SokraTUM at SemEval-2026 Task 3: A hybrid cascade of Label Distribution Learning, RAG supported generative extraction and contrastive metric learning for dimensional sentiment analysis

Denis Laschenko*[1] **Albert Korotyk*[1]**
[1]Technical University of Munich,
(d.laschenko, albert.korotyk)@tum.de

Abstract

The Dimensional ABSA (DimABSA) shared task extends traditional aspect-based sentiment analysis from categorical polarity to continuous valence–arousal (VA) prediction. We present our system for all three subtasks: Dimensional Aspect Sentiment Regression (DimASR), Dimensional Aspect Sentiment Triplet Extraction (DimASTE), and Dimensional Aspect Sentiment Quad Prediction (DimASQP).

Due to the cascading nature of the different subtasks, we built a modular interlocking pipeline that uses classical Machine Learning and NLP methods.

Experiments across domains show consistent gains in regression accuracy and structured extraction performance. Our results demonstrate the effectiveness of distribution-aware regression, retrieval-augmented generation, and contrastive prototype learning for dimensional sentiment analysis.

1 Introduction

Aspect-Based Sentiment Analysis (ABSA) typically models sentiment polarity as discrete labels (positive, negative, neutral) (Pontiki et al., 2014) (Zhang et al., 2022). However, human affect is inherently continuous. The valence–arousal (VA) framework (Russell, 1980) represents sentiment along two real-valued dimensions, capturing both polarity and intensity. The Dimensional ABSA (DimABSA) shared task integrates this framework into structured sentiment analysis (Lee et al., 2026; Yu et al., 2026).

Consider the following restaurant review: *"average to good thai food, but terrible delivery."* The Dimensional ABSA (DimABSA) framework captures these differences using continuous Valence-Arousal (VA) scores. For example, the **thai food** is assigned (Valence: 6.75, Arousal: 6.38) to represent moderately high-energy positive satisfaction, while the **delivery** receives (Valence: 2.88, Arousal: 6.62) to represent high-energy negative frustration. This example illustrates that modeling sentiment as a continuous dimensional space provides a richer understanding of user experiences compared to discrete labels (A complete, structured JSONL input-output example is provided in Appendix D).

The task consists of three subtasks (Yu et al., 2026): (1) **DimASR**, predicting VA scores for given aspects; (2) **DimASTE**, extracting (Aspect, Opinion, VA) triplets; and (3) **DimASQP**, extracting (Aspect, Category, Opinion, VA) quadruplets. Compared to categorical ABSA, these tasks introduce additional challenges: subjective regression targets, span instability in generative extraction, and category imbalance. We address these challenges with task-specific designs. For DimASR, we apply Label Distribution Learning (LDL) (Geng, 2016) to model annotation uncertainty and optimize Concordance Correlation Coefficient (CCC) loss to better capture correlation and scale agreement (Trigeorgis et al., 2016). For DimASTE, we combine Parameter-Efficient Fine-Tuning (PEFT) (Hu et al., 2021) with Retrieval-Augmented In-Context Learning (RAG-ICL) (Lewis et al., 2021; Dong et al., 2024) to enhance large language model predictions and apply fuzzy grounding (Ratcliff et al., 1988) to align generated spans with the input text. For DimASQP, we use Supervised Contrastive Learning (SupCon) (Khosla et al., 2021) to structure the category embedding space and perform nearest centroid inference for robust prediction under few-shot and imbalanced settings.

Our approach provides a unified yet flexible pipeline for structured dimensional sentiment analysis and achieves competitive performance across subtasks.

2 Background

2.1 Task Description

The task evaluates systems on Dimensional ABSA across three hierarchical subtasks (Yu et al.,

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1919–1929

2026) given a source sequence $S = (w_1, ..., w_n)$. **DimASR (Regression):** Given S and predefined aspect terms $A = \{a_1, ..., a_k\}$, the system learns a regression mapping $f_{ASR} : (S, A) \rightarrow y_i$, where $y_i = (v_i, a_i) \in [1.00, 9.00]^2$ represents continuous Valence-Arousal (VA) scores. **DimASTE (Triplet Extraction):** The system learns $f_{ASTE} : S \rightarrow T$, extracting triplets $T = \{(a_i, o_i, y_i)\}$, introducing the opinion term o_i that modifies a_i. **DimASQP (Quad Prediction):** The system learns $f_{ASQP} : S \rightarrow Q$, extracting quadruplets $Q = \{(a_i, c_i, o_i, y_i)\}$, integrating the aspect category $c_i \in C$ from a predefined ontology (e.g., `FOOD#QUALITY`).

2.2 Datasets and Prior Work

We utilize the official English laptop (4,076 samples) and restaurant (2,284 samples) datasets (Lee et al., 2026), which feature multi-aspect reviews and continuous VA annotations per aspect. While traditional ABSA (Pontiki et al., 2014) and recent unified extraction methods (Sun et al., 2019) assume categorical labels, DimABSA introduces the novel challenge of continuous sentiment regression (Yu et al., 2026). We address this transition by replacing standard classifiers with distribution-aware regression, retrieval-augmented generation, and contrastive prototype learning. Related work has shown that hybrid and retrieval-augmented architectures are effective for complex NLP pipelines (Kolli et al., 2025; Üyük et al., 2024), while large-scale social media analysis highlights the importance of modeling information dynamics in real-world data (Kolli and Khajeheian, 2018).

