

LOGSIGMA at SemEval-2026 Task 3: Uncertainty-Weighted Multitask Learning for Dimensional Aspect-Based Sentiment Analysis

Baraa Hikal¹, Jonas Becker^{1,2,*}, Bela Gipp^{1,*}

¹University of Göttingen, Germany; ²LKA NRW, Germany, * Advisory Role
{baraa.hikal, jonas.becker, gipp}@uni-goettingen.de

Abstract

This paper describes LOGSIGMA¹, our system for SemEval-2026 Task 3: Dimensional Aspect-Based Sentiment Analysis (DimABSA). Unlike traditional Aspect-Based Sentiment Analysis (ABSA) that predicts discrete labels, DimABSA requires predicting continuous Valence-Arousal (VA) scores on a 1–9 scale. A key challenge is that Valence and Arousal exhibit different prediction difficulties across languages and domains. We address this through learned homoscedastic uncertainty (Cipolla et al., 2018), where the model learns task-specific log-variance parameters ($\log \sigma^2$) that automatically balance each regression objective. Combined with language-specific encoders and multi-seed ensembling, LOGSIGMA achieves 1st place on five datasets across both tracks. The learned variance weights vary substantially across languages due to differing Valence-Arousal difficulty profiles—from $0.66\times$ for German to $2.18\times$ for English—demonstrating that optimal task balancing is language-dependent and cannot be determined *a priori*.

1 Introduction

The SemEval-2026 Task 3 introduces Dimensional Aspect-Based Sentiment Analysis (DimABSA) (Yu et al., 2026), which reformulates traditional aspect-based sentiment analysis as a regression problem (Lee et al., 2026). Instead of predicting categorical sentiment polarity (positive, negative, neutral) (Wilson et al., 2009), systems must predict continuous Valence-Arousal (VA) scores on a 1–9 scale (Nicolaou et al., 2011), where Valence captures sentiment polarity and Arousal captures emotional intensity (Mohammad, 2016). This dimensional representation preserves fine-grained distinctions lost in categorical classification (Reed et al., 2016)—for instance, “I love this dish” and “the food is

acceptable” both express positive sentiment but differ substantially in intensity (Diener et al., 1985). The task spans two tracks: Track A covers user reviews across six languages (Lee et al., 2026) (English, Japanese, Russian, Tatar, Ukrainian, Chinese), while Track B addresses stance detection across five languages (Becker et al., 2026) (German, English, Nigerian Pidgin, Swahili, Chinese).

A key challenge in DimABSA is that Valence and Arousal exhibit different prediction difficulties—Arousal is consistently harder across all languages (average Pearson correlation gap of 0.29 across all 15 datasets). Accurate arousal prediction is critical for downstream applications such as crisis detection, brand monitoring, and content moderation, where emotional intensity determines response urgency (Mohammad, 2016). The natural fix—manually tuning per-task loss weights—requires knowing the optimal V/A ratio in advance (Chen et al., 2018a), yet this ratio varies up to $3\times$ across languages (within a single seed) and cannot be estimated without labeled data. Rather than manually tuning per-task loss weights (Jha et al., 2020), We propose LOGSIGMA, a system that learns task-specific log-variance parameters ($\log \sigma^2$) via homoscedastic uncertainty (Cipolla et al., 2018), automatically balancing the contribution of each regression objective. We combine this with language-specific pretrained encoders and multi-seed ensembling.

The learned uncertainty weights encode linguistic structure: related languages converge to near-identical loss landscapes (English and Pidgin within $\Delta \leq 0.001$ despite per-language training), unrelated languages diverge up to $3\times$ within a single seed, and the directional stability of the learned ratio tracks dataset-level V/A difficulty separation (Figure 2). LOGSIGMA achieves 1st place on five datasets across both tracks, and language-specific encoders outperform a single multilingual model by 8.2% on Track A.

¹  [Source Code](#)

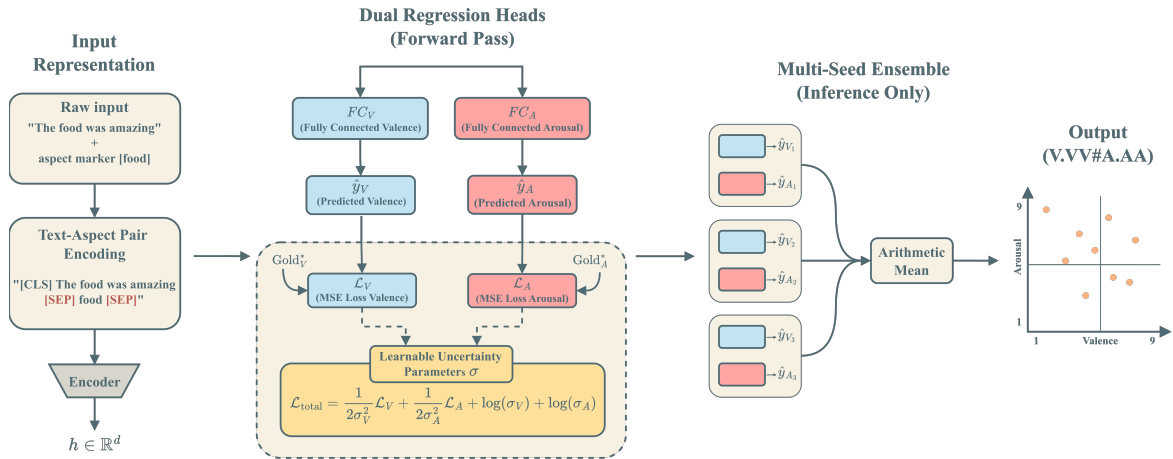


Figure 1: Overview of LOGSIGMA. **Left:** Input text and aspect are encoded as [CLS] text [SEP] aspect [SEP] and passed through a language-specific pretrained encoder to obtain $\mathbf{h} \in \mathbb{R}^d$. **Middle:** Dual fully-connected heads predict Valence and Arousal independently; during training (dashed region), learnable log-variance parameters $s = \log \sigma^2$ balance the MSE losses via precision weighting. **Right:** At inference, predictions from three seeds are averaged to produce the final V.VV#A.AA output.

Our core contributions are:

1. We introduce LOGSIGMA, a multitask learning framework that combines learned homoscedastic uncertainty for automatic V/A loss balancing with language-specific encoders and multi-seed ensembling (§3).
2. We show that the learned σ^2 values act as a zero-cost diagnostic, recovering language similarity, encoder pretraining effects, and dataset heterogeneity without explicit linguistic supervision (§5.3).
3. We achieve 1st place on five datasets across both tracks, with code publicly available. (§5).

2 Background

Task Description. The task comprises two tracks. Track A (DimABSA) includes three subtasks: Subtask 1 (VA regression given text+aspect), Subtask 2 (triplet extraction), and Subtask 3 (quadruplet extraction). Track B (DimStance) includes Subtask 1 only. We participate in Subtask 1 for both tracks, as it isolates the core continuous VA regression challenge.

Track A covers user reviews across four domains and six languages (Lee et al., 2026) (Chinese, English, Japanese, Russian, Tatar, and Ukrainian). Track B covers stance detection across five languages (Becker et al., 2026; Mohammad et al., 2016) (English, German, Chinese, Nigerian Pidgin, and Swahili). Each dataset provides training and

development splits; final evaluation uses a held-out blind test set.

The primary evaluation metric is Root Mean Square Error (RMSE, Equation 1):

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N [(\hat{V}_i - V_i)^2 + (\hat{A}_i - A_i)^2]} \quad (1)$$

where \hat{V}_i and \hat{A}_i denote predictions and lower RMSE indicates better performance. We participate in English for Track A and all five languages for Track B.

Dataset statistics are provided in Tables 4 and 5 in Appendix A. Track A comprises 10 datasets across 6 languages and 4 domains (892–2,507 instances each), while Track B has 5 datasets in stance detection (876–1,245 instances). Representative examples across languages and sentiment profiles are in Table 7 in Appendix C.

Related Work. Dimensional sentiment analysis builds on Russell’s circumplex model (Russell, 1980), representing affect along valence (pleasantness) and arousal (activation) (POSNER et al., 2005). Resources like Chinese EmoBank (Lee et al., 2022; Buechel and Hahn, 2017) provide VA annotations, and the SIGHAN-2024 shared task (Lee et al., 2024) pioneered dimensional ABSA for Chinese. Traditional ABSA focuses on discrete polarity labels (Zhang et al., 2023); DimABSA extends this to continuous dimensions.

For multi-task learning, [Cipolla et al. \(2018\)](#) propose learned homoscedastic uncertainty to balance losses when tasks have different noise levels—we adapt this for VA regression. An alternative is Grad-Norm ([Chen et al., 2018b](#)), which dynamically adjusts gradient magnitudes across tasks; we chose homoscedastic uncertainty for its simplicity, requiring only two scalar parameters rather than per-step gradient monitoring. For regression from text, [Akhaouri et al. \(2025\)](#) propose a generative text-to-text paradigm; we found discriminative fine-tuning substantially more effective for VA prediction (Section 5).

Cross-lingual sentiment has benefited from multilingual transformers ([Conneau et al., 2020](#); [Barbieri et al., 2022](#)), though language-specific models often outperform on tasks requiring cultural nuance ([Zhang et al., 2024](#)), consistent with our findings.

