Optimizer Choice and Calibration for QARiB on Arabic-Script Social Media Offensive Language Detection

Auda M. Elshokry
University College of Applied Sciences, Gaza, Palestine
ashokry@ucas.edu.ps

Mohammed Alhanjouri
Islamic University of Gaza, Gaza, Palestine
mhanjouri@iugaza.edu.ps

Abstract

Optimizer choice is a central hyperparameter in fine-tuning transformer models, yet its impact remains under-studied for Arabic-script social media classification under class imbalance. We compare Adam, AdamW, and SGD for fine-tuning QARiB on two Arabic offensive-language benchmarks, OffensEval20 and MPOLD, using a controlled grid over learning rate, weight decay, and warmup, and report test-set performance as mean (std) over three random seeds.

Minority-class discrimination is evaluated using macro-F_1 and AUC-PR_{OFF}, while calibration is assessed via expected calibration error (ECE), reliability diagrams, and proper scoring rules (Brier score and negative log-likelihood, NLL). Across both datasets, AdamW and Adam are consistently strong and closely matched when properly tuned, whereas SGD substantially underperforms under the same tuning budget and exhibits higher seed sensitivity.

We observe non-trivial miscalibration across optimizers; post-hoc temperature scaling offers a low-cost adjustment, yielding modest, dataset-dependent changes in calibration while preserving ranking-based discrimination. We further evaluate a practical decision-rule step by optimizing the classification threshold on the validation set and applying it to test predictions, and provide qualitative examples illustrating typical optimizer-dependent confidence behaviors. In practice, for Arabic offensive-language detection under imbalance, we recommend starting from a tuned AdamW or Adam baseline; when calibrated probabilities are required for thresholding or triage, temperature scaling can be applied. We will release a reproducible pipeline to support further evaluation of optimizer–calibration trade-offs in Arabic-script safety tasks.

1 Introduction

Detecting offensive and abusive content in Arabic social media is both societally important and technically challenging due to dialectal diversity, code-mixing, noisy orthography, and strong class imbalance between benign and offensive content (Zampieri et al., 2020; Mubarak et al., 2021). Dedicated Arabic PLMs such as AraBERT, MARBERT, and QARiB are now standard tools, but it remains unclear how fine-tuning choices—especially optimizer choice and post-hoc calibration—affect both discrimination and confidence in this setting (Antoun et al., 2020; Abdul-Mageed et al., 2021; Abdelali et al., 2021; Guo et al., 2017; Shen et al., 2024; Xie et al., 2024).

Arabic-script user-generated text can further amplify sensitivity through rich morphology and cliticization, dialectal spelling variation, and orthographic inconsistency, which complicate tokenization/normalization and increase sparsity (Attia, 2007; Habash et al., 2018; Alhafni et al., 2024). In this work, we therefore keep preprocessing conservative (minimal normalization) to preserve real-world script variation; this choice better reflects deployed settings but can amplify training noise, making optimizer behavior and calibration effects more consequential.

Why Abjad (Arabic-script) relevance? AbjadNLP targets Arabic and related Arabic-script languages (AbjadNLP Organizers, 2026). Because real-world Arabic-script UGC often contains orthographic and Unicode inconsistencies that complicate normalization/tokenization (Doctor et al., 2022), we study optimizer–calibration behavior under minimal normalization to better reflect deployed conditions.

Despite the central role of QARiB-like

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 26–36
March 28, 2026. ©2026 Association for Computational Linguistics

models in Arabic toxicity detection, optimization choices during fine-tuning remain under-examined in this setting. Adam-style methods (Kingma and Ba, 2015) and AdamW (Loshchilov and Hutter, 2019) dominate NLP practice, yet work in other domains suggests that SGD with momentum can sometimes yield different generalization behavior (Wilson et al., 2017). For imbalanced toxicity datasets such as OffensEval 2020 (Zampieri et al., 2020), the minority OFF class is small but critical, making macro-F_1 and precision–recall-based measures especially relevant. Post-hoc calibration methods such as temperature scaling can adjust predicted confidence without changing score rankings and may interact with optimizer choice and training dynamics (Guo et al., 2017).

We address this gap by systematically comparing Adam, AdamW, and SGD when fine-tuning QARiB for offensive-language detection on two Arabic datasets: OffensEval 2020 (Arabic) and MPOLD. Using a controlled grid of learning rate, weight decay, and warmup settings with matched seeds, we report discrimination and calibration metrics, and analyze how optimizer choice interacts with post-hoc calibration.

