transformer_1376 at SemEval-2026 Task 9: A Multi-Stage Pipeline with Calibrated Ensembles and Lexical Post-Processing for Online Polarization Detection in Bengali

Shuvodwip Saha, Pritha Saha
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology
shuvodwipsaha@gmail.com, prithasaha2022@gmail.com

Abstract

The POLAR @ SemEval-2026 Task 9 deals with the detection of online polarization in a variety of multilingual and multicultural environments. Our team participated in Subtask 1 of the POLAR @ SemEval-2026 Task 9, which mainly deals with binary classification of textual sequences for the detection of polarized stances. In this paper, we proposed a strong classification system for Bengali language based on fine-tuning the BanglaBERT Large model. The methodology used here involves a stratified five-fold cross-validation approach along with a performance-weighted ensemble method, combined with temperature scaling probability calibration and a set of lexical post-processing rules.

1 Introduction

Polarization is a major systemic risk to civic discourse wherein opinions split into opposing groups (Banim, 2025). The detection of this phenomenon involves identifying the presence of 'attitude polarization,' wherein group identity is associated with the vilification of opposing groups (Naseem et al., 2026b). This is further complicated by the presence of 'echo chambers' on the internet (Garrett, 2009) and 'micro agenda setters' on social media (Wohn and Bowe, 2016). The POLAR @ SemEval-2026 Task 9 tracks the performance of such detectors on a collection of 22 languages (Naseem et al., 2026a).

This paper describes our work on Subtask 1 of the Bengali language track of this competition. Bengali is a particularly challenging language for polarization detection because of the linguistic markers of hostility. We address this through a system based on the large-scale pretrained BanglaBERT encoder (Bhattacharjee et al., 2022). [1] Our methodology is based on a five-fold stratified ensemble method combined with temperature scaling and a post-processing layer to achieve precision through a combination of deep semantic understanding and rule-based precision. Our system achieved a **macro F1 of 0.8463** on the Bengali track of the competition. Our team achieved 5th place out of 36 teams in the leaderboard.

2 Background and Related Work

Polarization detection has evolved from sentiment analysis to "attitude polarization" and "out-group" vilification detection (Naseem et al., 2026b). Recent research has identified polarization as a "systemic risk" (Banim, 2025) marginalizing vulnerable groups through affective hostility, particularly in highly conflict-ridden societies (Ali et al., 2025).

Our research is a part of the POLAR @ SemEval-2026 Task 9, which involves the detection of multilingual, multicultural, and multi-event online polarization detection (Naseem et al., 2026a).

While existing research has successfully applied graph neural networks to detect online "ideological framing" (Hofmann et al., 2022) or "aspect-based sentiment analysis" for media tracking purposes (Miehling et al., 2025), our research extends the BanglaBERT architecture, which has been pretrained on a 27.5 GB corpus to detect nuances of native language (Bhattacharjee et al., 2022).Our system specializes in "othering" in low-resource environments, thus differing from other frameworks through a multi-stage pipeline.

2.1 Dataset Description

The benchmarking dataset for this research has been provided by the POLAR @ SemEval-2026 Task 9, specifically the POLAR dataset (Naseem et al., 2026b).Our research has been conducted exclusively on the Bengali language. The dataset has

[1] Code:https://github.com/Shuvodwip/SemEval-2026-Task-9-Subtask-1

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2082–2088

been obtained from event-driven social media comments, thus reflecting real-world data.

The dataset has been split into 3 sets: a training set of 3,333 instances, a development set of 166 instances, and a test set of 1,501 instances. Each data instance has 14–15 feature columns, out of which we have specifically used the "id," "text," and "polarization" columns for the purpose of Subtask 1.

The distribution of the training data shows a minor class imbalance, with 1,909 non-polarized samples (57.28%) and 1,424 polarized samples (42.72%).

3 System Overview

To address the problems associated with the small dataset size and class imbalance, we propose an architecture that uses a strong pre-trained encoder, cost-sensitive learning, weighted ensembling, probability calibration, and lexical post-processing. Figure 1 illustrates the end-to-end architecture of our proposed model.