3 System Overview

Figure 1 illustrates the LOGSIGMA architecture, comprising three stages: (1) input representation via language-specific encoders, (2) dual regression heads with learned uncertainty weighting during training, and (3) multi-seed ensemble averaging at inference.

3.1 Input Representation

Given input text and an aspect term, we construct a text-aspect pair encoded as [CLS] text [SEP] aspect [SEP], following [Sun et al. \(2019\)](#) (left panel of Figure 1). This sequence is passed through a pretrained transformer encoder, producing a hidden representation $\mathbf{h} \in \mathbb{R}^d$ from the [CLS] token.

3.2 Dual Regression Heads with Learned Uncertainty

Two fully-connected layers (FC_V and FC_A) independently project the shared encoder representation to scalar predictions (middle panel of Figure 1):

$$\hat{y}_V = \mathbf{w}_V^\top \mathbf{h} + b_V, \quad \hat{y}_A = \mathbf{w}_A^\top \mathbf{h} + b_A \quad (2)$$

The Problem with Fixed Weights. A naive approach minimizes $\mathcal{L} = \mathcal{L}_V + \mathcal{L}_A$, implicitly assuming equal task difficulty. Our experiments show this is violated: Arousal exhibits higher noise across all languages (avg. Pearson Correlation Coefficient (PCC) gap 0.29; Table 13), and optimal weights are language-dependent, making manual tuning impractical.

Learned Homoscedastic Uncertainty. Following [Cipolla et al. \(2018\)](#), we model the observation noise for each task as a learnable parameter. For a regression task with Gaussian likelihood, the negative log-likelihood is:

$$-\log p(y|\hat{y}, \sigma) \propto \frac{1}{2\sigma^2} \|y - \hat{y}\|^2 + \log \sigma \quad (3)$$

where σ^2 represents homoscedastic (input-independent) observation variance. Applying this to both tasks yields our training objective (dashed region in Figure 1):

$$\mathcal{L} = \frac{1}{2\sigma_V^2} \mathcal{L}_V + \frac{1}{2\sigma_A^2} \mathcal{L}_A + \log \sigma_V + \log \sigma_A \quad (4)$$

The $1/\sigma^2$ coefficients act as precision-based weights: a task with higher learned variance (more noise) contributes *less* to the total loss, preventing noisy gradients from destabilizing training. The $\log \sigma$ regularization terms prevent the trivial solution $\sigma \rightarrow \infty$.

Practical Implementation. For numerical stability, we parameterize $s = \log \sigma^2$ rather than σ directly:

$$\mathcal{L} = \frac{1}{2} e^{-s_V} \mathcal{L}_V + \frac{1}{2} e^{-s_A} \mathcal{L}_A + \frac{s_V + s_A}{2} \quad (5)$$

Parameters s_V and s_A are initialized as $s = \log(u)$, $u \sim \text{Uniform}(0.2, 1.0)$, breaking task symmetry from the start. Both are learned jointly with model weights; the system name LOGSIGMA reflects this parameterization.

3.3 Language-Specific Encoders

Rather than a single multilingual encoder, we select language-specific pretrained models (Table 1). We prioritize sentiment-pretrained models where available (Twitter-RoBERTa ([Liu et al., 2019](#); [Loureiro et al., 2022](#)) for English and Pidgin), use dedicated monolingual models where available (GBERT ([Chan et al., 2020](#)) for German), fall back to related-language encoders for low-resource languages (ruBERT ([Kuratov and Arhipov, 2019](#)) for Russian and Tatar), and use XLM-RoBERTa-Large ([Conneau et al., 2020](#)) only for Swahili, which lacks a monolingual model.

3.4 Multi-Seed Ensemble

For each language, we train the same encoder (Table 1) three times with different random seeds, pro-

| Language | Encoder |
|-----------|---|
| English | twitter-roberta-large-† topic-sentiment-latest |
| Japanese | bert-base-japanese-v3 |
| Russian | rubert-base-cased |
| Tatar | rubert-base-cased‡ |
| Ukrainian | ukr-roberta-base |
| Chinese | chinese-roberta-wwm-ext |
| German | gbert-large |
| Pidgin | twitter-roberta-large-† topic-sentiment-latest |
| Swahili | xlm-roberta-large* |

Table 1: Language-specific encoders. (†) sentiment-pretrained; (‡) related-language proxy; (*) multilingual fallback. This strategy outperforms a single multilingual encoder by 8.2% on Track A and up to 28.5% on Track B (Section 5).

ducing independent predictions ($\hat{y}_{V_i}, \hat{y}_{A_i}$). The final output is the arithmetic mean:

$$\hat{y}_V = \frac{1}{3} \sum_{i=1}^3 \hat{y}_{V_i}, \quad \hat{y}_A = \frac{1}{3} \sum_{i=1}^3 \hat{y}_{A_i} \quad (6)$$

formatted as V.VV#A.AA. This reduces RMSE by 3–5% over even the best individual seed by smoothing variance from random initialization, a well-established variance-reduction strategy (Lakshminarayanan et al., 2017) (Table 21 in Appendix L).

4 Experimental Setup

We fine-tune each encoder independently using AdamW (Loshchilov and Hutter, 2019) with learning rate 2×10^{-5} (except English environmental: 1×10^{-5} , selected via validation), batch size 16 with 4-step gradient accumulation, and early stopping on validation RMSE (patience 3). The log-variance parameters s_V and s_A use a separate learning rate of 5×10^{-2} —approximately 2,500× higher than the encoder—enabling rapid adaptation to task-specific noise levels while the encoder preserves pretrained representations. We train three models per configuration with seeds {21, 99, 42} (Track A) and {42, 12, 73} (Track B), averaging predictions across seeds. All models are trained on a single NVIDIA L4 or L40S GPU (fp16, PyTorch 2.x, Hugging Face Transformers 4.x); Full hyperparameters are in Table 6 in Appendix B; an epochs ablation (Table 14 in Appendix H) informed our choice of early stopping over fixed epoch counts, and a cross-domain transfer study (Table 15 in Appendix I) motivated language-specific fine-tuning over Track A initialization.

| Track | Dataset | System | RMSE ↓ | Rank |
|-------------------|-------------------|------------------|----------------------|----------------------|
| A | eng_laptop | LOGSIGMA | 1.241 | 1 st /33 |
| | | Kimi-K2 Thinking | 2.189 | 32 nd /33 |
| | | Qwen-3 14B | 2.809 | 33 rd /33 |
| | eng_restaurant | LOGSIGMA | 1.104 | 1 st /37 |
| | | Kimi-K2 Thinking | 2.146 | 36 th /37 |
| | | Qwen-3 14B | 2.643 | 37 th /37 |
| B | deu_politics | LOGSIGMA | 1.342 | 1 st /12 |
| | | Mistral-3 14B | 1.591 | 10 th /12 |
| | | mBERT | 2.329 | 12 th /12 |
| | eng_environmental | LOGSIGMA | 1.473 | 1 st /15 |
| | | Mistral-3 14B | 1.643 | 7 th /15 |
| | | mBERT | 2.699 | 15 th /15 |
| | pcm_politics | LOGSIGMA | 1.127 | 2 nd /12 |
| | | Mistral-3 14B | 1.739 | 9 th /12 |
| | | mBERT | 3.215 | 12 th /12 |
| | swa_politics | LOGSIGMA | 1.796 | 1 st /12 |
| | | Mistral-3 14B | 2.299 | 10 th /12 |
| | | mBERT | 2.784 | 12 th /12 |
| zho_environmental | LOGSIGMA | 0.646 | 7 th /14 | |
| | Mistral-3 14B | 0.740 | 12 th /14 | |
| | mBERT | 1.276 | 14 th /14 | |

Table 2: Blind test results with baselines from the dataset papers (Lee et al., 2026; Becker et al., 2026). LOGSIGMA outperforms all baselines across every dataset, ranking 1st on five of seven submitted datasets. See Appendix O for a discussion of the performance gap.

5 Results

5.1 Main Results

LOGSIGMA places 1st on five of seven submitted datasets across both tracks, demonstrating that learned uncertainty weighting with language-specific encoders generalizes from development to blind test (Table 2). On Track A, we submitted predictions for 2 of 10 datasets and ranked 1st on both: English laptop (1st/30 teams) and English restaurant (1st/33 teams). On Track B, we ranked 1st on 3 of 5 datasets—German politics (1st/11), English environmental (1st/14), and Swahili politics (1st/11)—and 2nd on Pidgin politics (2nd/11). Chinese environmental is our weakest result (7th/13), suggesting that our Chinese-RoBERTa encoder, effective for Track A review domains, may not transfer as well to environmental stance text.

