SRCB at SemEval-2026 Task 3: Boosting DimASR via Contrastive LLM-Based Data Augmentation

Hongyu Li, Yuming Zhang, Junyu Zhou, Yongwei Zhang, Shanshan Jiang, Bin Dong
Ricoh Software Research Center (Beijing) Co., Ltd
{Hongyu.Li, Yuming.Zhang1, Junyu.Zhou,
Yongwei.Zhang, Shanshan.Jiang, Bin.Dong}@cn.ricoh.com

Abstract

We present our system for the DimASR subtask of SemEval-2026 Task 3: DimABSA, targeting dimensional sentiment regression of Valence-Arousal scores in English restaurant reviews. Our approach leverages Qwen3 large language models combined with contrastive LLM-based data augmentation to enrich training data and capture subtle affective variations. Experiments show that this data augmentation framework significantly improves performance on DimASR task, particularly in capturing subtle affective shifts at the aspect level. Finally, our system achieves a score of 1.227 RMSE on the test set.

1 Introduction

The field of Aspect-Based Sentiment Analysis (ABSA) has traditionally relied on categorical labels—such as positive, negative, and neutral—to characterize user opinions at the aspect level (Pontiki et al., 2014, 2015, 2016). While effective for general classification, this approach often fails to capture the subtle lexical intensities and affective nuances inherent in natural language. To address these limitations, the SemEval-2026 Task 3: DimABSA (Yu et al., 2026) introduces a shift toward dimensional sentiment representation, modeling opinions along the continuous axes of Valence (V) and Arousal (A). In this paper, we describe our system for Subtask 1: Dimensional Aspect Sentiment Regression (DimASR). Given a sentence and its identified aspect terms, the goal of DimASR is to predict real-valued VA scores that reflect fine-grained sentiment intensity. This subtask utilizes the DimABSA dataset (Lee et al., 2026), a large-scale multilingual resource spanning six languages—Chinese, English, Japanese, Russian, Tatar, and Ukrainian—and multiple domains including customer reviews (Hotel, Laptop, Restaurant) and financial reporting (France).

In this work, we focus specifically on the English language subset of the restaurant domain within Subtask 1. While traditional regression approaches often utilize encoder-based architectures like BERT (Devlin et al., 2019), the recent success of large-scale generative models suggests that decoder-only Transformers possess superior contextual understanding. Our proposed system leverages the Qwen3 series (Yang et al., 2025), specifically the 14B and 32B variants, as the backbone for feature extraction. To bridge the gap between generative pre-training and continuous value prediction, we augment the backbone with a lightweight regression head and a Sigmoid-based scaling mechanism to map hidden states directly to the [1,9] Valence-Arousal scale.

The primary contribution of our work is a contrastive LLM-based data augmentation framework. To overcome data scarcity and capture subtle affective shifts, we employ GLM-4.7 (Team et al., 2025) to generate four directional variants (higher/lower V or A) for each training instance. To ensure the reliability of these pseudo-labels, we implement a rigorous quality control pipeline consisting of consistency filtering via 5-fold cross-validation (to ensure prediction stability), monotonicity checking (to enforce correct affective direction), and a magnitude constraint (Δ Filtering) (to guarantee sufficiently large and informative affective differences). Furthermore, we apply an LLM-based restoration step to align the pre-tokenized training text with the natural casing of the test set. Our experiments demonstrate that the proposed contrastive LLM-based data augmentation framework significantly improves performance on the DimASR task.

2 System Overview

2.1 Task Definition

Dimensional Aspect Sentiment Regression (DimASR) is formulated as follows: given a sentence

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 290–294

Qwen3 Input Prompt Template

Given the sentence: {text} and the aspect: {aspect}, use one emotional word to present the valence and arousal for the aspect:

Table 1: Prompt used in fine-tuning / inference.

$S = [w_1, \cdots, w_T]$ and a predefined aspect term a (a contiguous substring of S), the objective of Subtask 1 is to predict the valence and arousal (VA) scores associated with a. These scores are continuous values within the range $[1, 9]$, representing the intensity and polarity of sentiment along the dimensional affective space.