3.1 Base Architecture

We use BanglaBERT Large (csebuetnlp/banglabert) as our primary feature extractor and classifier. A linear classification layer is added on top of the [CLS] token representation to output the binary class predictions. To prevent overfitting on the relatively small dataset, we set the hidden dropout probability to 0.1 and attention probabilities dropout probability to 0.1. All the hyperparameters used during the training process are provided in Section 4 and Appendix A.

3.2 Handling Class Imbalance

To address class imbalance bias towards the majority Non-Polarized class due to the observed fold imbalance during the training process, we replace the traditional loss function with Weighted Cross-Entropy Loss. The class weights are calculated dynamically by scaling the class frequencies inversely. A larger penalty is assigned to the back-propagation loss for misclassified Polarized class instances.

3.3 Cross-Validation and Weighted Ensembling

To improve the reliability of our model and its ability to perform well on the unseen test set, we incorporate a Stratified 5-Fold Cross-Validation approach. Instead of taking the arithmetic mean of all five models' predictions, we implement Weighted Ensemble.

For each of the five folds, we test the model on its corresponding internal validation set and compute its Macro F1 score. The final probability output by the ensemble model for a particular instance is calculated by taking a weighted sum of the softmax probability outputs from all five models. The weights are proportional to the square of the validation set F1 score for each fold. Squaring the F1 score gives more preference to better-performing folds while diminishing the effect of poor-performing folds:

$$W_i = \frac{F1_i^2}{\sum_{j=1}^{5} F1_j^2}$$

$$\hat{P}_{ensemble} = \sum_{i=1}^{5} W_i \cdot P_i$$

3.4 Calibration and Threshold Tuning

Predictions from large-scale models exhibit the bias of being over-confident. This issue can be resolved by implementing temperature scaling ($T = 1.3$). Because the ensemble is obtained in probability space, scaling will be performed on the log-transformed probabilities of the ensemble due to multi-model architecture:

$$P_{calibrated} = softmax\left(\frac{\log(\hat{P}_{ensemble} + \epsilon)}{T}\right) \quad (1)$$

where $\epsilon = 10^{-10}$ ensures numerical stability. The threshold itself was fine-tuned using grid search in the development set (threshold of 0.45–0.60) rather than the standard value of 0.50. Mathematical rationale for this approach is included in the Appendix (C).

3.5 Lexical Post-Processing

Large pre-trained language models are known to miss explicit cultural markers of polarization that fall outside their training distribution. In order to solve this, we introduce a deterministic post-processing module based on a manually constructed Bengali polarization lexicon. The lexicon includes 13 terms carefully selected by us, native Bengali speakers, using linguistic intuition and familiarity with polarization discussions in Bengali social media. Terms were picked such that they are strong signals of hostility, group-based vilification, and conflict—all crucial to the problem defi-

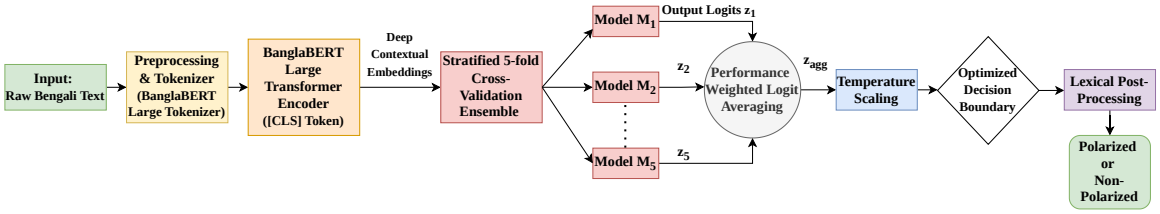

Figure 1: End-to-end architecture of our proposed system.

nition. All entries underwent independent evaluation, achieving consensus on the choice. The lexicon is intentionally small, acting as an experimental module to check if the error of a model can be fixed using a minimal number of terms, an assumption that appears true from a $\Delta F1$ of $+0.0008$ on the test set (Table 3). The full lexicon can be seen in Table 1.

The post-processing layer operates under three deterministic rules:

- **High-Confidence Override:** If $P_{calibrated} > 0.85$, the instance is classified as Polarized regardless of lexicon presence.
- **Low-Confidence Correction:** If $P_{calibrated} < 0.65$, the presence of any lexicon term flips the prediction to Polarized.
- **Borderline Adjustment:** For probabilities within ± 0.05 of the optimal threshold, instances containing two or more lexicon terms are shifted to Polarized; those without are shifted to Non-Polarized.