5.2 Ablation Study

To isolate the effect of learned uncertainty, we compare the same encoder with and without uncertainty weighting (Table 3). Uncertainty improves RMSE on all 5 same-encoder pairs, with relative RMSE reductions of 0.5–11.2%. The improvement is largest where V/A difficulty is most imbalanced (Pidgin: −11.2% RMSE); Figure 5 in Appendix I shows the Pidgin model progression from multilingual

| Dataset | Encoder | w/o | w/ | Δ |
|-----------------|------------|-------|--------------|----------|
| jpn_hotel (A) | XLM-R-L | 0.672 | 0.619 | -7.9% |
| jpn_finance (A) | XLM-R-L | 0.673 | 0.606 | -9.9% |
| deu_pol (B) | GBERT-L | 0.944 | 0.940 | -0.5% |
| eng_env (B) | Tw-RoBERTa | 1.180 | 1.086 | -8.0% |
| pcm_pol (B) | Tw-RoBERTa | 0.702 | 0.624 | -11.2% |

Table 3: Uncertainty weighting ablation on the dev set (RMSE \downarrow). Each row compares the same encoder with and without learned uncertainty. The “w/” column includes 3-seed ensembling. Gains are largest for Pidgin (-11.2%), where the V/A difficulty imbalance is greatest.

baseline to our final system. Full per-dataset development results are in Tables 8 and 9 in Appendix D;

5.3 Analysis

English–Pidgin sigma convergence. The learned σ^2 values for English and Pidgin converge to within $\Delta \leq 0.001$ across all three seeds (e.g., seed 42: $\sigma_V^2 = 0.91$ for both, $\sigma_A^2 = 0.42$ for both), despite per-language training. This quantitatively confirms that the model perceives these two languages as similarly difficult, consistent with Nigerian Pidgin being an English-based creole (Figure 3 in Appendix E shows all three seeds).

Arousal is universally harder than Valence. Across all 15 datasets, Valence PCC (0.84 avg) consistently exceeds Arousal PCC (0.55 avg), yielding an average gap of +0.29. The gap is largest for Chinese environmental (+0.44) and Russian restaurant (+0.42), and smallest for Chinese restaurant (+0.14); see Table 13 in Appendix G for the full breakdown. This aligns with the psychological observation that valence (positive vs. negative) maps more directly to lexical sentiment, while arousal (intensity) depends on subtler pragmatic cues (Russell, 1980; Warriner et al., 2013).

Learned uncertainty reveals language-specific difficulty. Figure 2 reveals that German is the only language where $\sigma_A^2 > \sigma_V^2$ across all seeds, meaning the model down-weights arousal for German political text. One reason could be the source origin of the German data, which consists of texts by political parties that might publish carefully crafted responses. This contrasts with other languages, such as Nigerian Pidgin, whose source contains emotionally charged political discussions among voters, yielding higher valence noise ($\sigma_V^2 > \sigma_A^2$). Chinese shows the highest cross-seed instability, with no consistent V/A

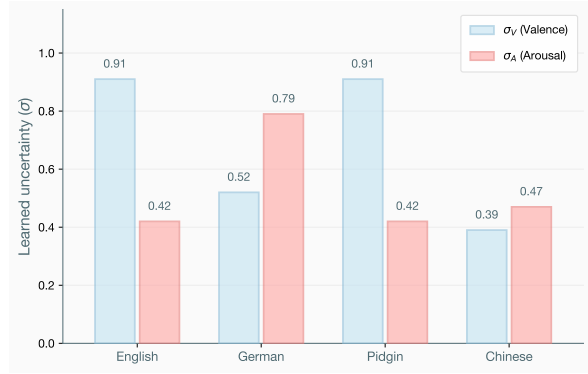


Figure 2: Learned task variance (σ^2) across Track B languages (seed 42). Higher σ^2 indicates greater task-specific noise, causing the model to down-weight that objective. English and Pidgin show $\sigma_V^2 > \sigma_A^2$ (arousal receives higher weight), while German shows the opposite ($\sigma_A^2 > \sigma_V^2$). English and Pidgin learn near-identical values despite per-language training, confirming their shared V/A difficulty profile.

weighting direction across seeds. Full per-seed σ^2 values are in Appendix E; the Track A extension shows that whether an encoder is sentiment-pretrained or general-purpose—not language family or review domain—is the primary driver of learned V/A difficulty ratios. Additional experiments including negative results, encoder comparisons, and post-competition ablations are in Appendix F–Appendix N.

6 Conclusion

We presented LOGSIGMA, a system that learns task-specific log-variance parameters to automatically balance Valence and Arousal regression losses for dimensional aspect-based sentiment analysis; combined with language-specific encoders and multi-seed ensembling, it ranks 1st on five of seven submitted datasets across both tracks. Optimal V/A task weighting varies up to $3\times$ across languages, and sigma profiles converge across related languages (English–Pidgin) and languages sharing the same pretraining regime (Japanese–Russian)—showing that encoder pretraining type is the primary driver of learned V/A difficulty, not language family or review domain. These findings confirm that adaptive loss balancing discovers language-specific difficulty without manual intervention and doubles as a reproducible dataset-level diagnostic probe. Future work should explore heteroscedastic uncertainty, σ^2 transfer to reduce cold-start cost for new languages, and stronger encoders for low-resource settings such as Tatar and Swahili.

Acknowledgements

This work was partially supported by the Lower Saxony Ministry of Science and Culture and the VW Foundation.

References

- Yash Akhauri, Bryan Lewandowski, Cheng-Hsi Lin, Adrian N. Reyes, Grant C. Forbes, Arissa Wongpanich, Bangding Yang, Mohamed S. Abdelfattah, Sagi Perel, and Xingyou Song. 2025. [Performance prediction for large systems via text-to-text regression](#). *Preprint*, arXiv:2506.21718.
- Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. [XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.
- Jonas Becker, Liang-Chih Yu, Shamsuddeen Hassan Muhammad, Jan Philip Wahle, Terry Ruas, Idris Abdulummin, 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.
- 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.
- Branden Chan, Stefan Schweter, and Timo Möller. 2020. [German’s next language model](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. 2018a. [Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks](#). In *International conference on machine learning*, pages 794–803. PMLR.
- Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. 2018b. [GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks](#). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 794–803. PMLR.
- Roberto Cipolla, Yarin Gal, and Alex Kendall. 2018. [Multi-task learning using uncertainty to weigh losses for scene geometry and semantics](#). In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. [Deepseek-v3.2: Pushing the frontier of open large language models](#). *Preprint*, arXiv:2512.02556.
- Ed Diener, Randy J Larsen, Steven Levine, and Robert A Emmons. 1985. Intensity and frequency: dimensions underlying positive and negative affect. *Journal of personality and social psychology*, 48(5):1253.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing](#). In *The Eleventh International Conference on Learning Representations*.
- Ankit Jha, Awanish Kumar, Biplab Banerjee, and Subhasis Chaudhuri. 2020. [Adamt-net: An adaptive weight learning based multi-task learning model for scene understanding](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 706–707.
- Yuri Kuratov and Mikhail Arkipov. 2019. [Adaptation of deep bidirectional multilingual transformers for russian language](#). *Preprint*, arXiv:1905.07213.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. [Simple and scalable predictive uncertainty estimation using deep ensembles](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- 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):1–18.
- 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.

- Lung-Hao Lee, Liang-Chih Yu, Suge Wang, and Jian Liao. 2024. [Overview of the SIGHAN 2024 shared task for Chinese dimensional aspect-based sentiment analysis](#). In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 165–174, Bangkok, Thailand. Association for Computational Linguistics.
- 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.
- Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. [TimeLMs: Diachronic language models from Twitter](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.
- Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. [SemEval-2016 task 6: Detecting stance in tweets](#). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.
- Saif M Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237. Elsevier.
- Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105.
- JONATHAN POSNER, JAMES A. RUSSELL, and BRADLEY S. PETERSON. 2005. [The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology](#). *Development and Psychopathology*, 17(3):715–734.
- Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 49–58.
- James A. Russell. 1980. [A circumplex model of affect](#). *Journal of Personality and Social Psychology*, 39(6):1161–1178.
- 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.
- Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. [Norms of valence, arousal, and dominance for 13,915 English lemmas](#). *Behavior Research Methods*, 45(4):1191–1207.
- Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433.
- 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, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. [Sentiment analysis in the era of large language models: A reality check](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.
- Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2023. [A survey on aspect-based sentiment analysis: Tasks, methods, and challenges](#). *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11019–11038.

A Dataset Statistics

| Language | Domain | Size |
|-----------|------------|-------|
| English | Restaurant | 2,507 |
| | Laptop | 1,869 |
| Japanese | Hotel | 1,892 |
| | Finance | 1,245 |
| Russian | Restaurant | 1,987 |
| Tatar | Restaurant | 892 |
| Ukrainian | Restaurant | 1,124 |
| Chinese | Restaurant | 2,156 |
| | Laptop | 1,678 |
| | Finance | 1,432 |

Table 4: Track A dataset statistics (6 languages, 4 domains, 10 datasets). Sizes are total aspect-level instances across train and dev splits. Tatar is the smallest dataset (892), contributing to its high prediction difficulty.

Table 5 presents the corresponding statistics for Track B, which covers stance detection across five languages.

| Language | Domain | Size |
|-----------------|---------------|-------|
| German | Politics | 1,245 |
| English | Environmental | 987 |
| Nigerian Pidgin | Politics | 1,156 |
| Swahili | Politics | 876 |
| Chinese | Environmental | 1,089 |

Table 5: Track B dataset statistics (5 languages, stance detection). All datasets are smaller than Track A, with Swahili being the most resource-constrained (876 instances).