3 System Overview

3.1 Pipeline Architecture

We propose a modular cascading framework to address the three DimABSA subtasks (Yu et al., 2026). Each subtask is modeled with a dedicated component tailored to its objective, while intermediate representations are reused when beneficial. Subtask 2 (triplet extraction) extracts aspect–opinion pairs and stets VA scores to 0#0, Subtask 1 (VA regression) provides a standalone regression module predicting VA, and Subtask 3 (category prediction) extends extracted triplets with aspect categories. Our design separates continuous regression, generative extraction, and discrete category prediction to allow task-specific optimization while maintaining a coherent pipeline.

3.2 Subtask 1: Dimensional Aspect Sentiment Regression

We formulate valence and arousal prediction as a distribution learning problem rather than direct scalar regression. Given an input sequence, we format the aspect and the surrounding context using the standard RoBERTa tokenizer schema (Liu et al., 2019) (i.e., `<s> Aspect </s></s> Sentence </s>`). Instead of relying solely on the `<s>` (CLS) token, a pretrained RoBERTa encoder processes this input, and we concatenate the hidden states from the final four Transformer layers. We apply mean-pooling (Gao et al., 2021) across all non-padded tokens to produce a robust global contextual representation, h_{pool}. Two independent linear prediction heads (W_v and W_a) then project this representation into logits (z^v, z^a) over B discretized bins covering the interval [1,9].

3.2.1 Distributional Regression Objective

Instead of minimizing mean squared error, we model each gold VA score as a Gaussian-shaped label distribution over B bins. The model predicts a discrete distribution via a softmax over the logits, and the final scalar prediction is deterministically computed as the mathematical expectation over the predicted bin centers.

To explicitly penalize scale shifts and capture trend alignment, we combine a Kullback-Leibler divergence objective ($\mathcal{L}_{LDL}$) (Geng, 2016) with the Concordance Correlation Coefficient (CCC) (Trigeorgis et al., 2016), a standard metric in affective computing. (Formal mathematical definitions for both components are provided in Appendix A).

Because early validation indicated Arousal exhibits a higher variance and is significantly harder to predict than Valence, we explicitly up-weight the Arousal penalty in our final composite loss:

$$\mathcal{L} = \sum_{d \in \{v,a\}} w_d \Big(\mathcal{L}_{LDL}^{(d)} + \lambda (1 - \text{CCC}(y^{(d)}, \hat{y}^{(d)})) \Big)$$

where $y^{(d)}$ and $\hat{y}^{(d)}$ are the gold and predicted scalars for dimension d. We empirically set the correlation penalty $\lambda = 0.5$ and the dimension weights to $w_v = 1.0$ and $w_a = 2.0$. This architecture ensures local distributional fidelity while forcing the model to prioritize the more challenging affective dimension.

3.3 Subtask 2: Dimensional Triplet Extraction

We formulate the extraction phase of DimASTE as an instruction-following structured generation task.

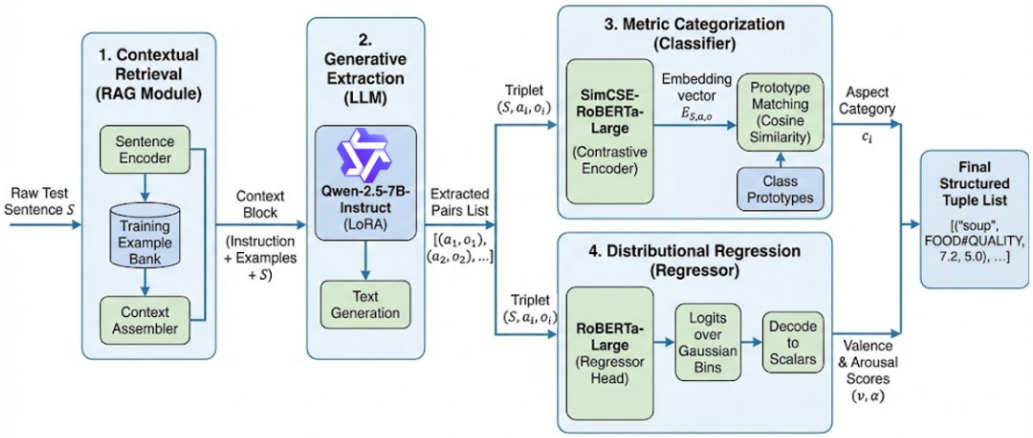

Figure 1: Modular Cascading Pipeline

Given a source sentence, the model is prompted to output intermediate triplets utilizing a placeholder for the continuous values: (Aspect,Opinion,0#0).

We fine-tune an instruction-tuned Large Language Model (LLM) using parameter-efficient LoRA adaptation (Hu et al., 2021) under 4-bit quantization (QLoRA). Crucially, we intentionally offload the continuous VA prediction to the downstream DimASR regressor. Therefore, during the SFT phase (Qwen et al., 2025), the LLM is explicitly optimized only to strictly extract the categorical aspect and opinion spans from the input text. This decoupled design prevents the generative model from expending representational capacity on numerical regression—a task it inherently struggles with—while ensuring seamless integration with our dedicated Label Distribution Learning (LDL) module.