Our findings offer evidence-based recommendations for Arabic offensive-language detection, and we provide a reproducible pipeline (to be released upon acceptance) to encourage further investigation of optimizer–calibration interactions in Arabic-script NLP.

Contributions. We:

- provide a controlled optimizer comparison (Adam/AdamW/SGD) for QARiB on two Arabic offensive-language benchmarks (OffensEval20 and MPOLD), jointly reporting discrimination (macro-F_1, AUC-PR_{OFF}, ROC AUC) and calibration (ECE, reliability diagrams);
- quantify sensitivity to common fine-tuning choices (warmup and weight decay) under matched random seeds and selection by validation macro-F_1;
- show that post-hoc temperature scaling yields modest, dataset-dependent ECE changes while preserving ranking-based discrimination (Guo et al., 2017);
- provide a configuration-driven, reproducible pipeline; code/configs will be released upon acceptance.

2 Related Work

Optimizers for Transformers. Adam (Kingma and Ba, 2015) is the default optimizer for pre-trained language models such as BERT (Devlin et al., 2019) because it is easy to tune and robust to noisy gradients. A long-standing debate contrasts adaptive methods with SGD with momentum: adaptive optimizers can converge faster yet sometimes generalize worse than SGD (Wilson et al., 2017; Zhou et al., 2020). For Transformer architectures, heavy-tailed gradient noise and block-wise Hessian structure make SGD systematically underperform Adam-style methods, even when SGD is competitive on other architectures (Zhang et al., 2020, 2024). AdamW decouples weight decay from the adaptive update and yields more predictable regularization (Loshchilov and Hutter, 2019). We adopt this perspective and ask, in an Arabic social-media toxicity setting, how much is gained by deviating from a tuned AdamW baseline (Gkouti et al., 2024; Halfon et al., 2024).

Optimizer choice in NLP fine-tuning. For BERT-like models on small datasets, fine-tuning can be unstable across seeds, with optimization and scheduling choices playing a central role (Mosbach et al., 2021). Systematic studies on GLUE (General Language Understanding Evaluation) and related benchmarks find that, once learning rates are tuned, adaptive optimizers behave similarly and that tuning the learning rate alone often delivers strong baselines (Gkouti et al., 2024; Halfon et al., 2024). Our work aligns with this literature but focuses on a concrete, imbalanced Arabic offensive-language detection task with explicit calibration analysis.

Calibration and evaluation under imbalance. In safety-critical applications, calibrated probabilities matter alongside F_1. Modern neural networks are often overconfident, motivating the use of ECE and temperature scaling (Guo et al., 2017). For LLMs, adaptation steps can degrade calibra-

tion, and auxiliary or adaptive temperature-scaling methods help restore it without harming accuracy (Shen et al., 2024; Xie et al., 2024; Murugesan et al., 2024). For imbalanced tasks such as offensive vs. non-offensive detection, precision–recall analysis is more informative than ROC curves: ROC can be overly optimistic when the positive class is rare, whereas precision–recall curves and AUC-PR better capture minority-class performance (Saito and Rehmsmeier, 2015). We therefore evaluate macro-F_1 and weighted F_1, AUC-PR for the OFF class, and ECE, tying optimizer choice to both discrimination and calibration on Arabic toxicity datasets.

3 Method

3.1 Task and Datasets

We study optimizer choice for Arabic offensive-language detection, formulated as binary text classification with labels `OFF` (offensive) and `NOT` (non-offensive). We evaluate on two Arabic user-generated text datasets that differ in platform and genre: OffensEval 2020 (Arabic) and MPOLD.

OffensEval 2020. We use the Arabic portion of SemEval-2020 Task 12 (OffensEval) and focus on Sub-task A (binary `OFF` vs. `NOT`) (Zampieri et al., 2020; Mubarak et al., 2021). Using the publicly available labeled Arabic data, we construct disjoint stratified train/validation/test split, preserving the class prior in each split.

MPOLD. We use MPOLD (Chowdhury et al., 2020) under a binary mapping to `OFF`/`NOT` and use the fixed train/validation/test split in our experiments (Table 1). Because MPOLD includes multi-platform, user-generated comments and may contain annotation noise, we treat results as benchmark guidance rather than ground-truth for deployment.

We apply minimal preprocessing uniformly across datasets: (i) we preserve emojis, elongations, and non-standard spellings; (ii) we strip URLs and user mentions (platform artifacts); (iii) we normalize whitespace only (no Arabic letter normalization and no diacritics removal beyond what is present).