B Hyperparameters

| Hyperparameter | Value |
|-------------------------------|-----------------------------|
| Optimizer | AdamW |
| Encoder learning rate | 2×10^{-5} |
| Uncertainty lr (s_V, s_A) | 5×10^{-2} |
| Weight decay | 0.01 |
| Batch size / grad. accum. | 16 / 4 (eff. 64) |
| Max epochs (Track A / B) | 25 / 15 |
| Early stopping patience | 3 |
| Dropout | 0.1 |
| Gradient clipping | 1.0 |
| Warmup | 10% of total steps |
| Max seq. length | 128 |
| Seeds (Track A / B) | {21, 99, 42} / {42, 12, 73} |

Table 6: Full training hyperparameters. All languages share the same configuration except English environmental (Track B), which uses a lower encoder learning rate of 1×10^{-5} . The uncertainty parameters receive a $2,500\times$ higher learning rate than the encoder to enable rapid adaptation to task-specific noise levels.

C Dataset Examples

| Language | Domain | Text (excerpt) | Aspect | V | A |
|---------------------------------------|---------------|---|---|------|------|
| <i>Track A — Multi-Domain Reviews</i> | | | | | |
| English | Restaurant | <i>I am so disappointed in the service...</i> | service | 3.38 | 5.75 |
| English | Laptop | <i>I got this laptop and am very impressed...</i> | laptop | 8.00 | 8.00 |
| Japanese | Hotel | 客室のトイレが古すぎて臭かったです... | トイレ (<i>toilet</i>) | 2.62 | 6.50 |
| Japanese | Hotel | 施設のスタッフ...とても丁寧でした | スタッフ (<i>staff</i>) | 8.00 | 7.83 |
| Russian | Restaurant | Сервис вполне хороший. | Сервис (<i>service</i>) | 6.50 | 5.00 |
| Chinese | Restaurant | 真的可以算是...賣像最差的pizza | 賣像 (<i>appearance</i>) | 3.33 | 6.67 |
| <i>Track B — Stance Detection</i> | | | | | |
| German | Politics | <i>Grundversorgung gehört nicht in staatliche Hand.</i> | Grundversorgung (<i>basic supply</i>) | 1.90 | 7.40 |
| Pidgin | Politics | <i>A new Nigeria is coming guys.</i> | Nigeria | 7.23 | 4.57 |
| Swahili | Politics | <i>asante kwa taarifa tumezichukua...</i> | taarifa (<i>information</i>) | 7.03 | 5.70 |
| Chinese | Environmental | 人類活動造成嚴重環境污染... | 人類活動 (<i>human activity</i>) | 3.30 | 6.30 |

Table 7: Representative examples from Track A and Track B datasets. Valence (V) and Arousal (A) are gold-annotated scores on a 1–9 scale. The examples illustrate why V/A imbalance is non-trivial: high arousal can co-occur with either high valence (excited approval) or low valence (angry criticism).

D Full Development Set Results

| Dataset | RMSE ↓ | RMSE _V | RMSE _A | PCC _V / PCC _A |
|----------------|--------------|-------------------|-------------------|-------------------------------------|
| eng_restaurant | 0.754 | 0.742 | 0.766 | 0.919 / 0.684 |
| eng_laptop | 0.660 | 0.600 | 0.719 | 0.934 / 0.611 |
| jpn_hotel | 0.688 | 0.690 | 0.687 | 0.909 / 0.575 |
| jpn_finance | 0.650 | 0.801 | 0.500 | 0.828 / 0.489 |
| rus_restaurant | 1.028 | 0.898 | 1.158 | 0.856 / 0.433 |
| tat_restaurant | 1.285 | 1.434 | 1.136 | 0.519 / 0.330 |
| ukr_restaurant | 1.232 | 1.256 | 1.208 | 0.679 / 0.363 |
| zho_restaurant | 0.532 | 0.582 | 0.482 | 0.848 / 0.711 |
| zho_laptop | 0.579 | 0.619 | 0.539 | 0.861 / 0.667 |
| zho_finance | 0.439 | 0.464 | 0.414 | 0.814 / 0.498 |
| Average | 0.785 | 0.809 | 0.761 | 0.817 / 0.536 |

Table 8: Track A full development set results (language-specific encoders + learned uncertainty + 3-seed ensemble). Chinese datasets are consistently easiest (0.44–0.58 RMSE), while low-resource languages are hardest: Tatar (1.29) and Ukrainian (1.23). Arousal RMSE exceeds Valence RMSE on 3 of 10 datasets (both English and Russian); the remaining 7 have RMSE_V > RMSE_A, including Japanese finance and all Chinese datasets—see Appendix G for analysis.

Table 9 reports the corresponding Track B results, where each language uses a different encoder.

| Dataset | Encoder | RMSE↓ | RMSE _V | RMSE _A | PCC _V | PCC _A |
|-------------------------|------------------|--------------|-------------------------|-------------------|------------------|------------------|
| deu_politics | GBERT-Large | 0.940 | 0.887 | 0.993 | 0.884 | 0.534 |
| eng_environmental | Tw-RoBERTa-Large | 1.111 | 1.267 | 0.956 | 0.784 | 0.487 |
| pcm_politics | Tw-RoBERTa-Large | 0.624 | 0.567 | 0.681 | 0.977 | 0.746 |
| swa_politics | XLM-R-Large | 1.234 | <i>n/a</i> [†] | | 0.788 | 0.621 |
| zho_environmental | Zh-RoBERTa | 0.443 | 0.471 | 0.415 | 0.898 | 0.463 |
| Average (4 lang) | | 0.779 | 0.798 | 0.761 | 0.886 | 0.557 |

Table 9: Track B full development set results (best configuration per language). All use learned uncertainty with 3-seed ensembling except Swahili (multilingual ensemble, no uncertainty). The 4-language average excludes Swahili for fair comparison. Pidgin achieves the best RMSE (0.62) despite being a low-resource language, benefiting from the English-pretrained encoder. [†]Per-dimension RMSE not available for Swahili’s multilingual ensemble.

E Learned Uncertainty Parameters

| Language | Seed | σ_V^2 | σ_A^2 | σ_V^2/σ_A^2 |
|----------|------|--------------|--------------|-------------------------|
| German | 42 | 0.519 | 0.790 | 0.66 |
| | 12 | 0.800 | 0.944 | 0.85 |
| | 73 | 0.301 | 0.504 | 0.60 |
| English | 42 | 0.912 | 0.418 | 2.18 |
| | 12 | 0.621 | 0.582 | 1.07 |
| | 73 | 0.306 | 0.457 | 0.67 |
| Pidgin | 42 | 0.911 | 0.418 | 2.18 |
| | 12 | 0.620 | 0.581 | 1.07 |
| | 73 | 0.306 | 0.457 | 0.67 |
| Chinese | 42 | 0.394 | 0.471 | 0.84 |
| | 12 | 0.888 | 0.228 | 3.90 |
| | 73 | 0.769 | 0.495 | 1.55 |

Table 10: Full per-seed learned σ^2 values (Track B). Higher σ^2 indicates greater task-specific noise, causing the model to down-weight that objective. German is the only language where $\sigma_A^2 > \sigma_V^2$ across all seeds. English and Pidgin converge to within $\Delta \leq 0.001$ on every seed. Chinese shows the highest cross-seed instability (σ_V^2 : 0.39–0.89), suggesting sensitivity to initialization for this language.

Seed Stability Analysis

We quantify cross-seed stability by the range of σ_V^2/σ_A^2 across three seeds. A narrow range means the model assigns a consistent relative difficulty to V and A regardless of initialization. A wide range means the learned weighting is sensitive to random seed.

German (stable direction). The ratio stays between 0.60 and 0.85 across all seeds—always below 1. Every seed independently concludes that arousal is harder than valence for German political text. The absolute magnitudes differ (σ_V^2 : 0.30–0.80), but the qualitative weighting direction never changes. This suggests a structurally consistent difficulty signal in the German data.

English and Pidgin (co-variant instability). The ratio ranges from 0.67 to 2.18, and seed 73 crosses the decision boundary (ratio < 1) while seeds 42 and 12 do not. The weightings are therefore not direction-stable across seeds. However, English and Pidgin track each other within $\Delta \leq 0.001$ on every seed despite being trained separately. Whatever initialization-dependent trajectory English takes, Pidgin follows identically—a direct consequence of their linguistic proximity as an English-based creole.

Chinese (high instability). Chinese shows the widest range: ratio 0.84–3.90. Seed 12 assigns $\sigma_V^2 = 0.888$ and $\sigma_A^2 = 0.228$, strongly down-weighting valence; seed 42 produces a near-balanced assignment (0.394 vs. 0.471); seed 73 sits in between (0.769 vs. 0.495). The three seeds therefore optimize under substantially different loss weightings. This likely reflects the linguistic diversity within the Chinese environmental stance dataset, which may span a wider range of registers than the other four languages. Even so, Chinese still achieves the best Track B RMSE (0.44) despite this instability—possibly because the three divergent weighting strategies complement each other in the ensemble.