3.3.1 Retrieval-Augmented In-Context Learning

To improve structural consistency and reduce formatting hallucinations, we apply Retrieval-Augmented In-Context Learning (RAG-ICL) (Lewis et al., 2021; Dong et al., 2024) at inference. For each test sentence, we retrieve k semantically similar training examples using embedding similarity and prepend them to the prompt as demonstrations. This dynamic prompting provides implicit structural guidance and significantly improves the LLM's extraction stability without requiring further weight updates.

3.3.2 Fuzzy Span Grounding

Despite explicit instruction tuning, generative models inherently suffer from span instability, frequently producing boundary hallucinations that do not verbatim match the source text. Because the DimASTE evaluation metrics require exact span extraction, we introduce a post-hoc fuzzy span grounding mechanism.

If a generated span s_g of length n does not perfectly align with a substring in the source sentence S, we perform a dynamic sliding-window search. We extract contiguous candidate substrings s_{sub} from S with lengths ranging from $n-2$ to $n+2$. We then select the grounded span s^* that maximizes the Ratcliff-Obershelp sequence similarity ratio (Gestalt Pattern Matching) (Ratcliff et al., 1988), defined as $s^* = \text{argmax}_{s_{sub} \in S_{window}} \text{Sim}_{RO}(s_g, s_{sub})$, where S_{window} represents the set of extracted sliding-window chunks. Finally, we apply an empirical confidence threshold. If $\text{Sim}_{RO}(s_g, s^*) > 0.85$, the hallucinated prediction is replaced by the exact boundaries of s^*. This deterministic heuristic successfully recovers misaligned boundaries without modifying the underlying LLM weights.

3.4 Subtask 3: Dimensional Quad Prediction

Subtask 3 extends triplets with aspect category prediction. Instead of a softmax classifier, we adopt Supervised Contrastive Learning (SupCon) (Gao et al., 2021) to learn a structured embedding space.

Each instance is explicitly encoded as a unified sequence string (e.g., `<s> Aspect </s> Opinion </s> Sentence </s>`) and mapped to an L_2-normalized embedding z. For a batch of embeddings, we optimize the SupCon loss (Khosla et al., 2021) (the formal mathematical definition is provided in Appendix A) to pull embeddings of the same class together while pushing apart different classes.

3.4.1 Prototype-Based Inference

After training, we compute class centroids $c_k = \frac{1}{|D_k|}\sum_{i \in D_k} z_i$, where D_k contains embeddings of class k. At inference, prediction is performed via nearest centroid: $\hat{k} = \text{argmax}_k z \cdot c_k$.

4 Experimental Setup

4.1 Data and Splits

We utilize the official datasets (Lee et al., 2026) provided by the DimABSA shared task organizers for the `restaurant` and `laptop` domains (both English). To maximize the volume of training instances for the final evaluation phase, we concatenated the official training and development sets into a single consolidated dataset.

From this consolidated data, we dynamically generated our own internal splits, allocating 90% of the instances for model optimization and holding out 10% for internal validation and early stopping. Crucially, to prevent data leakage during this split—a scenario where a single review sentence containing multiple aspects might inadvertently bridge the training and validation sets—we applied a Group Shuffle Split (Pedregosa et al., 2011) grouped by the raw source sentence. The official, unlabelled test set was kept strictly isolated and used exclusively for the final leaderboard submission. No external labeled datasets were incorporated.

4.2 Preprocessing

All sentences are lowercased and tokenized using the tokenizer corresponding to each pretrained backbone. To maintain consistency with our mathematical formulation, Subtask 1 inputs are treated as a sentence-pair regression task by constructing an auxiliary sentence from the aspect term (Sun et al., 2019). formatted using the standard RoBERTa schema (i.e., $x = [$<s>,Aspect,</s>,</s>,Text,</s>$]$).

For Subtasks 2 and 3, instruction-style prompts are constructed to enforce structured output formatting. Extracted spans are post-processed using our fuzzy span grounding heuristic based on the difflib sequence-matcher (Ratcliff et al., 1988) to ensure exact metric matching with the source text.

Valence and arousal scores are discretized into 9 evenly spaced bins covering the interval $[1,9]$ for Label Distribution Learning (Geng, 2016). During inference, final scalar predictions are computed as the mathematical expectation over the predicted bin distributions.

4.3 Model Configurations

Subtask 1 (DimASR): We utilize `roberta-large` (Liu et al., 2019) as the regression encoder. Training is performed using the AdamW optimizer (Loshchilov and Hutter, 2019) with an initial learning rate of $1.5e^{-5}$ and a batch size of 128. To preserve the pre-trained representations in the lower layers of the encoder, we apply layer-wise learning rate decay with a decay factor of 0.9, alongside a standard weight decay of 0.01. Models are trained for up to 20 epochs, with early stopping enforced using a patience of 3 epochs based on development set CCC validation (Trigeorgis et al., 2016). The final objective is a weighted combination of KL divergence (LDL) (Geng, 2016) and CCC loss. To ensure robustness against initialization variance, we apply multi-seed ensembling across three random seeds (100, 42, 2026).