3.2 Models and Fine-tuning Setup

We fine-tune the Arabic Transformer encoder *QARiB* (Abdelali et al., 2021), a BERT-style model (Devlin et al., 2019) with 12 layers, 12 attention heads, and hidden size 768. We attach a linear classification head on the [CLS] representation and train end-to-end with cross-entropy loss.

Unless otherwise stated, all experiments use batch size 16, maximum sequence length 256, and $E_{\max} = 5$ epochs, with global gradient-norm clipping at 1.0. Weight decay is applied to all non-bias, non-LayerNorm parameters.

Optimization details. For Adam/AdamW we use β_1=0.9, β_2=0.999, ϵ=10^{-8}; for SGD we use momentum 0.9. We evaluate once per epoch and apply early stopping on validation macro-F_1 with patience p=2; we report the best checkpoint per run. Each configuration is run with three random seeds and we report mean ± std.

3.3 Optimizers and Hyperparameter Grid

We compare SGD, Adam, and AdamW (Kingma and Ba, 2015; Loshchilov and Hutter, 2019). Following prior work (Mosbach et al., 2021; Gkouti et al., 2024; Halfon et al., 2024), we first run a coarse learning-rate sweep $(1, 2, 3, 5) \times 10^{-5}$ under fixed weight decay and warmup for each optimizer. Based on validation macro-F_1 and stability, we select $\eta_{\text{AdamW}} = 2 \times 10^{-5}$, $\eta_{\text{SGD}} = 10^{-4}$, and set $\eta_{\text{Adam}} = \eta_{\text{AdamW}}$.

When $\lambda = 0$, AdamW reduces to Adam because the decoupled weight-decay term vanishes (Loshchilov and Hutter, 2019).

The main grid compares:

- **AdamW:** $\eta = 2 \times 10^{-5}$, $\lambda \in \{0, 0.01\}$, $w \in \{0, 0.06\}$;
- **Adam:** same (η, λ, w) as AdamW;
- **SGD:** $\eta = 10^{-4}$, $\lambda = 0.01$, $w \in \{0, 0.06\}$;

with seeds $s \in \{42, 1978, 2025\}$ for all optimizers and for both datasets.

3.4 Learning Rate Schedule and Warmup

All runs use the standard linear warmup + linear decay schedule (Devlin et al., 2019) with

Table 1: Dataset splits and class distribution for OffensEval 2020 (Arabic) and MPOLD under our binary setup (`OFF` vs. `NOT`). Percentages are within each split.

Dataset	Split	NOT	OFF	Total	%OFF
OffensEval20	Train	5,402	1,341	6,743	19.9
	Val	1,158	288	1,446	19.9
	Test	1,158	288	1,446	19.9
MPOLD	Train	2,327	473	2,800	16.9
	Val	499	101	600	16.8
	Test	499	101	600	16.8

warmup ratio $w \in \{0, 0.06\}$; the schedule is held fixed across optimizers (only η, λ, and optimizer dynamics vary).

3.5 Evaluation Metrics and Calibration

Because offensive content is relatively rare, raw accuracy can be misleading. Following best practices for imbalanced classification (Saito and Rehmsmeier, 2015), we report macro-F_1 and weighted F_1 and AUC-PR for the offensive class (AUC-PR_{OFF}), which is more informative than ROC AUC in this regime.

To assess probability calibration, we compute expected calibration error (ECE) (Guo et al., 2017) on the test set for the class probability $p(\text{OFF})$. We partition $p(\text{OFF}) \in [0, 1]$ into M equal-width bins (we use M=15) and compute ECE as the weighted average of the per-bin absolute gap between empirical accuracy and mean confidence.

ECE limitation. ECE is sensitive to the choice of binning scheme and the number of bins, and commonly used plug-in estimators can be biased in finite samples (Roelofs et al., 2022; Kumar et al., 2019). We therefore treat ECE as a comparative diagnostic (with fixed M=15 across all runs) and complement it with reliability diagrams (Section 4.3); for visualization we use quantile binning in the diagrams to reduce sampling artifacts, while keeping equal-width binning for the scalar ECE to preserve comparability with prior work.

Given that modern neural networks tend to be over-confident (Guo et al., 2017), and optimizer choice can affect calibration, we report ECE alongside macro-F_1, weighted F_1, and AUC-PR_{OFF}. All experiments are launched through a configuration-driven script that logs all hyperparameters and metrics to JSON files; code and configuration files will be released after acceptance in an anonymized repository.