Bengali	Transliteration	English
শত্রু	śatru	enemy
দুশমন	duśman	foe
বিরোধী	birōdhī	opponent
ধ্বংস	dhbaṃsa	destruction
ঘৃণা	ghṛṇā	hate
চরমপন্থী	caramapanthī	extremist
উগ্রবাদী	ugrabādī	radical
বহিষ্কার	bahiṣkāra	expulsion
বিদ্বেষ	bidbeṣa	hostility
দাঙ্গা	dāṅgā	riot
হিংসা	hiṃsā	violence
বৈরিতা	bairitā	enmity
গোষ্ঠীগত	gōṣṭhīgata	group-based

Table 1: Bengali polarization lexicon used in the lexical post-processing layer.

3.6 Pipeline Walkthrough Example

To see how the pipeline handles overt hostility, consider this polarized example from the dataset: "সময় টিভি একটা ভুয়া চ্যানেল তাদের খবর মানেই বানানো মিথ্যা কথা যাকে বলে পা চাটা কুকুর সময় টিভি এর মালিক" (*Somoy TV is a fake channel, their news is made-up lies, the owner is a bootlicking dog*).

When this text passes through the BanglaBERT ensemble, the model detects several strong indicators of attitude polarization and out-group vilification. Specifically, it picks up on aggressive markers like "ভুয়া" (fake), "মিথ্যা কথা" (lies), and the highly derogatory phrase "পা চাটা কুকুর" (bootlicking dog). Because of this extreme vocabulary, the ensemble outputs a high calibrated probability (e.g., $P_{calibrated} = 0.89$). During post-processing, this score easily clears our high-confidence override threshold (> 0.85), allowing the system to confidently lock the final prediction to Polarized (1).

4 Experimental Setup

4.1 Dataset Preparation

Prior to training, we applied a strict data preprocessing protocol:

- **Label Formatting:** All polarization labels have been formatted into their corresponding numeric form, and instances with missing labels have been removed, while target variables have been cast into integers.
- **Text Handling:** Empty text strings have been explicitly replaced with the [EMPTY] token for maintaining dimensional consistency.
- **Split Utilization:** The training split has been utilized for 5-fold cross-validation, while the development split has been exclusively reserved for tuning the thresholds and calibrating the probabilities, and the test split has been exclusively reserved for evaluating the performance.

4.2 Hyperparameters and Reproducibility

In order to ensure reproducibility, random seeds for all environments have been set to 42.

BanglaBERT has been fine-tuned for 4 epochs with a learning rate of 1.5e-5 and an effective batch size of 16. The hyperparameter configuration has been included in the appendix A.

4.3 Implementation Frameworks

The computational pipeline has been implemented in Python 3.10, which has utilized PyTorch (v 2.1.0) and Hugging Face transformers (v4.36.0)[2] for architecture, while Scikit-learn (v1.3.2) has been utilized for supporting cross-validation, and Pandas and NumPy have been utilized for data handling.

4.4 Evaluation Measures

In accordance with the official guidelines for SemEval-2026 Task 9, the main evaluation metric used for ranking the systems is the Macro F_1-score metric. The Macro F_1-score is defined as the unweighted mean of the F_1-scores for the Polarized and Non-Polarized classes. The proposed metric is particularly suitable for this task because it guarantees the same evaluation of the models on the minority Polarized and majority Non-Polarized classes. For a detailed analysis of the results of our ablation study, we report the total Accuracy, Macro Precision, and Macro Recall as well.

5 Results and Analysis

5.1 Main Quantitative Findings

We performed an evaluation of the final proposed system on the official unseen test set. To test the effectiveness of our approach, the optimized BanglaBERT-large model was compared with other prominent pre-trained models: bert-base-uncased[3], google/muril-base-cased[4], XLM-Roberta-large[5], and the official task Baseline model.