Subtask 2 (DimASTE): Triplet extraction is implemented by fine-tuning a `Qwen2.5-7B-Instruct` (Qwen et al., 2025) LLM using QLoRA (Hu et al., 2021). We set rank $r = 32$, alpha $\alpha = 64$, and optimize using a cosine learning rate scheduler with an initial rate of $1.5e^{-5}$. Early stopping patience is set to 5 validation steps. For dynamic prompt augmentation at inference (RAG-ICL) (Lewis et al., 2021; Dong et al., 2024), we retrieve the top $k = 3$ semantically similar training examples using the `BAAI/bge-base-en-v1.5` (Chen et al., 2024) embedding model.

Subtask 3 (DimASQP): For categorical prediction, we utilize `sup-simcse-roberta-large` as the contrastive encoder (Gao et al., 2021). To mitigate the severe class imbalance observed in the aspect category taxonomy, training employs balanced $P - K$ sampling with $P = 4$ classes and $K = 4$ instances per batch. We optimize the Supervised Contrastive Loss (Khosla et al., 2021) using a temperature of $\tau = 0.07$. Following training, class centroids are computed across the embedding space for nearest-centroid inference (Snell et al., 2017).

Implementation Details: Hyperparameters were optimized based on development set performance. Our pipeline is implemented in PyTorch (Paszke et al., 2019) utilizing HuggingFace Transformers (Wolf et al., 2020), PEFT (LoRA), BitsAndBytes (4-bit quantization), and SentenceTransformers (Reimers and Gurevych, 2019) (RAG retrieval). All models were trained on an NVIDIA RTX 4000 Ada

GPU utilizing mixed-precision (bf16).[1]

5 Results and Analysis

We report performance using the official SemEval-2026 Task 3 (Yu et al., 2026) CodaBench metrics: RMSE for Subtask 1, and continuous F1 (cF1) for Subtasks 2 and 3 (Lee et al., 2026). Tables 1 and 2 summarize the performance of our cascading pipeline.

Tables 1 and 2 summarizes the performance of our cascading pipeline across all three subtasks.

Domain	PCC_V ↑	PCC_A ↑	RMSE_VA ↓
Laptop	0.8531	0.5379	1.2942
Restaurant	0.8870	0.6462	1.3011
System Average	-	-	1.4937
Average	*0.8700*	*0.5920*	*1.2976*

Table 1: DimASR (Subtask 1) regression performance. In comparison to the system paper using GPT-oss 120B Supervised Fine-Tuning (Lee et al., 2026) ↓ indicates lower is better; ↑ indicates higher is better.

Task & Domain	cPrecision	cRecall	cF1
DimASTE (Subtask 2)			
Laptop	0.6373	0.5050	0.5635
Restaurant	0.6855	0.5873	0.6326
System Average	0.5249	0.4737	0.4979
Average	*0.6614*	*0.5461*	*0.5980*
DimASQP (Subtask 3)			
Laptop	0.2867	0.2235	0.2512
Restaurant	0.6082	0.5210	0.5612
System Average	0.3896	0.3547	0.3712
Average	*0.4474*	*0.3723*	*0.4062*

Table 2: System performance on continuous Triplet (DimASTE) and Quadruplet (DimASQP) extraction. In comparison to the system paper using GPT-oss 120B Supervised Fine-Tuning (Lee et al., 2026)

On DimASR (Subtask 1), our LDL-based regressor proved robust against subjective annotation variance, achieving an RMSE of 1.2942 (laptop) and 1.3011 (restaurant), significantly outperforming the massive GPT-OSS 120B Supervised Fine-Tuning (SFT) baseline, which achieved 1.5269 and 1.4605 respectively (Lee et al., 2026). This confirms that explicit distribution modeling is vastly superior to LLM-based scalar generation.

[1]Code and models are publicly available at https://github.com/denislaschenko/HCC_dimABSA/tree/main/src

For DimASTE (Subtask 2), our modular Qwen-7B pipeline achieved an average cF1 of 0.5980. This demonstrates the viability of Retrieval-Augmented In-Context Learning (RAG-ICL) as a resource-efficient alternative to the GPT-OSS 120B baseline established by the task organizers on both the laptop (0.5635 vs. 0.4515) and restaurant (0.6326 vs. 0.5442) domains.(Lee et al., 2026). For DimASQP (Subtask 3), the system achieved an average cF1 of 0.4062. While performance naturally dropped on the densely overlapping `laptop` domain (cF1 0.2512), our contrastive nearest-centroid approach still strictly outperformed the best-performing Llama-3.3 70B (Grattafiori et al., 2024) baseline (cF1 0.2483). Furthermore, on the orthogonal `restaurant` domain, our system achieved a cF1 of 0.5612, surpassing the 70B baseline (0.5048).

5.1 Architectural Merits and Error Analysis

Although our system does not claim state-of-the-art absolute performance across all metrics, the results validates several critical architectural advantages of our decoupled, modular approach over monolithic end-to-end generative models.

Efficiency and Modularity: The official DimABSA baselines indicate that maximizing cF1 often requires fine-tuning models with 70B to 120B parameters. In contrast, our pipeline achieves competitive extraction capabilities using only a 7B-parameter generative model. By decoupling the extraction phase from the continuous sentiment scoring, we entirely bypass the issue of numerical hallucination in LLMs (Huang et al., 2025). The regression is handled by a lightweight, mathematically bounded RoBERTa model, allowing for rapid iteration and deployment on consumer-grade hardware.