4 Results

We report results on **OffensEval20** and **MPOLD** using the same training and evaluation pipeline. For each dataset and optimizer family, we select the *best configuration* by validation macro-F_1 and report test-set metrics as mean (std) over three random seeds $s \in \{42, 1978, 2025\}$. Because `OFF` is the minority class (Table 1), we emphasize macro-F_1 and AUC-PR_{OFF} (alongside ROC AUC and calibration via ECE), following prior recommendations for imbalanced evaluation and probability calibration (Saito and Rehmsmeier, 2015; Guo et al., 2017).

Table 2 summarizes best-per-optimizer test performance for both datasets. Overall, AdamW and Adam are competitive, while SGD lags substantially in minority-class discrimination (macro-F_1 and AUC-PR_{OFF}), especially on MPOLD, and shows higher seed variability (e.g., larger std in AUC-PR_{OFF}).

Calibration reporting convention. For comparability across optimizers, the ECE column in Table 2 reports **temperature-scaled ECE** (ECE (TS)): for each run, we fit a single temperature on the validation set and compute ECE on the test set (Guo et al., 2017). Because temperature scaling does not change the argmax class prediction and preserves score ordering, it leaves macro-F_1 and ranking-based metrics (AUC-PR, ROC AUC) unchanged (Guo et al., 2017); we therefore report the same discrimination metrics alongside ECE (TS).

Selected configurations. Selection by validation macro-F_1 yields: on MPOLD, AdamW (2e-5, WD=0.01, w=0.00), Adam

Table 2: Best-per-optimizer test performance (mean ± std over 3 seeds), per dataset. *ECE (TS)* denotes temperature-scaled expected calibration error.

Dataset	Optimizer	macro-F_1	AUC-PR_{OFF}	ROC AUC	ECE (TS)	Accuracy
MPOLD	AdamW	0.8174 ± 0.0182	0.7812 ± 0.0064	0.9369 ± 0.0041	0.0710 ± 0.0300	0.9028 ± 0.0025
	Adam	0.8356 ± 0.0232	0.7891 ± 0.0148	0.9352 ± 0.0085	0.0640 ± 0.0243	0.9117 ± 0.0060
	SGD	0.6091 ± 0.0277	0.4252 ± 0.0897	0.7408 ± 0.0381	0.2849 ± 0.0048	0.6983 ± 0.0312
OffensEval20	AdamW	0.9305 ± 0.0069	0.9397 ± 0.0046	0.9780 ± 0.0024	0.0413 ± 0.0045	0.9555 ± 0.0052
	Adam	0.9305 ± 0.0069	0.9397 ± 0.0046	0.9780 ± 0.0024	0.0413 ± 0.0045	0.9555 ± 0.0052
	SGD	0.8670 ± 0.0103	0.7971 ± 0.0609	0.9428 ± 0.0069	0.1045 ± 0.0348	0.9112 ± 0.0076

(2e-5, WD=0.00, w=0.06), and SGD (1e-4, WD=0.01, w=0.06); on OffensEval20, Adam/AdamW both select (2e-5, WD=0.00, w=0.00) while SGD selects (1e-4, WD=0.01, w=0.06).

4.1 Minority-class behavior: precision–recall analysis

Figure 1 plots PR curves for the OFF class for the best configuration of each optimizer. As discussed in Section 3.5, PR analysis is especially informative under class imbalance (Saito and Rehmsmeier, 2015). For readability, we plot a representative seed (s=42 when available), while Table 2 reports mean (std) over seeds.

Adam vs. AdamW. On OffensEval20, the best validation setting uses λ=0 for both methods (Selected configurations), so AdamW reduces to Adam when the decoupled weight-decay term is inactive (Loshchilov and Hutter, 2019); accordingly, their PR curves overlap (Figure 1) and test metrics are indistinguishable. On MPOLD, the selected warmup/weight-decay settings differ (Selected configurations), and the PR curves separate slightly, with small but consistent differences in macro-F_1 and AUC-PR_{OFF}.

SGD. SGD yields substantially worse PR curves on both datasets and shows higher variability, matching the larger AUC-PR_{OFF} std in Table 2.

4.2 Sensitivity to weight decay and warmup

To assess robustness, we evaluate Adam and AdamW across weight decay $\lambda \in \{0, 0.01\}$ and warmup ratio $w \in \{0, 0.06\}$. Figure 2 reports validation macro-F_1 (mean ± std over 3 seeds) for each (λ, w) setting on both datasets.

