RPI Team at SemEval-2026 Task 3: An LLM-Encoder Ensemble for Coarse-to-Fine Valence-Arousal Sentiment Prediction

Mohammed Shahid Modi and Boleslaw Szymanski
Rensselaer Polytechnic Institute
Troy, New York, USA
modim2@rpi.edu, szymab@rpi.edu

Abstract

We present our coarse-to-fine Valence-Arousal (VA) ensemble system for subtask 1 of task 3 (DimABSA) which covers aspect-level VA prediction. We use a pair of trained Qwen 3 8B LoRA-tuned LLMs to predict coarse bins between 1 and 8, providing ordinal VA guidance signals along with distributional features. We then train an instruction-style, multilingual E5 encoder model with a multitask head using these LLM-derived guidance features to produce continuous VA predictions. At inference time, the same guidance signals are generated for the test set by the trained LLMs and fed into the trained encoder. This approach leverages the LLM as a high-level prior while relying on the encoder for precise calibration across languages and domains. Our system achieves an RMSE_{VA} of 1.20 across six languages and five domains. We compare the joint VA model to separated valence and arousal models trained on coarsened ground truth data, showing that it outperforms them, particularly on arousal correlations.

1 Introduction

The field of sentiment analysis has evolved in favor of continuous models that allow for learning granular distinctions between sentences even if they convey similar emotions. In this vein, SemEval-2026 Task 3 is focused on Dimensional Aspect-Based Sentiment Analysis or DimABSA (Yu et al., 2026), requiring systems to predict continuous valence (negative-to-positive axis) and arousal (calm-to-excited axis) scores for aspects, provided a short text of a review and each aspect in that snippet. This task spans six languages (English, Japanese, Russian, Ukrainian, Chinese, Tatar) and five domains (laptop, restaurant, hotel, device, finance). It presents a multilingual and multi-domain regression problem. Specifically, this task requires that we predict aspect-level valence and arousal scores on a 1–9 scale.

Aspect-Based Sentiment Analysis (ABSA) deals with aspects, which are individual subjects of interest within a text. While categorical labels simplify emotional states into general polarities, they remove important information within those categories. For instance, a mildly disappointed review and a furious one both fall under the "negative" label categorically. ABSA instead represents affect as a continuous point in a two-dimensional space, following Russell's circumplex model of affect (Russell, 1980) where emotion is characterized along the axes of valence and arousal. The DimABSA shared task (Yu et al., 2026) and the accompanying dataset (Lee et al., 2026) provide a baseline for aspect-level continuous valence and arousal predictions for this task.

We found that predicting continuous dimensions in this task was challenging for three reasons. (1) the output space has no clearly defined boundaries, meaning models that learn hard categorical boundaries are punished and what separates close scores may not be intuitive, (2) valence and arousal have different distributions, with arousal actually being between 3-9 for the data in the training set, and (3) multilingual generalization is a struggle for low-resource languages, such as Tatar in this dataset.

We address these challenges with a two-stage coarse-to-fine approach. First, we train two LLMs to predict ordinal bins, providing high-level guidance. Second, a multilingual encoder uses these predictions as input features to produce final continuous scores. We find that LLM-derived distributional features provide effective soft supervision while jointly modeling valence and arousal substantially improves arousal prediction, creating an ensemble that can predict good VA scores across multiple languages, while using models that are smaller and require less compute than many mainstream LLMs. Our main challenge was achieving consistent performance across all languages.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3218–3224

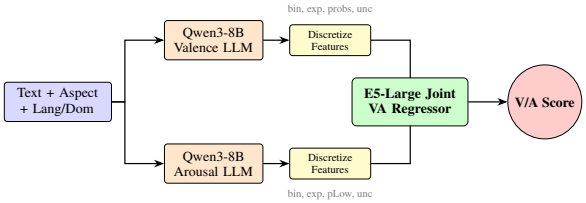

Figure 1: System architecture. Two LoRA-tuned Qwen3-8B LLMs produce coarse ordinal predictions for valence and arousal. Their outputs (bin labels, expected scores, probabilities, uncertainty) are discretized and concatenated with the input text to guide a multilingual E5 encoder that jointly regresses continuous VA scores.

2 Background

The coarse-to-fine approach has appeared across domains and use-cases. In (Mok and Chung, 2022), the authors align medical images at low resolutions, then refine the alignment further by increasing resolution slowly. In (Panousis et al., 2024), the authors discover high-level whole-image concepts first, followed by low-level sub-region concepts. In (Dong and Lapata, 2018), the authors take natural language utterances and generate rough sketches of their overall meaning, then use the sketch and the utterance to fill in the missing information. There are other examples and different variations of the coarse-to-fine approach (Charniak and Johnson, 2005; Kiddon and Domingos, 2011).