2.2 Overall Framework

To address the DimASR task, we propose a framework built upon large pre-trained causal language models. Our method leverages the strong contextual representation capability of decoder-only Transformers while introducing minimal task-specific parameters for dimensional sentiment regression.

2.2.1 Backbone Language Model

We employ Qwen3 (Yang et al., 2025) series models (14B and 32B variants) as the backbone. These models follow the GPT/Llama-style autoregressive architecture, consisting of stacked transformer decoder layers with causal attention masks. Given an input sentence $S = [w_1, \cdots, w_T]$ and a predefined aspect term a which is a contiguous substring of S, the input is formatted according to the Qwen3 Input Prompt Template (Table 1). This prompt design is crafted to constrain the model to output a single "emotional word," it forces the LLM to leverage its internal semantic knowledge to map complex aspect-based sentiments into a unified representative token that encapsulates both valence and arousal. The model then processes the tokenized input and produces contextualized hidden representations for each token position.

2.2.2 Feature Extraction

Let $\mathbf{H} \in \mathbb{R}^{T \times d}$ denote the last-layer hidden states output by the backbone model, where d is the hidden dimension. To obtain a fixed-dimensional representation for aspect-specific sentiment prediction, we extract the hidden state corresponding to the last token of the input sequence, denoted as $\mathbf{h}_{\text{last}} \in \mathbb{R}^d$. This representation aggregates contextual information from the entire sequence through the causal attention mechanism and serves as the input to the regression head.

2.2.3 Regression Head

A lightweight regression module is attached to the backbone to predict continuous valence and arousal scores. We adopt a unified regression head that outputs two-dimensional predictions:

$$\hat{\mathbf{y}} = \text{Sigmoid}(\mathbf{W}_r \mathbf{h}_{\text{last}} + \mathbf{b}_r) \quad (1)$$

where $\mathbf{W}_r \in \mathbb{R}^{2 \times d}$ and $\mathbf{b}_r \in \mathbb{R}^2$ are learnable parameters. The Sigmoid activation function constrains the raw outputs to the range $(0, 1)$, which are subsequently scaled to the target range $[1, 9]$ during post-processing. This design choice provides numerical stability and facilitates optimization.

To prevent gradient saturation in the Sigmoid function and ensure stable training, we initialize the regression head weights with small Gaussian noise ($\mathcal{N}(0, 0.01)$) and zero-initialize the bias terms. No additional non-linear layers are introduced between the hidden states and the output predictions, maintaining architectural simplicity and computational efficiency.

2.2.4 Training Objective

The model is optimized exclusively using Mean Squared Error (MSE) loss between the predicted valence-arousal pairs and the ground-truth annotations:

$$\mathcal{L} = \frac{1}{2N} \sum_{i=1}^{N} \left[(v_i - \hat{v}_i)^2 + (a_i - \hat{a}_i)^2 \right] \quad (2)$$

where N is the batch size, (v_i, a_i) are the target scores, and $(\hat{v}_i, \hat{a}_i)$ are the predictions. For the 14B variant, we perform full-parameter fine-tuning, while for the 32B variant, we adopt LoRA (Hu et al., 2022) to enable efficient adaptation.

2.2.5 Inference

During inference, the model takes a test instance and directly produces the valence and arousal predictions through a single forward pass. The Sigmoid outputs are linearly mapped from $(0, 1)$ to $[1, 9]$ via:

$$v_{\text{pred}} = 1 + 8 \cdot \hat{v}, \quad a_{\text{pred}} = 1 + 8 \cdot \hat{a} \quad (3)$$

2.3 Query Preprocessing

We observe an inconsistency between the training set and the validation/test sets: texts in the training set are pre-tokenized and lowercased, while the validation and test sets retain their original casing and natural word forms. To ensure consistency across all splits, we employ a large language model to restore the training instances to their original, untokenized form.