As shown in Table 2, the final proposed BanglaBERT-large model achieved a Macro F_1 of 0.8463, which is significantly better than the performance of the standard bert-base-uncased model (0.7407) and outperforms the performance of MuRIL (0.8345) and XLM-R (0.8355). However, the official Baseline model achieved outstanding performance by achieving the highest Macro F_1 of 0.8528. This shows that although the performance of the language-specific model is significantly better than general multilingual models, the official Baseline model's performance is still outstanding.

Method/Model	Macro F_1 Score
bert-base-uncased	0.7407
google/muril-base-cased	0.8345
XLM-Roberta-large	0.8355
Official Submission (Ours)	**0.8463**
Baseline model	0.8528

Table 2: Comparison of our final model against other baseline architectures on the official test set.

5.2 Quantitative Analysis: Ablation Study

While that our final leaderboard-qualifying run achieved a Macro F_1-score of 0.8463, Table 3 summarizes an additive ablation analysis to assess the isolated effect of each step. The final ensembled and calibrated pipeline achieved a reproducible Macro F_1-score of 0.8425 on the test set. The small discrepancy between the peak-performing run and the ablation analysis is due to the inherent variability of fine-tuning large-scale encoder models and minor distributional shifts during the official test set inference. The results reveal some interesting observations:

- **High Baseline Saturation:** The performance of the Vanilla BanglaBERT baseline model was high with an Macro F_1-score of 0.8419, indicating that there is limited scope for architecture-level modifications since the pre-trained model's representation learning is already saturated with the Bengali text domain.
- **The Power of Class Weighting:** The class weighting step achieved the largest isolated performance gain with an improvement of $\Delta F_1 = +0.0027$.This confirms that addressing class imbalance is crucial for identifying the minority "Polarized" class without sacrificing overall accuracy.
- **Dev vs. Test Distribution Shifts:** The 5-fold ensemble and threshold tuning resulted in marginal test regressions of -0.0017 and -0.0012, respectively. Although validation set performance improved, the results indicate distributional changes and overfitting from the development set and over-optimization.
- **Efficacy of Lexical Post-Processing:** The performance drop was recovered by the rule-

[2] https://huggingface.co/docs/transformers/
[3] https://huggingface.co/bert-base-uncased
[4] https://huggingface.co/google/muril-base-cased
[5] https://huggingface.co/facebook/xlm-roberta-large

Model / Configuration	Accuracy	Precision	Recall	Macro F_1	ΔF_1
1. Baseline (Vanilla 1-Fold)	0.8454	0.8412	0.8426	0.8419	-
2. + Class Weights (1-Fold)	0.8481	0.8439	0.8454	**0.8446**	+0.0027
3. + 5-Fold Ensemble	0.8461	0.8417	0.8445	0.8429	-0.0017
4. + Calib & Threshold Tune	0.8454	0.8414	0.8420	0.8417	-0.0012
5. + Lexical Post-Processing (Final)	**0.8468**	**0.8435**	**0.8416**	0.8425	+0.0008

Table 3: Additive ablation study results on the official test set demonstrating the incremental impact of each pipeline component.

based post-processing step with an improvement of $\Delta\ F_1 = +0.0008$, showing that this step remains an effective solution to address edge cases by leveraging domain knowledge of polarization keywords.

5.3 Error Analysis

In order to get a better idea of the limitations of the model, the errors made in test-set predictions were analyzed. According to the confusion matrix (Figure 2), the model was able to predict 742 Non-Polarized samples and 530 Polarized samples correctly, whereas 126 False Positive predictions and 103 False Negative predictions indicated that there is a small sensitivity to the Polarized class.

The False Positives display a systematic pattern; the model misclassifies texts that employ aggressive or emotionally-charged language targeted against individuals but not texts containing ideological vilification towards the out-group. For example, the following non-polarized text (ID: ben_0cb1ccbc1d659aecc18c44219062858b) was incorrectly predicted as Polarized:

> "কাউয়া তোর সমস্যা এতটাই সময় শেষ এখন তোরে ভাবতে হবে যে আমি কি নিঃশ্বাস নেব নাকি নিশ্বাস নিব না সময় শেষ তোর আর তোর মন্ত্রী আত্মার পুলিশ কুত্তার মত দেশের মানুষ পিটাবে তোদের শেষ সময় পালনের সুযোগ পাবি না ইন্ডিয়া তোদের কে বাঁচাবে ওরা ইন্ডিয়া তাদের নিজের স্বার্থের জন্য করতেছে সময় শেষ ব্যাটা"
> (*"Your time is up, now you have to think whether you will breathe or not. Your time is up, your minister's soul, the police will beat the people of the country like dogs. You will not get the chance to flee. India will not save you, they are doing it for their own interest. Time is up, buddy."*)

Despite its extremely derogatory lexicon and having a threatening tone, the aggression does not target an ideological out-group but a certain political leader. The model makes the mistake of interpreting the individual-targeted hostility as group-based polarization because of its emotional nature.