LLMs are often used to make high-level decisions due to their contextual understanding. Tool use could be considered an example of this, as LLMs decide which tool to use, and then the tool solves the actual problem presented to them (Schick et al., 2023). For natural language tasks, this ability of LLMs to make high-level decisions suggests that they could excel at making coarse observations about the data, providing a useful signal for a more focused model to use for guidance.

In addition, our coarse-to-fine approach leverages the full probability distribution of LLM inferences over ordinal bins to provide soft guidance for a specialized regression model. This combines the linguistic understanding of LLMs with the calibration advantages of encoder-based regressors.

3 System Overview

We use the Qwen3-8B model (Team, 2025) for coarse prediction, which we chose due to its strong multilingual capabilities at a relatively compact size. For fine-grained regression, we use `intfloat/multilingual-e5-large-instruct`, a 1024-dimensional multilingual encoder optimized for instruction-following tasks (Wang et al., 2024). Thus, our system consists of a valence LLM predictor, an arousal LLM predictor and a multilingual encoder regressor. Figure 1 illustrates our pipeline.

3.1 LLMs for Coarse Bin Prediction

We fine-tune two separate Qwen3-8B models to predict integer bins from 1–8 for valence and from 3–8 for arousal (reflecting the data distribution where arousal never falls below 3.0).

Training Setup. We fine-tune our LLMs using Low-Rank Adaptation (LoRA) (Hu et al., 2022), a parameter-efficient method that adds small trainable low-rank matrices to selected transformer layers while keeping the original weights frozen. We use rank $r = 16$, $\alpha = 32$, and dropout 0.05, applying LoRA to the attention and MLP projection layers (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj).

Prompts include language, domain, text, and aspect information, with a special `<ANS>` token marking where the model predicts the digit (See A for prompt template). We train for 3 epochs with learning rate 3×10^{-5}, batch size 48, maximum sequence length 384, and use a weighted sampler to balance language and coarse-bin distributions.

Valence LLM Loss. For the valence model, we use cross-entropy loss with hard targets where the ground-truth valence is rounded to the nearest integer (clipped to 1–8) to produce a single target class. The loss is:

$$\mathcal{L}_V = \mathrm{CE}(p, y^*) + \lambda_{EV} \cdot \mathrm{MSE}(\mathbb{E}[d], v) \quad (1)$$

where $\mathrm{CE}(\cdot,\cdot)$ denotes cross-entropy, $\mathrm{MSE}(\cdot,\cdot)$ denotes mean squared error, $p = (p_1, \ldots, p_8)$ is the softmax distribution over digit tokens 1–8, p_i is the probability assigned to digit i, v is the ground-truth valence score, $y^* = \mathrm{round}(v) \in \{1, \ldots, 8\}$ is the hard target class obtained by rounding and clipping v, $E[d] = \sum_{i=1}^{8} p_i \cdot i$ is the expected digit under p, and $\lambda_{EV} = 0.25$ weights the expected-value regression term.

Arousal LLM Loss. For the arousal model, we use soft cross-entropy with Gaussian-smoothed targets to capture ordinal structure. The loss we use is:

$$\mathcal{L}_A = -\sum_i q_i \log p_i + \lambda_{EV} \cdot \mathrm{MSE}(\mathbb{E}[d], a) \quad (2)$$

where a is the ground-truth arousal score, $p = (p_3, \ldots, p_8)$ is the softmax distribution over digit tokens 3–8, p_i is the probability assigned to digit i, $q = (q_3, \ldots, q_8)$ is the Gaussian-smoothed soft target distribution with $q_i \propto \exp(-(i-a)^2/2\sigma^2)$ and $\sigma = 0.8$, $E[d] = \sum_{i=3}^{8} p_i \cdot i$ is the expected digit under p, MSE$(\cdot, \cdot)$ denotes mean squared error, and $\lambda_{EV} = 0.4$ weights the expected-value regression term.

In our early observations, we felt that soft targets encouraged the model to assign probability mass to neighboring bins, producing better-calibrated distributions that are more useful for their ultimate goal of guiding a regressor model.

Guidance Features. At inference, we extract the predicted digit, expected score, coarse-bin indicators, and uncertainty features from the LLM outputs. For valence, this includes A/B/C bin information with probabilities p_a , p_b and p_c. For arousal, we use a B/C-oriented coarse signal together with the probability mass assigned to the low range (p_{low}), along with the predicted digit, expected score, and uncertainty statistics. Note that the bin breakdown is bin A for values less than three, bin B for values three or greater but less than six, and bin C for values six and over.

