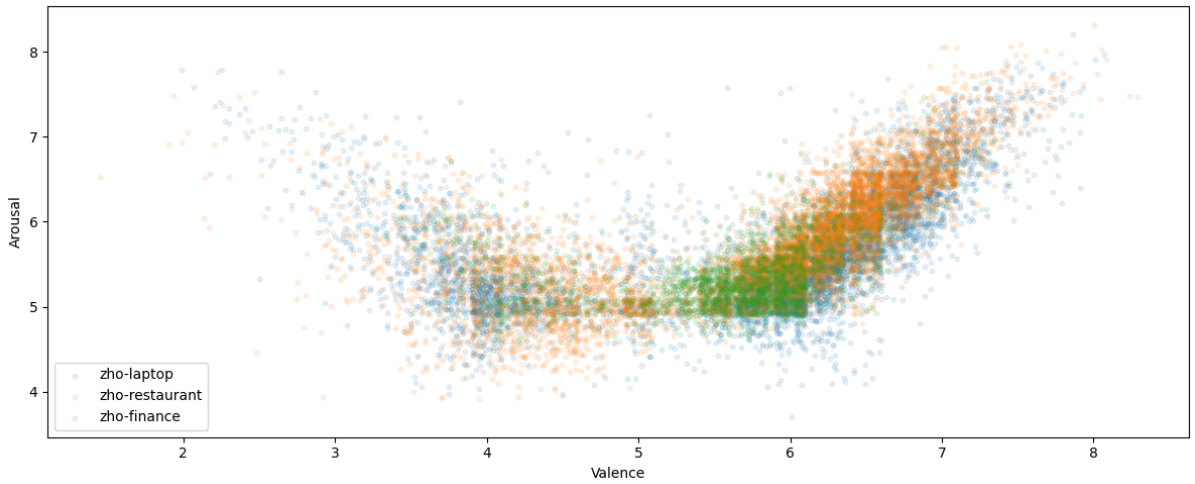

Figure 2: The data distribution of valence-arousal scores across multiple Chinese domains in the training set.

between the test predictions and the empirical training data.

3.3 Brief Overview of Strategies for Track A, Subtasks 2 and 3

While the methodologies applied in Subtasks 2 and 3 are not the central focus of this paper, we provide a brief overview for completeness.

In Subtasks 2 and 3 of Track A, alongside the prediction of valence-arousal (VA) scores, our PAI system is required to extract three additional fields: Aspect, Opinion, and Category.

We decompose the overarching prediction objective into a pipeline of granular sub-tasks. The process initiates with Aspect extraction. Subsequently, conditioned on the extracted Aspects, the corresponding Opinions are identified. Finally, leveraging the paired Aspects and Opinions, the Category

strategy	lang-domain	Score
Gaussian	rus-res	1.0628
Quantile	rus-res	1.1196
Sinkhorn	rus-res	1.0372
Gaussian	tat-res	1.4359
Quantile	tat-res	1.5217
Sinkhorn	tat-res	1.3535
Gaussian	ukr-res	1.1949
Quantile	ukr-res	1.161
Sinkhorn	ukr-res	1.0834

Table 4: Task 3 Track A Subtask 1 Dev Dataset Results in language of Russian, Tatar and Ukrainian in Stage II

and VA scores are predicted independently.

These sub-tasks are universally executed utilizing conventional Large Language Model (LLM)-based paradigms. Specifically, Aspect and Opinion extraction is formulated as a Named Entity Recognition (NER)-style task, Category prediction is addressed as a standard classification problem, and the estimation of VA scores is modeled as a linear regression task.

4 Experimental Setup

To ensure the reproducibility of our results, this section details the hyperparameter configurations and algorithmic parameters utilized for both the LLM fine-tuning and the subsequent distributional adaptation.

4.1 Implementation Details for LLM Fine-tuning

The open-source Qwen3 models were fine-tuned using the Supervised Fine-Tuning (SFT) paradigm integrated with Low-Rank Adaptation (LoRA). The training was conducted for 1 epoch with a learning rate initialized at 1×10^{-4}. We employed the AdamW optimizer and a cosine learning rate scheduler with zero warmup steps to manage the optimization process.

To balance computational efficiency and convergence stability, the training was performed using bfloat16 (BF16) mixed-precision. The effective batch size was configured to 8 by setting the per-device batch size to 1 and the gradient accumulation steps to 8. Additionally, the maximum gradient norm was clipped at 1.0 to prevent gradient explosion. For the LoRA configuration, we set the rank (r) to 8, the scaling factor (α) to 16, and a dropout rate of 0 to maximize parameter efficiency while maintaining the model's representative capacity.