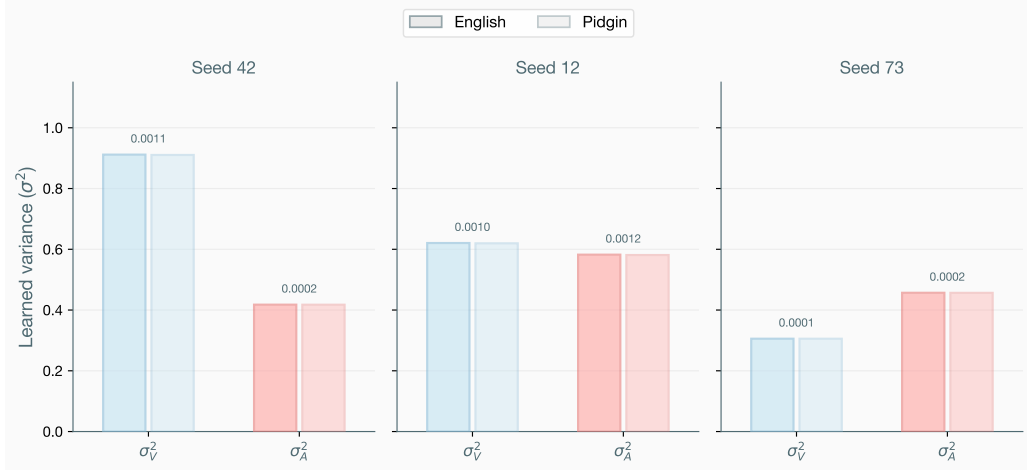


Figure 3: English–Pidgin σ^2 convergence across all three seeds. Despite per-language training on separate datasets, English and Pidgin learn near-identical variance parameters on every seed ($\Delta \leq 0.001$). This quantitatively confirms that the Twitter-RoBERTa encoder perceives these two languages as having identical V/A difficulty profiles, consistent with Nigerian Pidgin being an English-based creole.

Track A Sigma Values

| Language | Seed | σ_V^2 | σ_A^2 | σ_V^2/σ_A^2 |
|----------|------|--------------|--------------|-------------------------|
| English | 21 | 0.615 | 0.288 | 2.13 |
| | 99 | 0.859 | 0.559 | 1.54 |
| | 42 | 0.910 | 0.418 | 2.18 |
| Japanese | 21 | 0.354 | 0.752 | 0.47 |
| | 99 | 0.443 | 0.525 | 0.84 |
| | 42 | 0.392 | 0.469 | 0.84 |
| Russian | 21 | 0.355 | 0.755 | 0.47 |
| | 99 | 0.443 | 0.526 | 0.84 |
| | 42 | 0.394 | 0.472 | 0.84 |
| Tatar | 21 | 0.355 | 0.755 | 0.47 |
| | 99 | 0.443 | 0.527 | 0.84 |
| | 42 | 0.394 | 0.472 | 0.84 |

Table 11: Track A learned σ^2 values for four languages (seeds 21, 99, 42). English uses restaurant values; Japanese uses finance values (hotel is near-identical). English is the only language with $\sigma_V^2 > \sigma_A^2$ across all seeds, and the only Track A language using a sentiment-pretrained encoder (Twitter-RoBERTa). Japanese and Russian converge to near-identical ratios despite different encoders (BERT-Japanese vs. ruBERT), unrelated language families, and different review domains (finance/hotel vs. restaurant)—their shared factor is general-purpose encoder pretraining, suggesting that without sentiment pretraining, arousal is consistently the harder target. Russian and Tatar are trivially identical as they share the same ruBERT encoder. Ukrainian and Chinese values were not logged for this run.

Encoder pretraining type drives divergence. Japanese (BERT-Japanese) and Russian (ruBERT) converge to near-identical σ_V^2/σ_A^2 ratios across all three seeds ($\approx 0.47, 0.84, 0.84$), despite different encoders, unrelated language families (Japonic vs. Slavic), and different review domains (finance/hotel vs. restaurant). Their shared factor is encoder pretraining type: both use general-purpose encoders, and without sentiment pretraining, arousal is consistently the harder target. English is the sole outlier (ratio ≈ 1.95) and the only Track A language using a sentiment-pretrained encoder; Twitter-RoBERTa’s pretraining yields lower valence loss, so the model assigns higher σ_V^2 and concentrates optimization weight on the harder arousal task. These findings refine the Track B English–Pidgin observation: sigma convergence can arise from shared encoder architecture (Russian–Tatar), shared pretraining regime (Japanese–Russian), or shared linguistic and encoder profile (English–Pidgin), and disentangling these factors requires controlled cross-encoder comparisons.

F Encoder Comparison (Track B)

| Dataset | XLM-R-Large | XLM-R-Base | Afro-XLMR | mDeBERTa | XLM-Politics |
|-------------------|--------------|--------------|-----------|----------|--------------|
| deu_politics | 1.143 | 1.389 | 1.161 | 1.580 | 1.449 |
| eng_environmental | 1.383 | 1.523 | 1.418 | 1.628 | 1.602 |
| pcm_politics | 0.985 | 1.300 | 0.941 | 1.431 | 1.284 |
| swa_politics | 1.464 | 1.437 | 1.498 | 1.539 | 1.566 |
| zho_environmental | 0.464 | 0.734 | 0.465 | 0.636 | 0.633 |
| Average | 1.088 | 1.277 | 1.097 | 1.363 | 1.307 |

Table 12: Encoder architecture comparison on Track B development set (RMSE \downarrow , all without uncertainty weighting). XLM-RoBERTa-Large wins on 4 of 5 datasets. mDeBERTa-v3-base is the weakest encoder—near-zero arousal PCC on German (0.13) and English (0.18)—suggesting its pretraining does not transfer to VA regression. XLM-Twitter-Politics, despite being pretrained on political sentiment text, underperforms generic XLM-R-Large. XLM-R-Base wins only on Swahili, likely due to less overfitting on this low-resource language.

The encoder failures are concentrated on the arousal dimension: mDeBERTa and XLM-Twitter-Politics fall to near-zero arousal PCC on German and English while retaining moderate valence correlation (Figure 4), confirming that pretraining quality matters most for the harder regression target. The only exception is Swahili, where XLM-R-Base outperforms XLM-R-Large (1.437 vs. 1.464 RMSE)—likely because its lower capacity overfits less on the smallest dataset (876 instances).

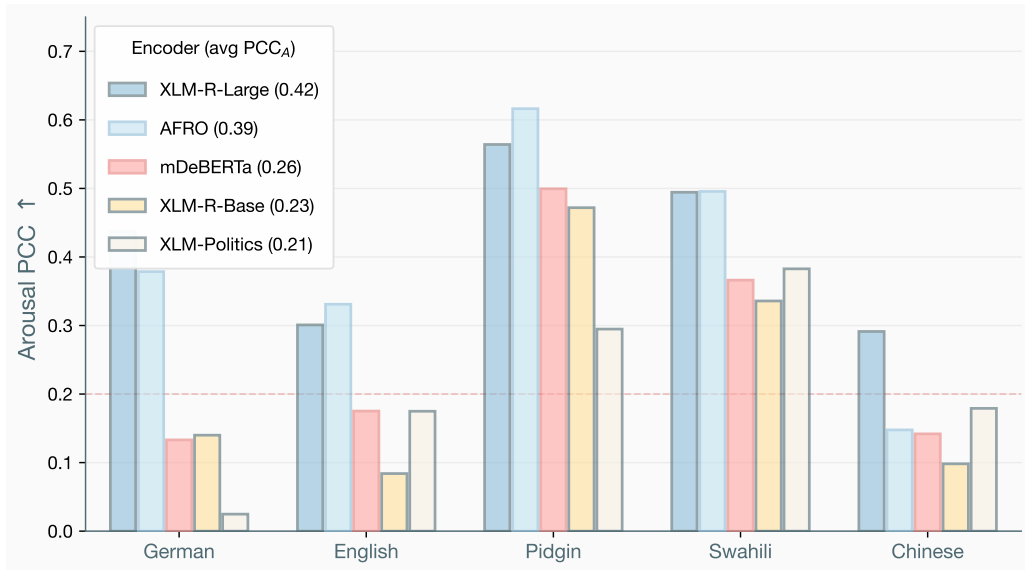


Figure 4: Arousal PCC per encoder across Track B languages. XLM-RoBERTa-Large and Afro-XLMR-Large achieve the highest arousal correlations on most datasets. mDeBERTa and XLM-Twitter-Politics fall below the near-zero threshold (dashed line) on German, demonstrating that their pretraining does not transfer to arousal regression. Pidgin is the only language where all encoders achieve reasonable arousal PCC (>0.29), likely because its English-based lexicon benefits all English-pretrained models.