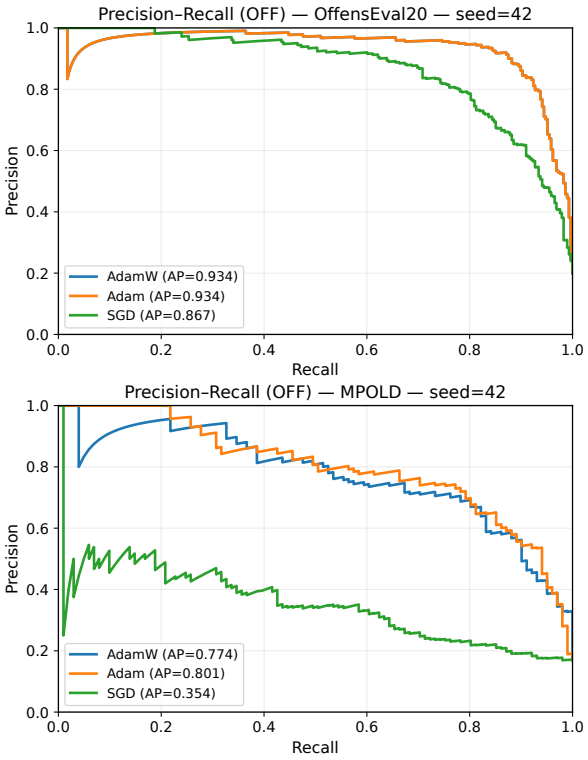
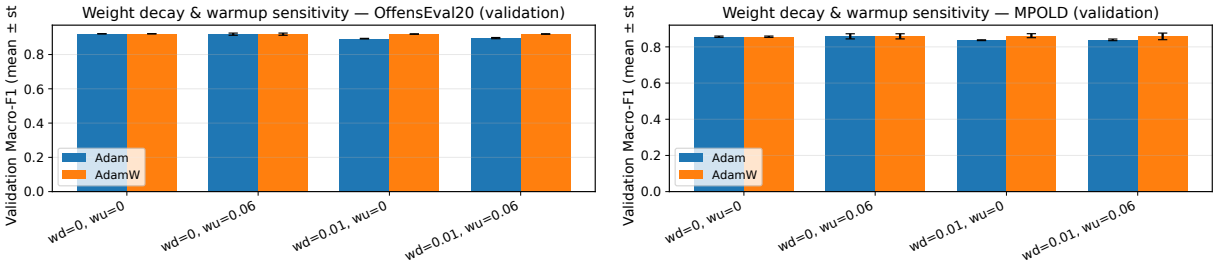


Figure 1: PR curves for OFF (best config per optimizer; representative seed).

Across OffensEval20 and MPOLD, differences are small relative to seed variance and the overall pattern AdamW ≈ Adam remains stable; on MPOLD, AdamW shows a slight advantage in some λ=0.01 settings, consistent with decoupled weight decay affecting optimization when $\lambda > 0$ (Loshchilov and Hutter, 2019).

4.3 Calibration analysis

We assess calibration using ECE and reliability diagrams (Guo et al., 2017), and we also report proper scoring rules (Brier and NLL) computed from saved probabilities (Table 3). Across both datasets, Adam/AdamW are substantially better calibrated than SGD (lower ECE (TS), Brier, and NLL; Tables 2 and 3),



(a) OffensEval20 (b) MPOLD

Figure 2: Validation macro-F_1 surface over weight decay λ and warmup ratio w for Adam and AdamW on OffensEval20 (a) and MPOLD (b). Values report mean±std over three seeds; higher is better. When $\lambda = 0$, AdamW reduces to Adam, explaining overlapping regions.

consistent with the reliability curves in Figure 3.

For Adam/AdamW, the reliability curves are closest to the diagonal at low-to-mid probabilities but fall below it at high p(OFF), indicating over-confidence in high-probability OFF predictions (Figure 3). Temperature scaling is a monotonic logit re-scaling and thus preserves ranking metrics (Guo et al., 2017); in our runs it yields modest, dataset-dependent ECE shifts. For AdamW (3 seeds), Raw→TS ECE changes are small (OffensEval20: 0.0450→0.0415; MPOLD: 0.0686→0.0727), so we interpret scalar ECE alongside the diagrams and scoring rules (Table 3).

Qualitative examples. AdamW often avoids overconfident SGD errors on noisy Arabic text: for a NON-OFFENSIVE example (gold=0), AdamW assigns p_{OFF}=0.0018 vs. SGD 0.5784; for an OFFENSIVE example (gold=1), AdamW assigns 0.9966 vs. SGD 0.5005 (see Table 4).