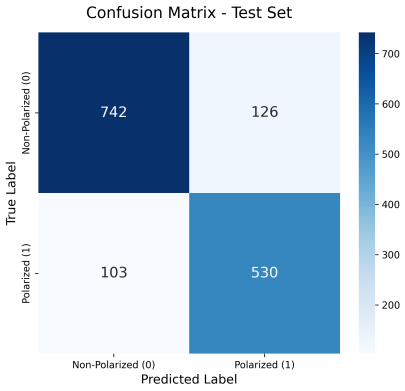

Figure 2: Confusion Matrix of the final model predictions on the test set.

The False Negatives reveal the complementary failure mode: polarization expressed through subtle identity-based othering, without any explicit hostility vocabulary. For example, the following polarized text (ID: ben_02b535475c1544cf00e28e1349da8232) was incorrectly predicted as Non-Polarized:

> "খবর পরিবেশনকারী আপনি মেয়ে নাকি ছেলে আপনি কি মুসলিম নাকি অমুসলিম মানুষ বুঝতে পারলাম না কোন হাদিসে র অনুসারে পোশাক পরিবর্তন করছেন"
> (*"Are you male or female? Are you Muslim or non-Muslim? I cannot tell. Which hadith justifies your clothing choices?"*)

This text encodes polarization through the simultaneous interrogation of a person's gender and religious identity — a form of implicit othering that contains no lexical signal detectable by either the transformer or the post-processing layer. The surface form of a mild question further obscures the polarized intent from the model.

While hard-label results have exposed certain problems, the probability distributions of the model show excellent results with high robustness. It is indicated by a discriminative power of the

ROC curve (Appendix B) having an Area Under the Curve (AUC) of 0.9143.

6 Conclusion

In our study, we propose a multi-stage pipeline for the POLAR SemEval-2026 Task 9 Bengali polarization detection task (Subtask 1) by utilizing the BanglaBERT Large model with a cost-sensitive weighted loss, stratified ensembles, and temperature scaling probability calibration. This helps us develop a reliable model for the detection of attitude polarization expressed in social media discourse.

The improvements in model accuracy can be attributed mainly to the use of class weights, which successfully address the imbalance in the training data. However, some other architectural features, such as the use of ensembling and probability scaling, led to marginal reductions in F_1 scores on the test set. This might be due to the distribution shift and the tendency of large transformer models, which are sensitive to the randomness of the model, to behave poorly on out-of-distribution data. In the future, we can also investigate the use of multi-task learning with fine-grained attributes and cross-lingual learning with the POLAR benchmark.

Limitations

Our work has a few limitations worth acknowledging. While our ensemble and threshold tuning consistently improved development set scores, this came at the cost of slight overfitting to the development distribution, reflected in a test regression of $\Delta F1 = -0.0029$ against the single-fold baseline. The official baseline (Naseem et al., 2026a), which fine-tunes LaBSE directly on the Bengali training split, ultimately performed better on the unseen test set. We attribute this to LaBSE's cross-lingual pre-training across 109 languages, which appears to produce more transfer-robust sentence representations, something our monolingual BanglaBERT encoder, despite its strong grounding in Bengali, struggled to match when faced with the stylistic variation of the test set.

Beyond this, our lexical post-processing layer covers only 13 terms, which we hand-picked based on our familiarity with Bengali polarization discourse. This means the layer inevitably misses subtler cases like texts where hostility is conveyed through irony, indirect phrasing, or idioms specific to certain communities. Building a richer, more systematic lexicon using corpus-driven approaches over polarized training examples is something can be explored further.