G Valence vs. Arousal Difficulty Gap

| Dataset | PCC_V | PCC_A | Gap ($V - A$) |
|---------------------|--------------|--------------|-----------------|
| eng_restaurant | 0.919 | 0.684 | +0.235 |
| eng_laptop | 0.934 | 0.611 | +0.323 |
| jpn_hotel | 0.909 | 0.575 | +0.334 |
| jpn_finance | 0.828 | 0.489 | +0.339 |
| rus_restaurant | 0.856 | 0.433 | +0.423 |
| tat_restaurant | 0.519 | 0.330 | +0.189 |
| ukr_restaurant | 0.679 | 0.363 | +0.316 |
| zho_restaurant | 0.848 | 0.711 | +0.137 |
| zho_laptop | 0.861 | 0.667 | +0.194 |
| zho_finance | 0.814 | 0.498 | +0.316 |
| deu_politics | 0.884 | 0.534 | +0.350 |
| eng_environmental | 0.784 | 0.487 | +0.297 |
| pcm_politics | 0.977 | 0.746 | +0.231 |
| swa_politics | 0.788 | 0.621 | +0.167 |
| zho_environmental | 0.898 | 0.463 | +0.435 |
| Average (15) | 0.837 | 0.546 | +0.291 |

Table 13: Valence vs. Arousal prediction difficulty across all 15 datasets. Arousal is harder to predict than Valence in every dataset, with an average PCC gap of +0.29. The gap is largest for Chinese environmental (+0.44) and Russian restaurant (+0.42), and smallest for Chinese restaurant (+0.14). Swahili has the smallest Track B gap (+0.17). However, PCC-based ranking can disagree with RMSE-based ranking: for Japanese finance and all Chinese Track A datasets, $RMSE_V > RMSE_A$ (Table 8), meaning the model makes larger absolute errors on Valence despite ranking Arousal samples less accurately. This discrepancy arises because PCC measures rank correlation while RMSE measures absolute error.

H Epochs Ablation

We compare 5-epoch and 10-epoch training to determine the optimal stopping point per dataset (Table 14).

| Dataset | 5 Epochs | 10 Epochs | Winner |
|----------------|--------------|--------------|--------|
| eng_laptop | 0.750 | 0.687 | 10ep |
| eng_restaurant | 0.834 | 0.745 | 10ep |
| jpn_finance | 0.673 | 0.770 | 5ep |
| jpn_hotel | 0.672 | 0.686 | 5ep |
| rus_restaurant | 1.002 | 0.947 | 10ep |
| tat_restaurant | 1.306 | 1.332 | 5ep |
| ukr_restaurant | 1.018 | 0.967 | 10ep |
| zho_laptop | 0.520 | 0.620 | 5ep |
| zho_restaurant | 0.540 | 0.544 | 5ep |
| zho_finance | 0.373 | 0.474 | 5ep |

Table 14: Epochs ablation on Track A development set (XLM-RoBERTa-Large, RMSE \downarrow). Five epochs wins on 6 of 10 datasets—all Japanese, all Chinese, and Tatar. Ten epochs wins on English, Russian, and Ukrainian. This suggests that CJK and low-resource languages are more prone to overfitting. Our final system uses early stopping (patience 3) rather than a fixed epoch count to adapt per language.

I Transfer and Adaptation Experiments

This section examines three transfer strategies: cross-domain initialization, encoder selection for low-resource languages, and translation-based adaptation.

Cross-domain transfer: Track A \rightarrow Track B. We test whether initializing from a Track A checkpoint (user reviews) improves Track B stance performance. Using XLM-RoBERTa-Large, Track A initialization improves 4 of 5 languages, with the largest gain on Chinese (-12.0% ; Table 15). English environmental is the only dataset where transfer hurts ($+7.8\%$), suggesting that language-specific factors beyond domain mismatch influence transferability.

| Dataset | Random Init | Track A Init | Δ |
|-------------------|--------------|--------------|-----------|
| deu_politics | 1.143 | 1.069 | -6.5% |
| eng_environmental | 1.383 | 1.490 | $+7.8\%$ |
| pcm_politics | 0.985 | 0.901 | -8.5% |
| swa_politics | 1.464 | 1.362 | -6.9% |
| zho_environmental | 0.464 | 0.409 | -12.0% |

Table 15: Cross-domain transfer from Track A to Track B, using XLM-RoBERTa-Large (RMSE \downarrow). Track A initialization improves 4 of 5 languages (up to -12% on Chinese). English environmental is the only dataset where transfer hurts ($+7.8\%$), despite Chinese environmental benefiting—suggesting that language-specific factors beyond domain mismatch influence transferability. This experiment uses a single XLM-R model, not our final language-specific setup.

Encoder selection for Nigerian Pidgin. Nigerian Pidgin (pcm) has no dedicated pretrained model and no Track A training data, making encoder choice critical. Figure 5 traces the full model progression. Switching from multilingual XLM-R to English-pretrained Twitter-RoBERTa accounts for -21.4% RMSE, while adding 3-seed ensembling and learned uncertainty contributes a further -19.4% . The uncertainty ablation isolates its contribution at -11.2% (Table 3), confirming that encoder selection, ensembling, and loss design are complementary.

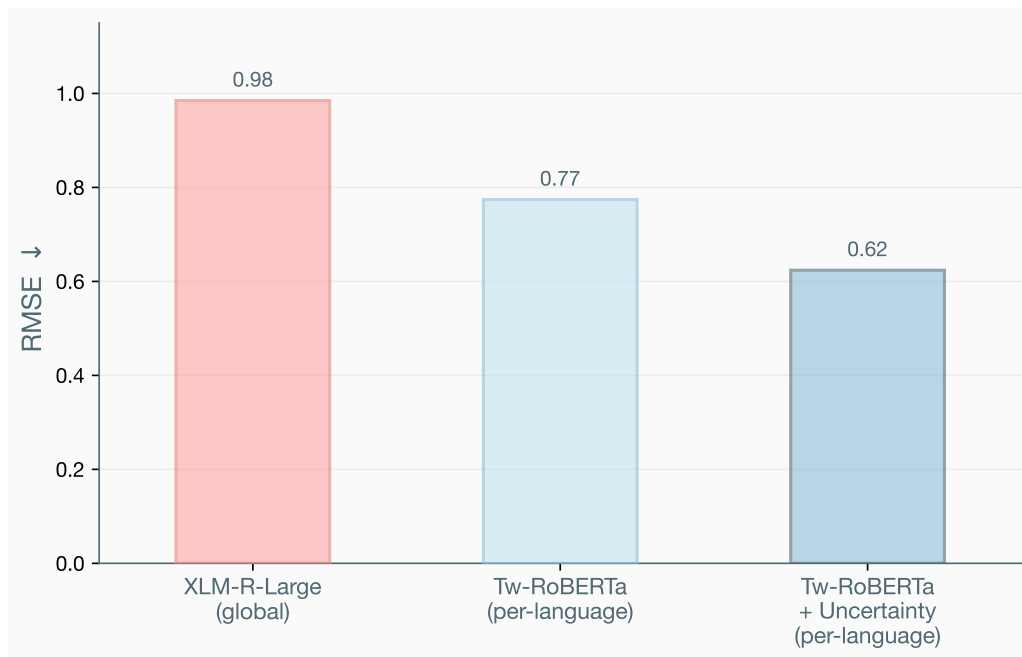


Figure 5: Nigerian Pidgin model progression (RMSE on dev set). Switching from multilingual XLM-R (0.98) to English-pretrained Twitter-RoBERTa yields -21.4% RMSE. Adding 3-seed ensembling and learned uncertainty yields a further -19.4% ($0.77 \rightarrow 0.62$), demonstrating that encoder selection, ensembling, and loss design each contribute substantially.

Translation. Can we bypass language-specific encoders by translating all data to English? We compare a native Russian model against one trained on machine-translated English data for Russian restaurant. Valence PCC is nearly preserved (0.938 translated vs. 0.934 native), but arousal PCC drops from 0.586 to 0.557 (Figure 6). This confirms that arousal is encoded in language-specific features—word choice, emphasis patterns, morphological intensity—that translation strips away, justifying our language-specific encoder design.

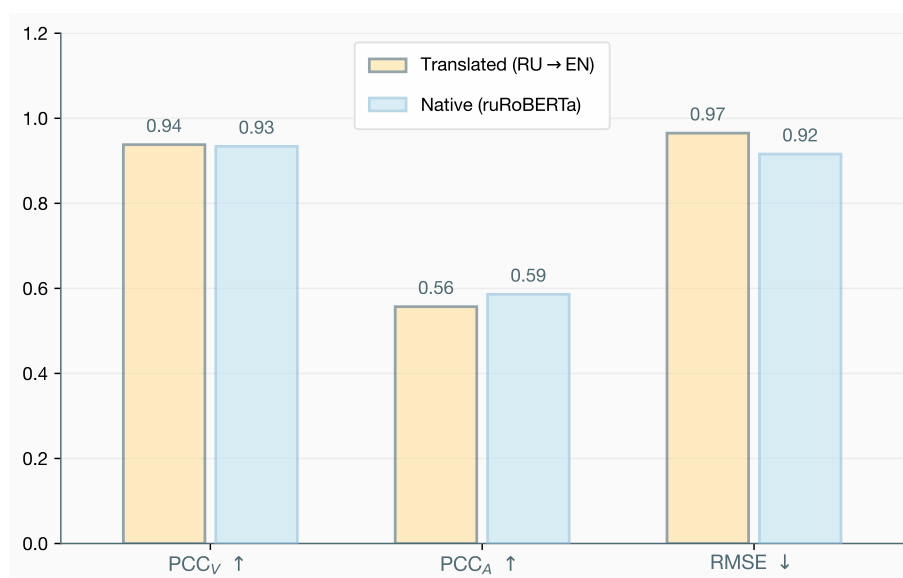


Figure 6: Translation vs. native model for Russian restaurant. Valence PCC is nearly identical (0.938 vs. 0.934), confirming that positive/negative sentiment transfers across languages. However, arousal PCC drops from 0.586 (native) to 0.557 (translated), and overall RMSE increases by 5.4%. This suggests that arousal is encoded in language-specific features—word choice, emphasis patterns, morphological intensity markers—that translation strips away.