4.2 Parameters for Distributional Adaptation

For the second stage of our framework, the Sinkhorn algorithm was implemented to perform optimal transport-based alignment. The cost matrix for the transport plan was computed using the squared Euclidean distance ($sqeuclidean$). To ensure a smooth transport plan and facilitate convergence, the entropic regularization parameter (ϵ) was set to 1. The optimization process was allowed a maximum of $10,000$ iterations to ensure that the algorithm reached a stable convergence state before termination.

5 Result

The final configuration of our proposed framework integrates the Qwen3-32B model(although Qwen3-14B achieved superior results on the Dev Dataset, we opted for the more generalizable Qwen3-32B model on the Test Dataset to ensure robustness) for Stage I and the Sinkhorn algorithm for Stage II. Our performance on the test set is summarized in Table 1 and Table 2 .

The predictions were submitted by team PAI under the account name archfool.

As specified in the task paper(Yu et al., 2026), two baselines were provided for both Track A and Track B. Specifically, Track A introduced Kimi K2 and Qwen3-14B as baselines, while Mistral-3-14B and mBERT were utilized for Track B. In our comparative analysis, we report the results of the superior-performing baseline for each track as shown in the table.

Our proposed methodology demonstrated competitive performance in Track A. Notably, our system attained first-place rankings across a total of eight sub-tracks:

Subtask-1 rus-res, tat-res, and ukr-res;
Subtask-2 rus-res, ukr-res, and zho-res;
Subtask-3 rus-res and ukr-res.

6 Conclusion

In this paper, we presented a robust dual-stage predictive strategy for valence and arousal score estimation within the Dimensional Aspect-Based Sentiment Analysis (DimABSA) framework. To address the inherent challenges of cross-lingual perceptual variance and annotator subjectivity, our approach effectively synergizes the linguistic reasoning of Large Language Models with advanced distributional alignment techniques.

Our empirical evaluation leads to several key findings. First, while proprietary models like Gemini-3 and Grok-4 demonstrate strong zero-shot capabilities through few-shot prompting, the fine-tuned open-source Qwen3-32B model consistently delivers superior performance by internalizing domain-specific emotional nuances via LoRA-based SFT. Second, we demonstrate that raw LLM predictions often deviate from the empirical distribution of human labels. By incorporating the Sinkhorn algorithm in the second stage, our system successfully aligns the joint distribution of valence and arousal, preserving their intrinsic correlation while significantly enhancing prediction stability.

References

Jonas Becker, Liang-Chih Yu, Shamsuddeen Hassan Muhammad, Jan Philip Wahle, Terry Ruas, Idris Abdulmumin, Lung-Hao Lee, Nelson Odhiambo, Lilian Wanzare, Wen-Ni Liu, Tzu-Mi Lin, Zhe-Yu Xu, Ying-Lung Lin, Jin Wang, Maryam Ibrahim Mukhtar, Bela Gipp, and Saif M. Mohammad. 2026. Dimstance: Multilingual datasets for dimensional stance analysis. *Preprint*, arXiv:2601.21483.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Google DeepMind. 2025. Gemini 3 flash model card.

Lung-Hao Lee, Liang-Chih Yu, Natalia Loukashevich, Ilseyar Alimova, Alexander Panchenko, Tzu-Mi Lin, Zhe-Yu Xu, Jian-Yu Zhou, Guangmin Zheng, Jin Wang, Sharanya Awasthi, Jonas Becker, Jan Philip Wahle, Terry Ruas, Shamsuddeen Hassan Muhammad, and Saif M. Mohammad. 2026. Dimabsa: Building multilingual and multidomain datasets for dimensional aspect-based sentiment analysis. *Preprint*, arXiv:2601.23022.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

xAI. 2025. Grok 4 model card: Advancing reasoning and agentic capabilities. Technical report, xAI.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Liang-Chih Yu, Jonas Becker, Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Lung-Hao Lee, Ying-Lung Lin, Jin Wang, Jan Philip Wahle, Terry Ruas, Alexander Panchenko, Ilseyar Alimova, Kai-Wei Chang, Lilian Wanzare, Nelson Odhiambo, Bela Gipp, and Saif M. Mohammad. 2026. SemEval-2026 task 3: Dimensional aspect-based sentiment analysis (DimABSA). In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.