Impact of Fuzzy Grounding: Our post-hoc fuzzy grounding proved crucial for bridging the gap between generative LLM outputs and strict extraction metrics. The relatively tight clustering of continuous precision (0.6614) and recall (0.5461) in DimASTE indicates that the heuristic successfully anchored the LLM's inherently flexible generated spans back to the exact textual boundaries of the input sequence.

6 Conclusion

We presented a decoupled pipeline for SemEval-2026 Task 3 DimABSA. Unlike monolithic LLMs, our modular architecture achieves competitive

performance through task-specific optimizations. For DimASR, combining LDL (Geng, 2016) with a CCC objective (Trigeorgis et al., 2016) successfully mitigated annotation noise. For DimASTE/DimASQP, QLoRA (Hu et al., 2021) paired with RAG-ICL (Lewis et al., 2021; Dong et al., 2024) on a 7B model proved a resource-efficient alternative to 120B-parameter baselines. Finally, post-hoc fuzzy grounding bridged the gap between generative flexibility and strict metrics. This work highlights that neuro-symbolic modularity and targeted representations can rival brute-force scaling in complex NLP tasks.

Limitations

While our modular pipeline provides a resource-efficient approach to Dimensional ABSA, it presents several inherent limitations. First, as a decoupled cascading system, it is susceptible to error propagation. If the generative Qwen-7B module fails to extract a valid aspect-opinion span during the initial phase, the downstream DimASR regressor cannot recover the missing instance, capping the maximum achievable recall.

Second, our Supervised Contrastive Learning (SupCon) (Khosla et al., 2021) module combined with nearest-centroid inference struggles with dense, long-tailed ontologies. While highly effective for orthogonal domains like `restaurant`, the system exhibits significant performance degradation in domains with heavily overlapping technical attributes, such as the `laptop` domain. Future iterations would require hierarchical contrastive boundaries to resolve these fine-grained semantic overlaps.

Finally, our system was developed and evaluated exclusively on the English (`eng`) track of the DimABSA dataset (Lee et al., 2026). Because our fuzzy span grounding heuristic relies on token-level string matching, its efficacy on morphologically rich or zero-space languages (e.g., Chinese or Japanese) remains untested and would likely require language-specific tokenizer adaptations. We aim to address these limitations in our future work.

Acknowledgments

The authors would like to thank the anonymous reviewers and task organizers. We thank the TUM and Prof. Jana Diesner and especially our mentor Shaghayegh Kolli.

References

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xin Geng. 2016. Label distribution learning. *Preprint*, arXiv:1408.6027.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2021. Supervised contrastive learning. *Preprint*, arXiv:2004.11362.

Shaghayegh Kolli and Datis Khajeheian. 2018. Social network analysis of pokemon go in twitter. In *2018 2nd National and 1st International Digital Games Research Conference: Trends, Technologies, and Applications (DGRC)*, pages 17–26.

Shaghayegh Kolli, Richard Rosenbaum, Timo Cavelius, Lasse Strothe, Andrii Lata, and Jana Diesner. 2025. Hybrid fact-checking that integrates knowledge graphs, large language models, and search-based

retrieval agents improves interpretable claim verification. In *Proceedings of the 9th Widening NLP Workshop*, pages 106–115, Suzhou, China. Association for Computational Linguistics.

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. Mohammed. 2026. Dimabsa: Building multilingual and multidomain datasets for dimensional aspect-based sentiment analysis. *Preprint*, arXiv:2601.23022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Preprint*, arXiv:1912.01703.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

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.

Qwen, :, 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, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

John W Ratcliff, David E Metzener, and 1 others. 1988. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, 13(7):46.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

J. A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. *Preprint*, arXiv:1703.05175.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. 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 380–385, Minneapolis, Minnesota. Association for Computational Linguistics.

George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204.

Cem Üyük, Danica Rovó, Shaghayeghkolli, Rabia Varol, Georg Groh, and Daryna Dementieva. 2024. Crafting tomorrow's headlines: Neural news generation and detection in English, Turkish, Hungarian, and Persian. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 271–307, Miami, Florida, USA. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *Preprint*, arXiv:2203.01054.

A Formal Mathematical Definitions

While the main text provides a high-level overview of our architecture, we include the formal mathematical definitions of our core objective functions here to ensure full reproducibility and to justify our deviation from standard mean squared error (MSE) regression.

A.1 Label Distribution and CCC Loss (DimASR)

Standard continuous sentiment analysis often treats target variables as absolute scalars, optimizing via MSE. However, human affect—especially in subjective domains like restaurant reviews—exhibits intrinsic variance. Therefore, we model the gold scalar Valence-Arousal (VA) scores as a Gaussian distribution over $B = 9$ discrete bins, effectively shifting the task from scalar regression to distribution learning.