3.2 Encoder for Fine Bin Prediction

Input Construction. We format inputs in E5's instruction style, converting most continuous LLM guidance features into coarse buckets (e.g., confidence and uncertainty levels) while keeping the predicted digit and expected score as numeric strings. (see B for full guidance details)

Model Architecture. The encoder output is mean-pooled, then passed through a projection layer (Linear $\rightarrow$ GELU $\rightarrow$ Dropout 0.1), then fed to separate valence and arousal regression heads. Outputs are constrained to [1,9] via sigmoid scaling: $y = 1 + 8 \cdot \sigma(z)$. We also include auxiliary 3-way bin classification heads for regularization.

Training. We use Smooth L1 loss for regression and auxiliary cross-entropy for coarse bin classification:

$$\mathcal{L} = \mathcal{L}_V^{\text{enc}} + \mathcal{L}_A^{\text{enc}} + \lambda_{bin}(\mathcal{L}_{V_{bin}} + \mathcal{L}_{A_{bin}}) \quad (3)$$

where L_V^{enc} and L_A^{enc} are the Smooth L1 regression losses ($\beta = 0.5$) for valence and arousal respectively, $L_{V_{\text{bin}}}$ and $L_{A_{\text{bin}}}$ are the auxiliary cross-entropy losses over the 3-way coarse bin classification heads for valence and arousal respectively, and $\lambda_{\text{bin}} = 0.15$ weights the auxiliary classification term. Training uses AdamW (lr=2×10^{-5}, weight decay=0.01), linear warmup (6% of steps), a batch size of 64, and language-balanced sampling.

Language	Domain	Total
English (eng)	Laptop	7,469
	Restaurant	5,503
	Subtotal	*12,972*
Japanese (jpn)	Hotel	4,222
	Finance	3,293
	Subtotal	*7,515*
Russian (rus)	Restaurant	4,205
Ukrainian (ukr)	Restaurant	4,205
Chinese (zho)	Laptop	8,733
	Restaurant	11,137
	Finance	5,550
	Subtotal	*25,420*
Tatar (tat)	Restaurant	4,205
Grand Total		**58,522**

Table 1: Dataset statistics by language and domain.

4 Experimental Setup

Data. The DimABSA dataset provides train/dev/test splits across 6 languages and 5 domains, totaling approximately 58K aspect-level annotations (Lee et al., 2026). Table 1 contains a summary of each language and domain. For our final submission, we trained the encoder on all available labeled data (train and dev) with a 90/10 internal split for validation. We cleaned aspects by replacing null aspect values with "general sentiment."

Evaluation. The official evaluation metric for the task is RMSE_{VA}, computed as:

$$\sqrt{\sum_{i=1}^{N} \frac{(V_p^{(i)} - V_g^{(i)})^2 + (A_p^{(i)} - A_g^{(i)})^2}{N}} \quad (4)$$

RMSE_{VA} or Root Mean Squared Error of Valence-Arousal measures the square root of the average squared distance between predicted (V_p, A_p) and ground truth values (V_g, A_g) of both valence (V) and arousal (A) across all datapoints (N).

In addition, the Pearson Correlation Coefficients for valence (PCC_V) and arousal (PCC_A) are also

Lang	Domain	E5-VA (Ours)			Separate V+A (Ablation)			Baselines ($RMSE_{VA}$)	
		$RMSE_{VA}$	PCC_V	PCC_A	$RMSE_{VA}$	PCC_V	PCC_A	Kimi-K2	Qwen3-14B
ENG	laptop	**1.283**	0.871	**0.521**	1.376	**0.886**	0.346	2.1893	2.8089
ENG	restaurant	**1.201**	0.896	**0.626**	1.341	**0.902**	0.487	2.1461	2.6427
JPN	finance	**0.825**	**0.868**	**0.624**	0.871	0.863	0.439	1.640	1.896
JPN	hotel	**0.641**	**0.949**	**0.741**	0.724	0.938	0.494	1.755	2.291
RUS	restaurant	1.485	0.905	**0.622**	**1.370**	0.905	0.461	1.777	2.153
TAT	restaurant	1.780	**0.784**	**0.522**	**1.741**	0.739	0.302	1.938	2.637
UKR	restaurant	1.549	0.891	**0.595**	**1.391**	**0.901**	0.451	1.781	2.212
ZHO	finance	**0.540**	0.848	**0.638**	0.655	0.848	0.604	1.965	1.471
ZHO	laptop	**0.701**	**0.896**	**0.726**	0.808	0.893	0.562	1.644	1.771
ZHO	restaurant	**0.960**	**0.862**	**0.579**	1.060	0.838	0.440	1.896	2.007
Macro Avg		**1.097**	**0.877**	**0.619**	1.134	0.871	0.459	1.800†	2.055†