Validation-set threshold tuning (F_1-OFF). Using the saved validation probabilities for each seed, we swept a decision threshold $t \in [0, 1]$ to maximize validation F_1-OFF, then applied the resulting t^* to the corresponding test probabilities. Table 5 shows that this post-processing step yields small and optimizer-dependent effects on OffensEval20: SGD benefits slightly on average, whereas Adam/AdamW do not consistently improve, indicating that the default 0.5 threshold is already near-optimal for the stronger optimizers in this setting.

5 Conclusion

We studied how optimizer choice affects fine-tuning QARiB for Arabic offensive-language detection under class imbalance on two benchmarks (OffensEval20 and MPOLD), reporting mean (std) test performance over three random seeds.

AdamW and Adam were consistently strong and closely matched when tuned, whereas SGD substantially underperformed under the same tuning budget and showed higher seed sensitivity, especially on minority-class metrics. On OffensEval20, the selected best settings use λ=0, for which AdamW reduces to Adam, explaining the overlapping behavior.

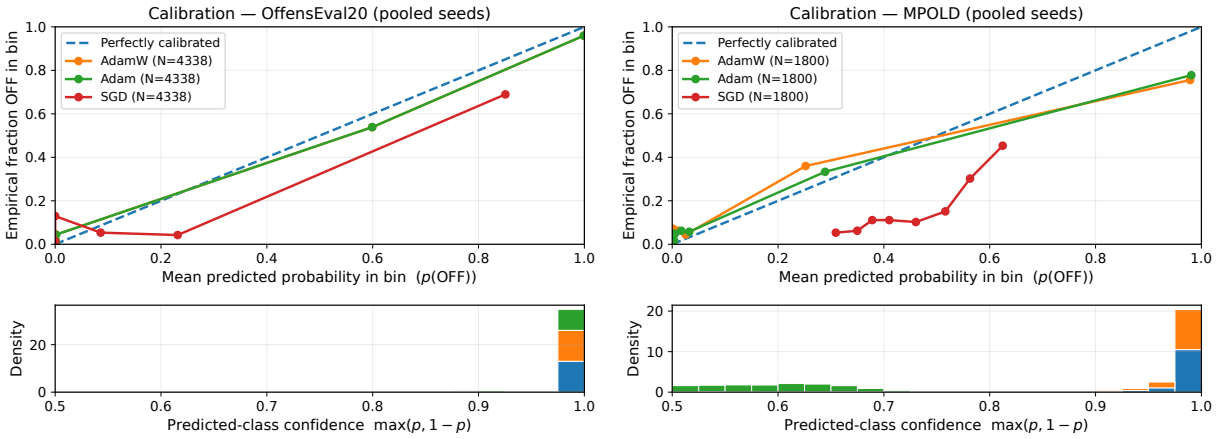
Calibration analysis shows non-trivial miscalibration across optimizers. Temperature scaling provides a practical post-hoc adjustment with modest, dataset-dependent changes in ECE while preserving ranking-based discrimination; we therefore interpret scalar ECE together with reliability diagrams and complementary proper scoring rules (Brier score and negative log-likelihood, NLL).

Practical takeaways. In our setting, we recommend the following for Arabic-script offensive-language detection:

- **Optimizer choice:** Start from a tuned AdamW or Adam baseline. SGD is a weaker choice under the same tuning budget and shows higher seed sensitivity.
- **Evaluation under imbalance:** Prioritize macro-F_1 and precision–recall–based metrics for the minority OFF class (e.g., AUC-PR_{OFF}) alongside accuracy and ROC AUC.

Table 3: Proper scoring-rule calibration metrics on the test set (mean±std over 3 seeds). Lower is better.

Dataset	Optimizer	Brier	NLL	N
MPOLD	AdamW	0.0830 ± 0.0088	0.3846 ± 0.1249	3
	Adam	$\mathbf{0.0791 \pm 0.0038}$	$\mathbf{0.3469 \pm 0.0851}$	3
	SGD	0.2059 ± 0.0041	0.6022 ± 0.0084	3
OffensEval20	AdamW	$\mathbf{0.0443 \pm 0.0037}$	$\mathbf{0.2201 \pm 0.0121}$	3
	Adam	$\mathbf{0.0443 \pm 0.0037}$	$\mathbf{0.2201 \pm 0.0121}$	3
	SGD	0.0844 ± 0.0014	1.3994 ± 0.9427	3



(a) OffensEval20 (b) MPOLD

Figure 3: Calibration summary on OffensEval20 (a) and MPOLD (b), pooled over three seeds. **Top:** reliability diagram for $p(\text{OFF})$ using quantile bins; perfect calibration follows the diagonal. **Bottom:** stacked histogram of predicted-class confidence $\max(p, 1-p)$, showing how probability mass concentrates at low/high confidence.