Let $\mathbf{p}$ be the gold label distribution and $\hat{\mathbf{p}}$ be the model's predicted softmax probability distribution over the bins. We optimize the Label Distribution Learning (LDL) loss using Kullback-Leibler (KL) divergence (Geng, 2016), which forces the model to capture the variance and uncertainty of the target rather than just its mean:

$$\mathcal{L}_{LDL} = \sum_{b=1}^{B} p_b \log\left(\frac{p_b}{\hat{p}_b}\right) \tag{1}$$

While LDL captures local distributional shape, it does not inherently prevent global scaling shifts or penalize inverse correlations. To enforce structural trend alignment between predictions and ground truth, we combine LDL with the Concordance Correlation Coefficient (CCC) (Trigeorgis et al., 2016), a standard reliability metric in affective computing. Let y and $\hat{y}$ denote the ground-truth and predicted scalars (derived via mathematical expectation over the bins), with means $\mu_y, \mu_{\hat{y}}$ and variances $\sigma_y^2, \sigma_{\hat{y}}^2$. The CCC is defined as:

$$\text{CCC} = \frac{2\rho\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2} \tag{2}$$

where ρ is the Pearson correlation coefficient. By explicitly penalizing deviations from the 45-degree line of perfect concordance (via the squared difference of the means), the CCC objective directly mitigates the "regression to the mean" effect common in subjective NLP tasks.

A.2 Supervised Contrastive Loss (DimASQP)

For Subtask 3 (DimASQP), our system must categorize extracted aspects into predefined ontologies (e.g., `FOOD#QUALITY`). We observed severe class imbalance and semantic overlap in the training data, rendering standard cross-entropy classifiers brittle. To address this, we implement Supervised Contrastive Learning (SupCon) (Khosla et al., 2021).

For a batch of L_2-normalized embeddings $\{\mathbf{z}_i\}$ with labels $\{y_i\}$, the SupCon loss pulls embeddings of the same semantic class together while forcefully pushing apart representations of different classes:

$$\mathcal{L}_{SupCon} = \sum_i \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \neq i} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)} \tag{3}$$

where $P(i)$ denotes the set of indices of all positive samples in the batch for instance i, and τ is a scalar temperature hyperparameter. This structurally enforces a margin of separation, allowing for highly robust nearest-centroid inference even for underrepresented classes.

B Evaluation Metrics

To facilitate direct comparison with future work, we strictly follow the official SemEval-2026 Task 3 guidelines (Yu et al., 2026) for all reported evaluations. We detail the continuous formulation of these metrics below.

B.1 Root Mean Square Error (DimASR)

Subtask 1 evaluates the pure regression capability of the pipeline using a joint Root Mean Square Error ($RMSE_{VA}$) across both affective dimensions. Where N is the total number of instances, $V_p^{(i)}, A_p^{(i)}$ denote the predicted values, and $V_g^{(i)}, A_g^{(i)}$ denote the gold values:

$$RMSE_{VA} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left((V_p^{(i)} - V_g^{(i)})^2 + (A_p^{(i)} - A_g^{(i)})^2\right)} \tag{4}$$

B.2 Continuous F1 (DimASTE & DimASQP)

Standard ABSA metrics rely on binary strict-matching (a triplet is either perfectly extracted or entirely wrong). Because DimABSA introduces

continuous targets, Subtasks 2 and 3 utilize a continuous F1 (cF1) score. This metric softly penalizes numerical deviations while strictly enforcing exact categorical span matching.

First, a normalized Euclidean distance is calculated between the predicted continuous tuple VA_p and the gold tuple VA_g:

$$dist(VA_p, VA_g) = \frac{\sqrt{(V_p - V_g)^2 + (A_p - A_g)^2}}{D_{max}} \tag{5}$$

where $D_{max} = \sqrt{8^2+8^2} = \sqrt{128}$ represents the maximum possible theoretical error distance in the bounded [1,9] Valence-Arousal coordinate space.

The standard True Positive is then relaxed into a continuous true positive (cTP) for a given extraction t:

$$cTP^{(t)} = \begin{cases} 1 - dist(VA_p^{(t)}, VA_g^{(t)}), & t \in P_{cat} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where P_{cat} denotes predictions where all textual and categorical elements (Aspect, Opinion, and optionally Category) match the gold annotation exactly. Continuous Precision (cPrec) and Recall (cRec) are calculated by dividing the summed cTP by the total number of predicted and gold tuples, respectively, resulting in the final harmonic mean (cF1).

C Ethical Considerations

While Dimensional Aspect-Based Sentiment Analysis (DimABSA) offers fine-grained insights into user opinions, it is critical to recognize the epistemological limits of text-based emotion recognition. As noted by the shared task organizers, individuals express attitudes towards entities in highly complex and idiosyncratic ways (Yu et al., 2026). The annotations used to train our models inherently capture *perceived* sentiment as interpreted by human annotators, which may differ significantly from the true, internal psychological state of the original author.

Because language serves as a shared mechanism for communication, perceived opinions correlate reliably with actual opinions at an aggregate level (e.g., evaluating general product reception across thousands of reviews). However, severe ethical risks arise if this architecture is applied at the micro-level. Because our DimASR module outputs precise, real-valued numbers (e.g., Valence: 2.88), there is an inherent risk of "false objectivity"—downstream users or automated decision systems might misinterpret these numerical predictions as definitive psychological metrics. We explicitly warn against deploying this architecture in sensitive, individual-facing domains (such as employment screening, judicial sentencing, or mental health triage) where the gap between perceived text and true human emotion could lead to tangible harm.