Table 2: Test set results by language-domain comparing our E5-VA model, the Separate V+A ablation, and official baselines. Bold indicates the better score between E5-VA and Separate V+A. E5-VA scores better on 7 of 10 pairs on $RMSE_{VA}$ and achieves 35% higher PCC_A overall.

used as a supplemental datapoint, defined as:

$$\frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}} \quad (5)$$

where, n is the number of aspects, x_i, y_i are the predicted valence (or) arousal values for aspect i and the corresponding ground-truth values, and $\overline{x}, \overline{y}$ are the means of the predicted valence (or) arousal values and the means of the ground-truth valence (or) arousal values respectively. This number measures the strength of the linear relationship between predicted and ground-truth set valence (or) arousal values. PCC values closer to 1 are better.

Variants. We compare two configurations: (1) **E5-VA**: single encoder jointly predicting both dimensions with full LLM guidance features; (2) **Separate V+A**: two independent E5 encoders for valence and arousal, using only coarse bin tokens without distributional features. The separate encoders serve as an ablation and they are trained using coarsened ground-truth values from the training dataset, but used the LLM's coarse bins (same as above) during test-time inference, showing how our approach of using LLM inferences for guidance during training instead of a ground-truth signal affects outcomes.

Tools. We use PyTorch, Hugging Face Transformers, PEFT for LoRA, as well as Flash Attention 2 (Dao et al., 2023) for efficient LLM training. We will make model weights available at https://github.com/modim2rpi/dimabsa_rpi.

5 Results

Table 2 shows test set performance comparing our E5-VA model, the Separate V+A ablation, and two official LLM baselines (Kimi-K2 Thinking and Qwen3-14B zero-shot). Our final E5-VA model achieves a macro-averaged $RMSE_{VA}$ of **1.097**, substantially outperforming both baselines and winning on 7 of 10 language-domain pairs against the separate-model ablation. (Note that there was a very slight difference in the TAT restaurant figures between the metrics calculated by us using the official metrics code and the result computed by codabench (1.784 vs 1,780 RMSE). Also note that the overall RMSE provided by codabench for our submission is 1.20.)

Ablation Analysis. The joint model improves arousal prediction (PCC_A: 0.619 vs 0.459) by 35% relative to the separated arousal model, while maintaining comparable valence performance. This suggests that sharing representations between valence and arousal is beneficial. Extreme values of valence convey strong emotions such as joy, grief, high satisfaction and anger, and it is likely that such emotions are expressed with more intensity relative to more moderate valences. This would explain why the joint model predicts VA values more accurate.

The joint model also benefits from distributional features from the LLMs (expected values, entropy, bin probabilities) rather than just the coarse ground-truth tokens used by the separate models during training. The separate models have to use LLM coarse bins for test set inference, and any advantage gained from perfectly accurate coarse bins during training is offset by receiving a poor inference

guidance signal.

That said, the joint model struggles more with languages for which training data is sparse, and is beaten by the separate models. As different languages use varied verbiage for emotional expression, the joint model's learned speech patterns for one language may not generalize to another. Neither the Qwen3 models nor the encoder may have picked up on subtle differences during training for the less-represented languages. The separate model outperformed in this case, which is possibly due to stronger alignment for less-represented languages attained from ground-truth coarse bins. It is possible that joint training may overfit to better-represented languages, and an architecture combining both separated and joint model approaches could yield further improvements.

6 Conclusion

We presented a coarse-to-fine approach for aspect-level VA prediction, using LLM-generated ordinal predictions as guidance for a multilingual encoder. Our key contributions are: (1) demonstrating that LLM probability distributions provide effective soft supervision, (2) showing that joint VA modeling improves arousal prediction through shared representations, and (3) highlighting the potential drawbacks of such an ensemble.

Limitations

A more detailed ablation study, such as the use of different LLMs, different loss functions and variables, and a different ML architecture for fine VA prediction, would reveal more advantages and disadvantages of our technique. Any inaccuracies in the LLM prior's predictions are propagated downstream in the joint architecture at least to some extent, and a detailed validation of the LLMs would show where further improvements could be made. We aim to perform such testing in the future.