3 Data Augmentation

To improve the robustness and generalization of DimASR, we introduce a contrastive LLM-based data augmentation framework. Starting from the original (sentence, aspect) pairs, we generate affect-controlled variants that explicitly manipulate valence and arousal. To ensure that the resulting pseudo-labeled data is reliable and beneficial for training, we design a multi-stage quality control pipeline consisting of Consistency Filtering (Std Filtering), Monotonicity Checking, and Magnitude Constraint (Δ Filtering). These mechanisms jointly enforce prediction stability, directional correctness, and sufficient affective deviation, respectively. Through this process, only high-confidence augmented samples are retained, thereby strengthening the model's sensitivity to dimensional sentiment shifts while minimizing noise from unreliable generations.

3.1 Data Generation

To enhance the training data for DimASR, we employ a data augmentation strategy leveraging the GLM-4.7 (Team et al., 2025) model. For each (sentence, aspect) pair in the training set—excluding samples where the aspect term is 'NULL'—we generate four distinct variants. These variants are explicitly conditioned to exhibit either higher or lower valence, or higher or lower arousal, relative to the original sample. This yields up to four augmented instances per original example, each intended to modulate a specific dimension of the affective space.

All generated samples are subsequently scored by our best-performing intermediate 5-fold cross-validation model (selected based on overall validation performance). This model produces valence and arousal predictions for each augmented instance, which serve as the basis for subsequent filtering and labeling steps.

3.2 Data Filtering

To ensure the quality and reliability of the generated data, we apply three sequential filtering criteria:

a. Consistency Filtering (Std Filtering) For each generated sample, we examine the standard deviation of the valence and arousal predictions produced by the five cross-validation models. Samples for which either valence or arousal exhibits a standard deviation exceeding a predefined threshold are discarded. We experiment with threshold values of 0.1, 0.2, and 0.3, and also consider the case of no filtering for comparison. This step ensures that retained samples have low inter-model disagreement, indicating stable and reliable pseudo-label estimates.

b. Monotonicity Checking This step verifies whether the intended directional relationship between the generated sample and its original counterpart is preserved. Specifically, for variants intended to have higher valence or arousal, we require that the mean of the five cross-validation predictions—computed after removing the maximum and minimum values to reduce outlier influence—is greater than the corresponding score of the original sample. Symmetric constraints apply to variants intended to have lower scores. Samples that fail to satisfy these monotonicity constraints are discarded. This criterion guarantees that augmented samples faithfully reflect the intended affective shift, preventing mislabeled or contradictory instances.

c. Magnitude Constraint (Δ Filtering) Even when the intended monotonic direction is satisfied, we further require that the change in valence or arousal be non-trivial. Samples whose predicted valence or arousal differs from the original by less than 0.25 (in absolute terms) are filtered out. This threshold ensures that only augmented instances with sufficiently distinct affective intensity are retained. This step ensures that retained samples exhibit meaningful affective differences, avoiding weak or redundant augmentations that contribute little to model learning.

3.3 Label Integration

After filtering, each retained augmented sample is assigned a pseudo-label for valence and arousal. To obtain a robust point estimate from the five cross-validation predictions, we compute the median of the five predicted scores for each dimension. The

median is chosen over the mean due to its resilience to extreme values, providing a stable and representative pseudo-ground truth for subsequent training.

4 Experimental Results

4.1 Experiment Settings

We conduct experiments on the English restaurant-domain dataset using five-fold cross-validation on the training set. The predictions from the five validation folds are concatenated and evaluated against the entire training set using Root Mean Square Error (RMSE). Samples with aspect label "NULL" are excluded from evaluation.

All experiments are conducted on four NVIDIA A100 GPUs (80GB each). The global batch size is 64, and models are trained for 5 epochs.

For full-parameter fine-tuning, the backbone learning rate is set to 1×10^{-5} and the regression head learning rate to 1×10^{-4}. For LoRA-based fine-tuning, backbone learning rates of 1.5×10^{-4}, 2.5×10^{-4}, and 3×10^{-4} are explored, while keeping the regression head learning rate fixed at 1×10^{-4}.

For parameter-efficient tuning, we adopt LoRA with rank 16, $\alpha = 32$, and dropout rate 0.05, applied to the attention and feed-forward projection layers.