D Detailed Input and Output Formulations

Given the multi-stage, hierarchical nature of the DimABSA dataset, data formatting is a non-trivial component of our pipeline. We explicitly decouple the numerical prediction from the textual extraction. To clarify how data flows through our cascading modules, the following is a complete JSON input-output formulation for the "Thai food" restaurant review referenced in Section 1.

Subtask 1: Dimensional Aspect Sentiment Regression (DimASR)

```
Input:
{
  "ID": "R001",
  "Text": "average to good thai food,
   but terrible delivery.",
  "Aspect": ["thai food", "delivery"]
}
Output:
{
  "ID": "R001",
  "Aspect_VA": [
    {"Aspect": "thai food", "VA": "6.75#6.38"},
    {"Aspect": "delivery", "VA": "2.88#6.62"}
  ]
}
```

Subtask 2: Dimensional Aspect Sentiment Triplet Extraction

```
Input:
{
  "ID": "R001",
  "Text": "average to good thai food,
   but terrible delivery."
}
Output:
{
  "ID": "R001",
  "Triplet": [
    {
      "Aspect": "thai food",
      "Opinion": "average to good",
      "VA": "6.75#6.38"
    },
    {
      "Aspect": "delivery",
      "Opinion": "terrible",
      "VA": "2.88#6.62"
    }
  ]
}
```

Subtask 3: Dimensional Aspect Sentiment Quad Prediction

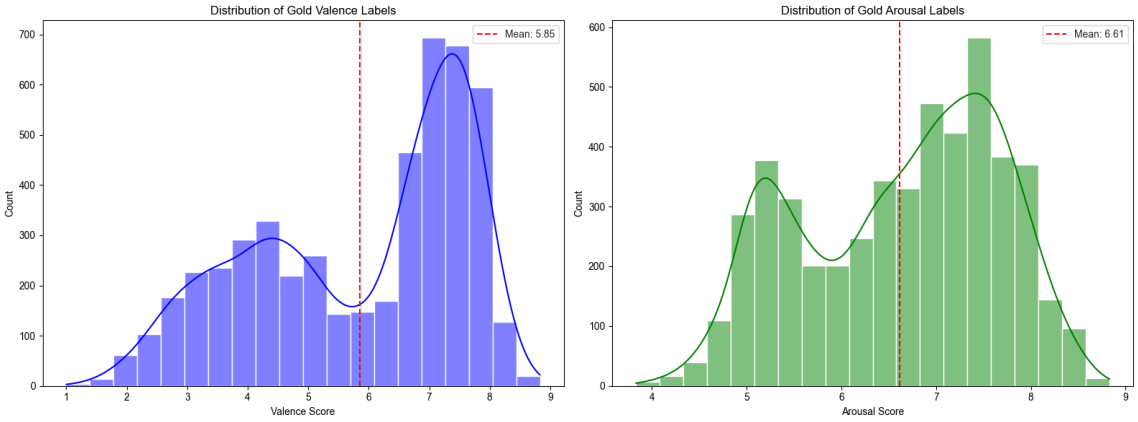

Figure 2: Distribution of gold Valence and Arousal labels across the consolidated training set. The complex, bimodal nature of Valence provides direct empirical justification for our distribution-aware modeling approach over standard scalar regression.

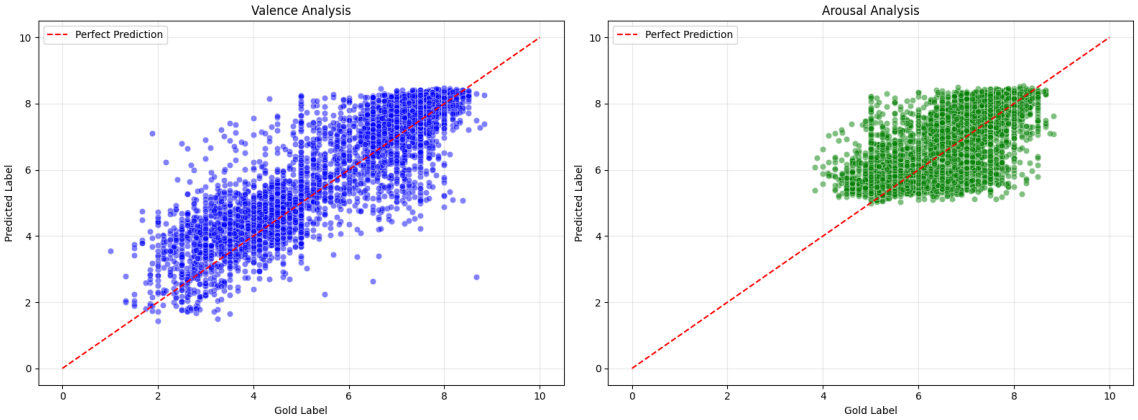

Figure 3: Predicted vs. Gold labels for Valence and Arousal. The red dashed line indicates perfect CCC concordance.

```
Input:
{
  "ID": "R001",
  "Text": "average to good thai food, but terrible
   delivery."
}
Output:
{
  "ID": "R001",
  "Quadruplet": [
    {
      "Aspect": "thai food",
      "Category": "FOOD#QUALITY",
      "Opinion": "average to good",
      "VA": "6.75#6.38"
    },
    {
      "Aspect": "delivery",
      "Category": "SERVICE#GENERAL",
      "Opinion": "terrible",
      "VA": "2.88#6.62"
    }
  ]
}
```

E Instruction-Tuning Prompt Template

For Subtask 2 (DimASTE) (Yu et al., 2026), we observed that base generative models frequently hallucinate boundaries or generate 'NULL' values when unsure. We designed a strict, rule-based prompt template to enforce the extraction of aspect-opinion pairs. Crucially, we offload numerical regression to the DimASR module, instructing the LLM only to format the output.