J What Did Not Work

Generative regression. To test whether framing VA prediction as sequence generation improves over discriminative models, we evaluate RegressLM (Akhaoui et al., 2025). We compare a pretrained T5-Gemma backbone (60M params, lr = 2e-5, early stopped at epoch 16) against a from-scratch 2-encoder-2-decoder variant (31.2M params, lr = 5e-5, early stopped at epoch 46); both are trained jointly on all languages with MSE loss.

The pretrained T5-Gemma variant achieves 1.48 avg RMSE, nearly double the 0.78 of our discriminative LOGSIGMA (Table 16). Augmenting its inputs with synthetic opinion terms narrows this gap to 1.34 (Table 19 in Appendix K), yet remains 71% behind. The from-scratch variant performs substantially worse: on Japanese, it produces negative PCC, meaning predictions are anti-correlated with gold labels (Table 17). Pretrained initialization alone accounts for a 15.7% RMSE reduction (1.59 \rightarrow 1.34), confirming the importance of transfer learning for low-resource multilingual regression.

| Dataset | RegressLM | LOGSIGMA | Δ |
|----------------|--------------|--------------|----------|
| eng_laptop | 1.438 | 0.660 | -54.1% |
| eng_restaurant | 1.334 | 0.754 | -43.5% |
| jpn_hotel | 1.333 | 0.688 | -48.4% |
| jpn_finance | 1.523 | 0.650 | -57.3% |
| rus_restaurant | 1.873 | 1.028 | -45.1% |
| tat_restaurant | 2.137 | 1.285 | -39.9% |
| ukr_restaurant | 1.736 | 1.232 | -29.0% |
| zho_restaurant | 1.109 | 0.532 | -52.0% |
| zho_laptop | 1.372 | 0.579 | -57.8% |
| zho_finance | 0.922 | 0.439 | -52.4% |
| Average | 1.478 | 0.785 | -46.9% |

Table 16: RegressLM (T5-Gemma, pretrained, 60M params) on Track A dev set (RMSE \downarrow), text + aspect input only. Δ shows the relative RMSE reduction of LOGSIGMA over RegressLM. The opinion-augmented variant is in Table 19 in Appendix K.

The from-scratch variant (Table 17) performs substantially worse, producing negative PCC on Japanese datasets.

| Dataset | PCC _V | PCC _A | RMSE |
|----------------|------------------|------------------|--------------|
| eng_laptop | 0.69 | 0.32 | 1.560 |
| eng_restaurant | 0.78 | 0.56 | 1.294 |
| jpn_finance | 0.03 | -0.03 | 1.610 |
| jpn_hotel | -0.05 | -0.11 | 1.969 |
| rus_restaurant | 0.32 | 0.23 | 1.997 |
| tat_restaurant | 0.41 | 0.24 | 1.951 |
| ukr_restaurant | 0.48 | 0.08 | 1.908 |
| zho_laptop | 0.07 | 0.16 | 1.395 |
| zho_restaurant | 0.12 | 0.16 | 1.355 |
| zho_finance | 0.14 | 0.10 | 0.829 |
| Average | 0.30 | 0.17 | 1.587 |

Table 17: From-scratch RegressLM (2E2D, $d_{\text{model}} = 512$, 31.2M params, lr = 5e-5, early stopped at epoch 46) on Track A dev set. Japanese datasets produce *negative* PCC on both dimensions, meaning predictions are anti-correlated with gold labels. Even the best language (English restaurant, PCC_V = 0.78) falls far short of pretrained encoders (PCC_V = 0.93 with LOGSIGMA).

Alternative encoders. mDeBERTa-v3-base (He et al., 2023) produces near-zero arousal PCC on German (0.13) and English (0.18), and averages 1.36 RMSE—worst among all encoders tested (Table 12 in Appendix F, Figure 4). XLM-Twitter-Politics, despite being pretrained on political sentiment text, underperforms generic XLM-RoBERTa-Large. Domain-specific sentiment pretraining does not transfer to VA regression.

Translation. Translating Russian to English degrades arousal while preserving valence; see Appendix I for the full analysis.

K Post-Competition Experiments

K.1 Synthetic Opinion Generation

Motivation: opinion terms drive VA variance. Subtask 1 provides text and aspect terms, while opinion annotations are available for all domains except the finance domain in two languages. An exploratory analysis of the Track A training data shows that, within the same aspect, opinion terms are the primary source of VA variability (Table 18). Grouping instances by aspect alone yields large within-group variance. Adding opinion to the grouping key substantially reduces this variance: on average, σ decreases by **79.1%** for Valence and **55.1%** for Arousal. An analysis of variance (ANOVA) further confirms this effect: opinion accounts for nearly all residual within-aspect variance, with $\eta_V^2=0.98$ and $\eta_A^2=0.95$. This suggests that VA scores are largely determined by opinion rather than aspect identity.

Illustrative example. As a concrete example, the aspect “*food*” in eng_restaurant spans $V = 1.17$ – 8.75 depending on the expressed opinion, where positive opinions such as “*delicious*” yield high scores (8.00) and negative opinions such as “*terrible*” produce low scores (1.17). Two datasets (jpn_finance, zho_finance) originally lack opinion annotations; for these we use synthetic DeepSeek-generated opinions (§K). Synthetic opinions reduce jpn_finance variance by 85.8%, on par with ground-truth datasets, while zho_finance shows a smaller 47.4% reduction—consistent with the domain shift observed in its RMSE ablation.

| Dataset | Aspect-only σ | | Aspect+Opinion σ | | $\Delta\%$ | |
|--------------------------|----------------------|-------------|-------------------------|-------------|--------------|--------------|
| | V | A | V | A | V | A |
| eng_laptop | 1.28 | 0.85 | 0.33 | 0.41 | −74.5 | −51.6 |
| eng_restaurant | 1.30 | 0.87 | 0.30 | 0.37 | −76.8 | −57.0 |
| jpn_hotel | 0.91 | 0.49 | 0.22 | 0.24 | −76.0 | −50.9 |
| jpn_finance [†] | 0.92 | 0.43 | 0.13 | 0.12 | −85.8 | −71.5 |
| rus_restaurant | 1.34 | 0.80 | 0.16 | 0.30 | −88.2 | −61.8 |
| tat_restaurant | 1.24 | 0.79 | 0.19 | 0.35 | −84.5 | −56.2 |
| ukr_restaurant | 1.28 | 0.78 | 0.18 | 0.33 | −86.1 | −57.7 |
| zho_restaurant | 0.77 | 0.55 | 0.23 | 0.29 | −70.8 | −47.4 |
| zho_laptop | 0.82 | 0.55 | 0.21 | 0.27 | −74.5 | −50.3 |
| zho_finance [†] | 0.39 | 0.31 | 0.21 | 0.19 | −47.4 | −37.9 |
| Average | 1.03 | 0.64 | 0.21 | 0.29 | −79.1 | −55.1 |

Table 18: Within-group standard deviation of VA scores when instances are grouped by aspect alone versus aspect+opinion. Adding opinion reduces σ_V by 47–88% and σ_A by 38–72%, confirming that opinion is the dominant source of within-aspect VA variance ($\eta_V^2=0.98$, $\eta_A^2=0.95$). (†) Synthetic DeepSeek-generated opinions; all other datasets use ground-truth labels.

Method: DeepSeek few-shot pipeline. For the two domains without opinion annotations (jpn_finance, zho_finance), we generate synthetic opinion terms via few-shot prompting with DeepSeek-V3.2 (DeepSeek-AI et al., 2025) (deepseek-chat, temperature 0.1, max 50 tokens). All other datasets use ground-truth opinion labels from the Track A training data.

For each text–aspect pair, the model extracts the exact opinion word or phrase, or outputs NULL if none is expressed. Since the finance domains lack native examples, few-shot prompts draw from related training sets (jpn_hotel and zho_restaurant, respectively), balanced across positive, negative, and null categories. Both languages use native-language system prompts; the English template is shown in Figure 7.

System

You are an expert at Aspect-Based Sentiment Analysis (ABSA). Given a text and an aspect term, extract the exact opinion word or phrase from the text that expresses sentiment about that aspect. The opinion should be a word or short phrase directly from the text expressing positive, negative, or neutral sentiment. If no explicit opinion is expressed, output “NULL”. Output ONLY the opinion word/phrase.

User

Example 1 (positive): Text: {pos_text}
Aspect: {aspect} → Opinion: {pos_opinion}

Example 2 (negative): Text: {neg_text}
Aspect: {aspect} → Opinion: {neg_opinion}

Example 3 (null): Text: {null_text}
Aspect: {aspect} → Opinion: NULL

Now extract the opinion for:
Text: {text} Aspect: {aspect}
Opinion:

Figure 7: English system prompt and few-shot template for synthetic opinion generation with DeepSeek. Each of the 6 languages uses an equivalent native-language prompt. Examples cover positive, negative, and null opinion categories.

Results: opinion ablation. We ablate the impact of including opinion terms—using ground-truth labels for eight datasets and synthetic DeepSeek-generated labels for the two finance domains—using a T5-Gemma baseline (Table 19). Appending these opinions to the input reduces average RMSE by **9.2%** (1.48 → 1.34).