- **Calibration for deployment:** When confidence scores are needed—e.g., for thresholding or human-in-the-loop triage—apply post-hoc temperature scaling and validate calibration using both ECE and reliability diagrams (not scalar ECE alone).
- **Threshold tuning:** The default 0.5 decision threshold is not universally optimal under class imbalance. When deployment requires a specific operating point (e.g., prioritizing OFF recall), tune the threshold on a held-out validation set and report its impact on test performance.

We also include qualitative examples to contextualize common confidence patterns and errors under different optimizers.

Future work should test the generality of these findings across more Arabic-script languages, domains, and model families, including parameter-efficient tuning.

6 Limitations

Generalizability. Our experiments evaluate QARiB fine-tuning on two Arabic offensive-language benchmarks: (i) the Arabic subset of OffensEval 2020, primarily tweet-based and tied to a specific collection period and annotation context (Zampieri et al., 2020; Mubarak et al., 2021), and (ii) MPOLD, a multi-platform Arabic news-comment dataset with platform- and domain-specific language patterns (Chowdhury et al., 2020). Accordingly, the observed discrimination and calibration trade-offs may not transfer to other platforms, genres, or later time periods where topic mix, dialect coverage, and community norms shift. More broadly, these effects may vary across Arabic PLM pretraining corpora and tokenization/normalization choices used in other model families.

Model and optimizer scope. We focus on an encoder-only Transformer (QARiB with a standard classification head) and compare Adam, AdamW, and SGD (Kingma and Ba,

Table 4: Qualitative examples illustrating probability gaps ($\Delta p = |p_{\text{off}}^{\text{AdamW}} - p_{\text{off}}^{\text{SGD}}|$) between AdamW and SGD on noisy Arabic social media text (raw, unscaled probabilities).

Text (truncated Arabic)	**Gold**	**AdamW p_{off}**	**SGD p_{off}**
... لا مش غلط يا كابتن محمود يا كبير ف ناس تعبانه بدني محتاجه @USER @USER	0	0.0018	0.5784
... هههههههه يلا مش مشكله يا روضه يا حبيبتي @USER @USER @USER @USER: RT	0	0.0023	0.5090
... وطلع كيونا بيطلع من الجيش وشيومين بيروح ف نفس اليوم حسبي الله	1	0.9966	0.5005
... إيش دخل الوهابية ياهووو دي بتاعتنا في السعودية و راضيين @USER @USER	1	0.9844	0.7571

Table 5: OffensEval20 validation threshold tuning for F_1-OFF (mean±std over 3 seeds). Adam (not shown) matches AdamW.

Opt.	t^*	F_1@0.5	F_1@t^*	Δ
AdamW	0.3410 ± 0.5660	0.8747 ± 0.0150	0.8669 ± 0.0177	-0.0079 ± 0.0307
SGD	0.5348 ± 0.0088	0.7903 ± 0.0157	0.7961 ± 0.0067	$+0.0058 \pm 0.0090$

2015; Loshchilov and Hutter, 2019). Results may differ for other Arabic PLMs and pretraining regimes (e.g., AraBERT (Antoun et al., 2020), MARBERT (Abdul-Mageed et al., 2021), or other QARiB variants (Abdelali et al., 2021)), which can change tokenization, domain coverage, and fine-tuning dynamics. Other training recipes (e.g., alternative schedules/regularization, larger models, adapters/LoRA), or optimizers (e.g., Adafactor, Lion, quasi-second-order methods) could yield different behavior. Thus, any optimizer "ranking" is conditional on this model family and fine-tuning regime.

Statistical limits. We report test metrics as mean (std) over three random seeds ($s \in \{42, 1978, 2025\}$). While this improves robustness over single-run reporting, three seeds provide limited power for strong claims, especially for calibration. For PR curves we visualize a representative seed for readability, while tables remain seed-aggregated; PR analysis is particularly informative under class imbalance (Saito and Rehmsmeier, 2015). For reliability diagrams we pool predictions across seeds *after* selecting the best run per (optimizer, seed) by validation macro-F_1 , which increases sample size per curve. Even so, small std does not guarantee robustness under broader initializations or wider hyperparameter ranges. Additionally, validation-based threshold selection can overfit: in OffensEval20 we observe high seed-to-seed variability in t^* for Adam/AdamW, and corresponding non-monotonic changes on test, so threshold tuning should be treated as a deployment-time choice requiring careful validation.