During our Retrieval-Augmented In-Context Learning (RAG-ICL) (Lewis et al., 2021; Dong et al., 2024) inference phase, up to $k = 3$ semantically similar training examples were prepended to this template to provide dynamic few-shot guidance. The base system prompt is structured as follows:

```
"""Given a textual instance [Text],
extract all (A, O, VA) triplets, where:
- A is an Aspect term (a phrase
describing an entity mentioned in [Text])
- O is an Opinion term
- VA is a Valence-Arousal score in the
format (valence#arousal)
```

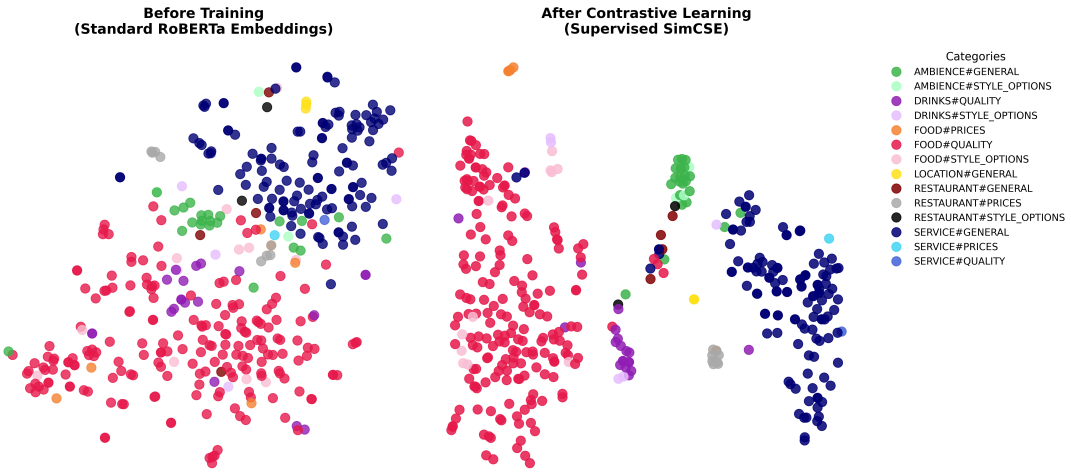

Figure 4: t-SNE visualization of the SupCon embedding space for the Restaurant domain. Distinct clustering of distinct categories enables highly robust, parameter-free nearest-centroid inference.

```
RULES:
1. EXTRACT EXACTLY: Copy the substring
exactly as it appears in the text,
including typos and capitalization.
2. NO NULLS: Every triplet must have
explicit Aspects and Opinions from the
text. Do not generate 'NULL'.
3. NO HALLUCINATIONS: If a word is not in
the text, do not extract it.
4. VA FORMAT: Predict 0#0 for all aspects.

Be exhaustive: extract every aspect and
opinion mentioned. You must preserve the
EXACT capitalization and spelling as it
appears in the [Text]."""
```

F Data Distribution and Regression Error Analysis

A core claim of our methodology is that continuous subjective sentiment cannot be accurately modeled with simple scalar regression due to inherent annotation variance. Figure 2 empirically validates this claim. The distribution of gold Valence annotations is highly non-normal and heavily bimodal, representing polarizing opinions. Standard Mean Squared Error (MSE) models struggle severely with bimodal targets, often predicting the "dead center" between the peaks. Our Label Distribution Learning (LDL) approach (Geng, 2016) successfully models these complex distribution shapes.

Furthermore, Figure 3 visualizes the residuals of our DimASR module, plotting predicted versus gold labels. The scatter plots offer a visual explanation for the performance discrepancies observed in our results table. While Valence predictions tightly hug the 45-degree perfect-concordance line, Arousal exhibits a noticeable "regression to the mean" effect at the extreme bounds. This visually confirms that predicting human emotional energy (Arousal) from text alone remains a significantly harder task than predicting polarity (Valence), necessitating the heavier loss weighting we applied to the Arousal dimension.

G Embedding Space Visualizations

Our decision to bypass standard classification heads in favor of Supervised Contrastive Learning (SupCon) (Khosla et al., 2021) for Subtask 3 (DimASQP) was driven by the need for semantic robust boundaries. Figure 4 visualizes the learned embedding space for the restaurant domain using t-SNE (Gao et al., 2021).

The visualization confirms our architectural hypothesis: the SupCon objective successfully forces orthogonal categories (such as `FOOD#QUALITY` vs. `SERVICE#GENERAL`) into distinct, tightly packed clusters. This clear margin of separation is what enables our computationally lightweight nearest-centroid matching inference to achieve high continuous F1 (cF1) performance, efficiently overcoming the task's inherent category imbalance.