Ethics Statement

This study is significant not only to automated moderation but also to the sociolinguistic study of polarization in the Bengali language community, bearing in mind the possible dangers of such a study. Considering the sensitivity of the model to the use of aggression, in contrast to ideological polarization, we emphasize the importance of responsible use.

Acknowledgements

We thank the SemEval-2026 organizers for providing the benchmark dataset used in this work and for their support throughout the competition.

References

Adem Chanie Ali, Seid Muhie Yimam, Abinew Ali Ayele, Chris Biemann, and Martin Semmann. 2025. Silenced voices: social media polarization and women's marginalization in peacebuilding during the northern ethiopia war. *i-com*, 24(2):407–432.

Guy Banim. 2025. Very large online platforms—how big is your polarization footprint? towards a metric to give eu citizens transparency around an online systemic risk driving conflict in our societies. Report, Build Up.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

R. Kelly Garrett. 2009. Echo chambers online?: Politically motivated selective exposure among internet news users1. *Journal of Computer-Mediated Communication*, 14(2):265–285.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Valentin Hofmann, Xiaowen Dong, Janet Pierrehumbert, and Hinrich Schuetze. 2022. Modeling ideological salience and framing in polarized online groups with graph neural networks and structured sparsity.

In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 536–550, Seattle, United States. Association for Computational Linguistics.

Daniel Miehling, Daniel Dakota, and Sandra Kübler. 2025. Investigating polarization in YouTube comments via aspect-based sentiment analysis. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 718–728, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acarürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, and 15 others. 2026a. SemEval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, CA, USA. Association for Computational Linguistics.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, and 24 others. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *arXiv preprint arXiv:2505.20624*.

Donghee Yvette Wohn and Brian J Bowe. 2016. Micro agenda setters: The effect of social media on young adults' exposure to and attitude toward news. *Social Media + Society*, 2(1):2056305115626750.

A Hyperparameters and Reproducibility

To ensure full reproducibility, all random seeds across the environments were fixed to 42. Our training configuration was optimized for the large parameter space of BanglaBERT:

- **Sequence Length:** Tokenization was truncated to a maximum length of 128 subwords.
- **Optimization:** We trained the models for 4 epochs using a learning rate of 1.5e-5, incorporating a warmup ratio of 0.1 and a weight decay of 0.01.
- **Batching:** Due to GPU memory constraints, we utilized a per-device training batch size of 8 combined with 2 gradient accumulation steps, yielding an effective batch size of 16. Mixed precision (FP16) was enabled.

B ROC-AUC Curve

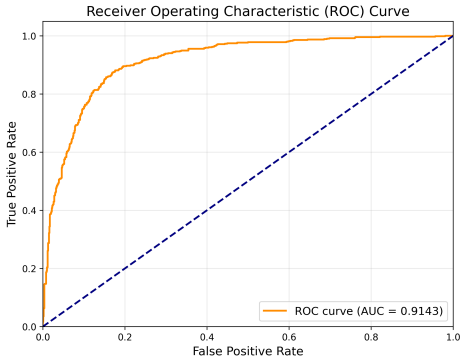

Figure 3: Receiver Operating Characteristic (ROC) Curve for the final ensembled model on the test set, demonstrating an AUC of 0.9143.

C Mathematical Formulation of Ensemble Calibration

For the single model, standard temperature scaling is defined as follows: $P = \sigma(z/T)$, where z stands for raw logits (Guo et al., 2017). In contrast, in our pipeline, the weighted ensembling scheme is utilized such that logits per each fold cannot be used to produce the final prediction in form of $\hat{P}_{ensemble}$.

To bridge this gap, we implement Post-Ensemble Probability Calibration. We recover a logit-like representation

($\hat{z}_{ensemble}$) from the ensembled probabilities using a log-transformation, where $\epsilon = 10^{-10}$ ensures numerical stability for values near zero:

$$\hat{z}_{ensemble} = \log(\hat{P}_{ensemble} + \epsilon) \quad (2)$$

The final calibrated probability can be derived by introducing the temperature parameter $T = 1.3$ into the recovered logit as follows:

$$P_{calibrated} = softmax\left(\frac{\hat{z}_{ensemble}}{T}\right) \quad (3)$$

This retains all the benefits offered by temperature scaling, such as the reduction of overconfidence of the model, within the current ensemble framework.