The model training was limited to the datasets provided by the task's organizers. For deployment of this technique in different scenarios, more training and finetuning is recommended. The model performs best when trained on data that provides language and domain information. A more generic version of this architecture would employ a separate model or technique to extract domain and aspects for each point of raw data. Indeed, the other subtasks in this track moved toward this goal.

Ethical Considerations

It may be said that valence-arousal prediction could potentially be misused for emotion surveillance or manipulation on social media platforms. Also, the lower performance on Tatar highlights the risk of NLP systems under-serving speakers of less-spoken languages.

Acknowledgments

We thank the DimABSA task organizers for creating this exciting task. We thank RPI's Future of Computing Institute for providing computing resources that made this research possible.

AI use statement: An AI tool was used for proofreading and grammar suggestions.

References

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173–180, Ann Arbor, Michigan. Association for Computational Linguistics.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Li Dong and Mirella Lapata. 2018. Coarse-to-fine decoding for neural semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–742, Melbourne, Australia. Association for Computational Linguistics.

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

Chloé Kiddon and Pedro Domingos. 2011. Coarse-to-fine inference and learning for first-order probabilistic models. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI-11)*, pages 1049–1056.

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.

Tony C. W. Mok and Albert C. S. Chung. 2022. Affine medical image registration with coarse-to-fine vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20835–20844.

Konstantinos P. Panousis, Dino Ienco, and Diego Marcos. 2024. Coarse-to-fine concept bottleneck models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Timo Schick, Jonathan Zhange, Hinrich Schütze, and Colin Raffel. 2023. Toolformer: Language models can teach themselves to use tools. In *International Conference on Learning Representations*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

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.

A Prompt Templates

We use structured prompts to elicit ordinal bin predictions from the Qwen3-8B LLMs. The prompts include language and domain metadata to enable cross-lingual learning.

Valence LLM Prompt:

```
You are a sentiment classification
system.

Task:
Given a text and an aspect, output a
single integer from 1 to 8
representing the valence expressed
toward the aspect.

Output rules:
- Output exactly ONE integer
- No explanation
- No additional text

Language: {language}
Domain: {domain}
Text: {text}
Aspect: {aspect}

Output: <ANS>
```

The <ANS> token marks the position where digit prediction occurs. We extract logits only for digit tokens 1–8 at this position.

Arousal LLM Prompt. The arousal prompt follows an identical structure, but requests integers from 3 to 8. It states:

```
Given a text and an aspect, output a
single integer from 3 to 8 representing
the arousal expressed toward the aspect.
```

E5 Encoder Input. The multilingual E5 encoder uses an instruction-style format that incorporates all LLM guidance features:

```
Instruct: Predict valence and arousal
(continuous) toward the aspect. Use the
guidance features if helpful, but rely
on the text.
Query: Lang={lang} Domain={dom}
V:bin={vbin} digit={vdig} exp={vexp}
  conf={vconf} pA={vpA} pB={vpB} pC={vpC}
  unc(top1={v_t1},marg={v_m},ent={v_e})
A:bin={abin} digit={adig} pLow={aplow}
  exp={aexp}
  unc(top1={a_t1},marg={a_m},ent={a_e})
Aspect: {aspect}
Text: {text}
```

B Guidance Feature Discretization

To prevent the E5 encoder from over-relying on exact LLM outputs, we discretize some continuous guidance features into categorical buckets. Table 3 lists the bucketing schemes.

Feature	Bucket Edges
Confidence	0.05, 0.15, 0.30, 0.50, 0.70, 0.85
Bin probs.	(same as confidence)
Margin	0.05, 0.15, 0.30, 0.50, 0.70, 0.85
Entropy	normalized then rounded to 7 levels

Table 3: Discretization bucket edges for LLM guidance features. Confidence, bin probabilities, top-1 probability, and margin use the same bucket scheme. Entropy is first normalized by the maximum entropy for the digit set, then mapped to 7 discrete levels. Expected score and digit are kept as numeric strings (rounded to 2 decimals for exp).

Guidance Features Summary. For each input, we extract the following from the valence and arousal LLMs:

- **Coarse bin**: A/B/C for valence (low/mid/high), B/C for arousal
- **Digit**: Argmax predicted integer (1–8 for V, 3–8 for A)
- **Expected score**: $\mathbb{E}[d] = \sum_i p_i \cdot i$
- **Bin probabilities**: p_A, p_B, p_C (V) or p_{low} (A)
- **Confidence**: Highest single digit probability (most confident predicted digit bin)
- **Margin**: Difference between top-1 and top-2 probabilities
- **Entropy**: $-\sum_i p_i \log p_i$ over the digit distribution