4.2 Main Results

As shown in Table 2, the 14B models used augmented data filtered at $\sigma \leq 0.2$ and $\sigma \leq 0.3$ demonstrate significant performance improvements over the baseline, showing the effectiveness of our contrastive LLM-based data augmentation framework for DimASR task. Due to time and resource constraints, some experiments for the 32B model are not conducted, including the full-parameter and LoRA fine-tuning baselines. The 32B model fine-tuned via LoRA with $\sigma < 0.2$ achieves a competitive performance (best 1.0210) that is closely comparable to the 14B augmented models.

We selected 12 models out of the fold models for the final test-time ensemble. Median aggregation was used to combine model predictions. The resulting system achieves an RMSE of 1.227 on the test set.

4.3 Additional Attempts

Regression Head Variants At the model architecture level, besides the unified regression head, we also explored two alternative designs: *two separate regression heads* and a *shared regression head*. The first design employs the same backbone language model but attaches two independent regression heads to decode Valence and Arousal separately. The second design uses a single shared regression head, where both Valence and Arousal are decoded through it, but controlled by two different prompts during training and inference. Experimental results showed that both alternative designs underperformed compared to the unified regression head.

Multi-Task Learning In addition to the primary regression task, we explored incorporating various auxiliary tasks, including classification tasks, a language modeling task (where the model generates Valence and Arousal scores using natural language), and learning-to-rank objectives (using Spearman or ListNet losses). Overall, these multi-task learning strategies did not yield the expected improvements. Most methods achieved performance comparable to the baseline, while some even resulted in significant degradation. Due to limited time, we did not extensively explore strategies for balancing the losses between multiple tasks.

Continued Pretraining Following prior work (Xu et al., 2024), we further pre-trained the backbone language model. Specifically, we used the texts of all training and validation samples to continue pretraining the Qwen-3-14B model. This approach yielded some improvements on certain folds, but overall performance across all five folds showed no significant difference compared to the baseline. Due to time constraints, we did not include results from this method in the final test submission.

5 Conclusion

In this work, we presented our system for the SemEval-2026 DimABSA task, focusing on dimensional sentiment regression in English restaurant reviews. Our system leverages the robust contextual representations of Qwen3 large language models, enhanced by a contrastive LLM-based data augmentation framework that explicitly modulates sentiment intensity. Rigorous filtering mechanisms, including consistency, monotonicity, and magnitude constraint checks, ensured the integration of high-quality pseudo-labeled data, which significantly boosted model performance. Ultimately,

Model Config	Aug. Threshold	LR	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Overall
M_{14B} (Baseline)	-	1×10^{-5}	1.0672	1.1453	1.0380	1.0652	1.0372	1.0727
$M_{14B} + \mathcal{D}_{aug}$	$\sigma \leq 0.2$	1×10^{-5}	**0.9650**†	1.0980†	0.9945†	1.0159†	**0.9803**†	**1.0131**
$M_{14B} + \mathcal{D}_{aug}$	$\sigma \leq 0.3$	1×10^{-5}	0.9960	1.0972	**0.9869**	**0.9940**	0.9898	1.0151
M_{32B} + LoRA $+\mathcal{D}_{aug}$	$\sigma < 0.2$	1.5×10^{-4}	0.9820	1.0967†	0.9984	1.0026†	1.0154	1.0210
		2.5×10^{-4}	0.9770	1.0973†	1.0489	1.0003†	1.0096	1.0281
		3.0×10^{-4}	0.9918	**1.0950**†	0.9941†	1.0076†	1.0168	1.0230

Table 2: Performance comparison across different model scales, data augmentation filtering std thresholds (σ), and learning rates (LR). Results are reported in RMSE. Fold models marked with † are selected for the final ensemble.

our ensemble of fine-tuned 14B and 32B models achieved a competitive RMSE of 1.227 on the test set, demonstrating the efficacy of our contrastive LLM-based data augmentation framework for DimASR task.

References

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

GLM Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, and 152 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *Preprint*, arXiv:2508.06471.

Hongling Xu, Delong Zhang, Yice Zhang, and Ruifeng Xu. 2024. HITSZ-HLT at SIGHAN-2024 dimABSA task: Integrating BERT and LLM for Chinese dimensional aspect-based sentiment analysis. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 175–185, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 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.