Improvements are consistent across both ground-truth and synthetic subsets, with the largest gains observed in zho_laptop (−25.0%) and rus_restaurant (−17.5%). Synthetic opinions also reduced error in jpn_finance (−9.5%), validating the few-shot extraction strategy for filling annotation gaps.

Performance did not improve in two cases: tat_restaurant and zho_finance. For Tatar, where ground-truth opinions were used, the +3.2% increase suggests that the baseline model struggled to map native-language opinion terms to VA scores. For zho_finance, the +6.5% regression likely stems from domain shift, as few-shot examples were drawn from restaurant data. Although the generative model remains 42% behind LOGSIGMA, the consistent improvements confirm that opinion terms encode meaningful VA signal, motivating future joint modeling.

| Dataset | −Opinion | +Opinion | Δ |
|--------------------------|--------------|--------------|--------|
| eng_laptop | 1.438 | 1.279 | −11.1% |
| eng_restaurant | 1.334 | 1.188 | −10.9% |
| jpn_hotel | 1.333 | 1.224 | −8.2% |
| jpn_finance [‡] | 1.523 | 1.379 | −9.5% |
| rus_restaurant | 1.873 | 1.545 | −17.5% |
| tat_restaurant | 2.137 | 2.205 | +3.2% |
| ukr_restaurant | 1.736 | 1.722 | −0.8% |
| zho_restaurant | 1.109 | 0.859 | −22.5% |
| zho_laptop | 1.372 | 1.029 | −25.0% |
| zho_finance [‡] | 0.922 | 0.982 | +6.5% |
| Average | 1.478 | 1.341 | −9.2% |

Table 19: Opinion ablation with RegressLM (T5-Gemma, pretrained, 60M params) on Track A dev set (RMSE ↓). **−Opinion**: text + aspect only (same baseline as Table 16); **+Opinion**: input includes opinion terms (Ground-Truth for 8 sets; Synthetic for 2 sets marked with ‡). Inclusion of opinion terms reduces average RMSE by 9.2%; see Table 16 for LOGSIGMA comparison.

K.2 LLM Fine-tuning

We evaluate whether scaling to a 4B-parameter LLM improves over our encoder-based system. To probe the upper performance bounds of Track A, we fine-tuned Qwen3-4B with two regression heads (5 epochs, MSE loss). As shown in Table 20, the LLM reduces RMSE by **21.4%** on rus_restaurant compared to LOGSIGMA (0.808 vs. 1.028). This suggests that for high-variance domains, the larger parameter count can better separate overlapping VA distributions.

Despite these gains, we did not scale the LLM approach across all 15 datasets due to three key factors. First, we observed diminishing returns: in jpn_finance, the improvement over LOGSIGMA was negligible (0.645 vs. 0.650), indicating that LLM scale does not always translate to better performance in specialized or lower-variance domains. Second, the fundamental V/A asymmetry remained unresolved; the model still struggled significantly more with Arousal than Valence (1.08 vs. 0.54 in Russian), suggesting this is a data-level challenge rather than a capacity issue. Finally, the extreme computational cost of fine-tuning 4B-parameter models made them impractical for the competition’s multilingual requirements, where our encoder-based discriminative models offered a much higher efficiency-to-performance ratio.

| Dataset | RMSE _V | RMSE _A | RMSE | LOGSIGMA |
|----------------|-------------------|-------------------|--------------|----------|
| rus_restaurant | 0.538 | 1.077 | 0.808 | 1.028 |
| jpn_finance | 0.835 | 0.455 | 0.645 | 0.650 |

Table 20: Post-competition Qwen3-4B fine-tuning (2 regression heads, 5 epochs, MSE loss) on two Track A dev datasets (RMSE ↓). While Qwen3-4B delivers a substantial improvement in the Russian domain, it exhibits diminishing returns in Japanese finance, leading to our decision to prioritize more efficient encoder models.

L Ensemble Ablation

Table 21 compares individual seed runs against the 3-seed ensemble for two representative datasets.

| Dataset | Track | Seed 1 | Seed 2 | Seed 3 | Ensemble | Δ |
|-------------------|-------|--------|--------|--------|--------------|----------|
| eng_restaurant | A | 1.277 | 1.211 | 1.214 | 1.153 | −4.8% |
| eng_environmental | B | 1.218 | 1.201 | 1.251 | 1.161 | −3.3% |

Table 21: Single-seed vs. 3-seed ensemble (RMSE ↓). Δ is the relative improvement from the best individual seed to the ensemble. Averaging predictions across three seeds consistently improves over even the best single run, smoothing variance from random initialization.

M Language Difficulty Ranking

Table 22 ranks all 15 dataset–language combinations by best development RMSE.

| # | Language | Track | Domain | Best RMSE ↓ |
|----|-----------|-------|---------------|-------------|
| 1 | Chinese | A | Finance | 0.373 |
| 2 | Chinese | B | Environmental | 0.409 |
| 3 | Chinese | A | Laptop | 0.520 |
| 4 | Chinese | A | Restaurant | 0.532 |
| 5 | Japanese | A | Finance | 0.606 |
| 6 | Japanese | A | Hotel | 0.619 |
| 7 | Pidgin | B | Politics | 0.624 |
| 8 | English | A | Laptop | 0.660 |
| 9 | English | A | Restaurant | 0.686 |
| 10 | Russian | A | Restaurant | 0.808 |
| 11 | German | B | Politics | 0.940 |
| 12 | Ukrainian | A | Restaurant | 0.967 |
| 13 | English | B | Environmental | 1.086 |
| 14 | Swahili | B | Politics | 1.234 |
| 15 | Tatar | A | Restaurant | 1.285 |

Table 22: All 15 dataset–language combinations ranked by best development RMSE across all configurations tested. Chinese is the easiest language across all domains and tracks (0.37–0.53). Tatar and Swahili are the hardest (>1.23), both lacking dedicated pretrained encoders. English environmental ranks 13th—harder than German politics—likely because environmental stance text is distributionally different from the sentiment-rich data that Twitter-RoBERTa was pretrained on. Note: some entries use earlier configurations (e.g., XLM-R-Large) that outperformed the final language-specific system on individual datasets; final system results are in Tables 8 and 9.

N Error Analysis

Domain transfer failure. Chinese-RoBERTa achieves 0.44–0.58 RMSE on Track A Chinese reviews (Table 8) but is our weakest submission on Track B Chinese environmental stance (7th/13 teams, 0.646 blind test RMSE). Same encoder, same language, different domain: the encoder’s sentiment representations do not transfer to stance-detection text. This tells us that domain-specific pretraining matters as much as language-specific pretraining.

Low-resource encoder gap. Tatar (1.285 RMSE) and Swahili (1.234) post our two worst results (Table 22). Both lack dedicated monolingual encoders: Swahili falls back to XLM-R-Large while Tatar uses ruBERT as a Russian proxy. The gap to the next-hardest dataset (English environmental: 1.086) is large, confirming that the encoder gap—not dataset size alone—is the primary driver. Developing monolingual encoders for these languages remains the clearest path to improvement.

O Comparison Against Dataset Paper Baselines

Table 2 includes official baselines from the Track A (Lee et al., 2026) and Track B (Becker et al., 2026) dataset papers: Kimi-K2 Thinking and Qwen-3 14B on Track A, and Mistral-3 14B and mBERT on Track B. Three factors explain the performance gap.

Track A baselines use token-based generation. Both Track A baselines predict VA scores auto-regressively as text tokens. Kimi-K2 Thinking uses zero/one-shot prompting; Qwen-3 14B is fine-tuned via supervised learning but remains a generative decoder. The Track A dataset paper explicitly notes that this produces “a random grid-like VA distribution, likely due to its token-based generation.” VA regression requires a continuous output space, which auto-regressive generation cannot provide. This structural mismatch—not model capacity—explains why Kimi-K2 Thinking scores 2.189 RMSE (32nd/33) on eng_laptop and 2.146 (36th/37) on eng_restaurant, while LOGSIGMA achieves 1.241 and 1.104 on the same sets using direct MSE regression.

All baselines apply one model across all languages. On Track B, Mistral-3 14B uses a regression head (Becker et al., 2026)—the same discriminative setup as LOGSIGMA—yet ranks 7th–10th because it uses a single LLM for all five languages. mBERT, a generic multilingual encoder, finishes last on every dataset (e.g., 2.329 vs. our 1.342 on German; 3.215 vs. our 1.127 on Pidgin). The Track B dataset paper’s own PLM experiments confirm this: XLM-R-Large, the strongest single multilingual encoder baseline, averages 1.641 RMSE across languages—more than twice our 0.779 (4-language average). LOGSIGMA uses dedicated pretrained encoders per language (Table 1), which alone accounts for 8.2% average RMSE improvement over the multilingual baseline (§5.2).

No baseline addresses V/A difficulty or variance. Arousal is harder than Valence in every dataset (avg. PCC gap of 0.29; Table 13). All baselines—whether generative or regression-based—apply uniform loss weighting, implicitly over-weighting the easier Valence objective. LOGSIGMA’s learned parameters (s_V, s_A) correct this automatically, with the largest gains where the V/A gap is widest: Pidgin (−11.2%), English environmental (−8.0%), Japanese finance (−9.9%) (Table 3). A further 3–5% comes from 3-seed ensembling, which no baseline employs (Table 21).