Calibration dependence. ECE and reliability plots depend on binning design and on the probability target (here, p(OFF)); common ECE estimators can also be sensitive in finite samples. To reduce reliance on binning-based ECE alone, we also report Brier score and NLL as complementary proper scoring rules. Consequently, calibration differences between optimizers (and the apparent effect size of temperature scaling) should be viewed as indicative. Temperature scaling preserves prediction rankings and can improve likelihood-based calibration, but its effect on ECE can be modest and dataset-dependent under different binning/aggregation choices (Guo et al., 2017; Roelofs et al., 2022; Kumar et al., 2019). Stronger conclusions would require more seeds, binning-sensitivity analysis, and cross-domain evaluation.

Ethical use. Automatic offensive-language detection is high-impact: errors can lead to over-filtering or under-detection, and performance may vary across dialects and writing styles that are under-represented in training data (Mubarak et al., 2021). Deploying language technologies in moderation settings can introduce or amplify harms without transparency and appropriate human oversight (Bender et al., 2021; Gallegos et al., 2023). Accordingly, the models studied here are not intended as stand-alone moderation systems; they should be treated as components in a broader pipeline with documented policies, human review, and appeal mechanisms.

References

Ahmed Abdelali, Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2021. QARiB: QCRI arabic and dialectal BERT. *arXiv preprint*.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

AbjadNLP Organizers. 2026. Abjadnlp workshop. `https://wp.lancs.ac.uk/abjad/`. Accessed: 2025-12-18.

Bashar Alhafni, Sarah Al-Towaity, Ziyad Fawzy, Fatema Nassar, Fadhl Eryani, Houda Bouamor, and Nizar Habash. 2024. Exploiting dialect identification in automatic dialectal text normalization. *arXiv preprint*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Mohammed Attia. 2007. Arabic tokenization system. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 65–72, Prague, Czech Republic. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*.

Sabit Hassan Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-Gyo Jung, Bernard J. Jansen, and Joni Salminen. 2020. A multi-platform arabic news comment dataset for offensive language detection. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Raiomond Doctor, Alexander Gutkin, Cibu Johny, Brian Roark, and Richard Sproat. 2022. Graphemic normalization of the perso-arabic script. *arXiv preprint arXiv:2210.12273*.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Nefeli Gkouti, Prodromos Malakasiotis, Stavros Toumpis, and Ion Androutsopoulos. 2024. Should I try multiple optimizers when fine-tuning a pre-trained transformer for NLP tasks? should I tune their hyperparameters? In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2555–2574, St. Julian's, Malta. Association for Computational Linguistics.

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*, pages 1321–1330.

Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Nadir Habash, Fahad Al-Obaidli, Mahmoud Al-Taie, and Wajdi Zaghouani. 2018. Unified guidelines and resources for arabic dialect orthography. In

Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Alon Halfon, Shai Gretz, Ofir Arviv, Artem Spector, Orith Toledo-Ronen, Yoav Katz, Liat Ein-Dor, Michal Shmueli-Scheuer, and Noam Slonim. 2024. Stay tuned: An empirical study of the impact of hyperparameters on LLM tuning in real-world applications. *arXiv preprint*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Ananya Kumar, Percy Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations (ICLR)*.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2021. Arabic offensive language on twitter: Analysis and experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 126–135, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Balamurali Murugesan, Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. 2024. Robust calibration of large vision–language adapters. *arXiv preprint*.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C. Mozer. 2022. Mitigating bias in calibration error estimation. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Takaya Saito and Marc Rehmsmeier. 2015. The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432.

Maohao Shen, Subhro Das, Kristjan Greenewald, Prasanna Sattigeri, Gregory W. Wornell, and Soumya Ghosh. 2024. Thermometer: Towards universal calibration for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 44687–44711.

Ashia C. Wilson, Rebecca Roelofs, Mitchell Stern, Nathan Srebro, and Benjamin Recht. 2017. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*.

Johnathan Xie, Annie S. Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. 2024. Calibrating language models with adaptive temperature scaling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18128–18138, Miami, Florida, USA. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J. Reddi, Sanjiv Kumar, and Suvrit Sra. 2020. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems*.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo.

2024. Why transformers need Adam: A hessian perspective. In *Advances in Neural Information Processing Systems*, volume 37.

Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Hoi, and Weinan E. 2020. Towards theoretically understanding why SGD generalizes better than Adam in deep learning. In *Advances in Neural Information Processing Systems*.