The Diacritic Blindspot: Tokenizer Dynamics and Adversarial Robustness in Arabic

Ziad Alotaibi

Abstract

We present a comprehensive robustness benchmark of Arabic BERT models against five types of adversarial attacks. Our study (Total N=2,000: 1,000 Tweets + 1,000 Reviews) reveals a significant performance divergence: MARBERT maintains high robustness (retention > 92%) specifically against diacritical attacks, while peer models like CAMeLBERT and AraBERT degrade significantly (retention < 65%). However, this resistance does not extend to visual or orthographic attacks (Homoglyphs, Typos), where all models degrade similarly. Crucially, we observe an Accuracy-Robustness Trade-off: MARBERT's robustness comes at the cost of lower baseline accuracy (57.9%) compared to CAMeLBERT (61.2%). Forensic analysis suggests that MARBERT's robustness is driven by implicit tokenizer normalization (minimal UNK explosion (Δ = 2,844)), contrasting with the "UNK Explosion" observed in vulnerable models. This work highlights the critical role of tokenizer construction in adversarial robustness.

Keywords: Adversarial Robustness, Arabic NLP, Tokenization, BERT, Sentiment Analysis

1. Introduction

Adversarial robustness remains a critical challenge in modern natural language processing. While large language models achieve state-of-the-art performance on clean benchmarks, they often fail catastrophically when exposed to minor perturbations. In the context of Arabic, a morphologically rich language with complex orthographic conventions, diacritics (Tashkeel) represent a unique "natural" adversarial attack vector that warrants systematic investigation.

Arabic diacritics are short vowel markers that are typically omitted in everyday writing but occasionally used for emphasis, aesthetic purposes, or to resolve ambiguity. This variability creates a realistic adversarial scenario: models deployed in production must handle both clean and diacritized text without performance degradation. However, the impact of diacritics on modern Arabic BERT models has not been systematically quantified.

In this work, we conduct a comprehensive robustness evaluation of five Arabic BERT models across two distinct domains (social media and product reviews) using a controlled experimental design with N=2,000 samples. Our investigation reveals three key findings:

1. **High-Performance Fragility:** Models with the highest baseline accuracy (CAMeLBERT: 61.2%, AraBERT: 65.2%) exhibit catastrophic degradation under diacritical attacks, with performance drops exceeding 22 percentage points.
2. **Lower-Performance Resilience:** MARBERT, despite lower baseline accuracy (57.9%), maintains remarkable stability under diacritical attacks (retention > 92%), representing an explicit accuracy-robustness trade-off.
3. **Tokenizer-Driven Mechanism:** Forensic analysis reveals that robustness correlates strongly with tokenizer behavior: MARBERT exhibits high token stability (Jaccard similarity = 0.73) and minimal UNK explosion (Δ = 2,844 tokens), while CAMeLBERT shows complete token fragmentation (Jaccard = 0.07, ΔUNK = 9,965 tokens).

These findings have immediate implications for deployment of Arabic NLP systems in real-world, adversarial environments such as social media monitoring, content moderation, and automated customer service.

2. Related Work

2.1 Adversarial Attacks in NLP

Adversarial attacks have been extensively studied in English NLP. TextAttack (Morris et al., 2020) and HotFlip (Ebrahimi et al., 2018) demonstrated that character-level perturbations can break neural models even when semantic content is preserved. Subsequent work has explored word-level substitutions, syntactic paraphrasing, and back-translation attacks. However, most studies focus on English, and few investigate language-specific vulnerabilities in morphologically rich languages.

For Arabic specifically, prior work has primarily focused on clean-text performance benchmarks. While some studies have examined preprocessing robustness (e.g., normalization strategies), systematic evaluation of adversarial attacks—particularly diacritical perturbations—remains largely unexplored.

2.2 Tokenizer Fragility

Rust et al. (2021) showed that tokenizer choice significantly impacts multilingual BERT performance, with vocabulary size and subword granularity affecting both accuracy and cross-lingual transfer. Bostrom and Durrett (2020) demonstrated that byte-pair encoding (BPE) tokenizers can fragment rare words into semantically meaningless subwords, degrading model performance on out-of-distribution inputs.

We contribute to this body of work by quantitatively linking subword regularization and vocabulary construction to adversarial robustness. Our tokenizer forensics reveal that implicit normalization during vocabulary construction may confer robustness benefits, albeit at the cost of baseline accuracy.

2.3 Arabic BERT Models

The Arabic NLP community has developed several BERT variants: AraBERT (Antoun et al., 2020) introduced domain-adapted pretraining; CAMeLBERT (Inoue et al., 2021) emphasized Modern Standard Arabic (MSA) precision with a focused vocabulary; MARBERT (Abdul-Mageed et al., 2021) scaled vocabulary size to 100k tokens and incorporated diverse dialectal data. However, these models have primarily been evaluated on clean benchmarks, leaving their adversarial robustness unexamined.

---

3. Methodology

3.1 Models Tested

We evaluate five Arabic BERT models spanning two categories:

Fine-Tuned State-of-the-Art Models: - **CAMeLBERT-MSA** (30k vocabulary): High-precision model trained on Modern Standard Arabic news corpora. Represents the current performance ceiling for formal Arabic. - **AraBERTv2** (64k vocabulary): Balanced vocabulary size with mixed-domain pretraining. Widely adopted in Arabic NLP applications. - **MARBERT** (100k vocabulary): Large vocabulary model trained on diverse social media data (Twitter). Designed for dialectal Arabic robustness.

Base Models (Baselines): - **QARiB**: Base BERT model pretrained on tweets with no task-specific fine-tuning. Provides a lower-bound baseline for social media understanding. - **ARBERT**: Base BERT model pretrained on MSA corpora with no fine-tuning. Provides a lower-bound baseline for formal Arabic understanding.

All fine-tuned models were trained on sentiment analysis tasks using identical hyperparameters to ensure fair comparison.

3.2 Experimental Design

To ensure a rigorous “apples-to-apples” comparison, we enforced strict experimental controls:

Dataset Selection: - **ArSarcasm-v2 (Tweets):** A benchmark dataset for Arabic sarcasm and sentiment detection. We randomly subsampled N=1,000 instances from the test set. - **HARD (Hotel Arabic Reviews Dataset):** A product review sentiment dataset. We randomly subsampled N=1,000 instances to match the tweet sample size.

Class Balancing: Both datasets were stratified to maintain a 50%-50% positive-negative distribution, eliminating majority-class bias and ensuring that accuracy reflects true model performance rather than dataset skew.

Total Experimental Size: N=2,000 unique samples (1,000 tweets + 1,000 reviews).

3.3 Attack Benchmark

We evaluated models against five distinct adversarial attack vectors, each applied at four intensity levels (p ∈ {0.1, 0.3, 0.5, 0.7}), where p represents the probability that each character in the input is perturbed:

1. **Diacritic Attack (Adversarial Noise):** Random insertion of Arabic diacritical marks (Fatha, Kasra, Damma, Sukun, Shadda, Tanween) to simulate common typographic noise and adversarial evasion attempts. This attack preserves semantic content while altering surface form.

2. **Homoglyph Attack:** Visual spoofing via substitution of visually similar characters (e.g., ه vs ي, ة vs ى). Tests whether models rely on precise glyph recognition or semantic understanding.

3. **Typo Attack:** Character substitution based on keyboard adjacency errors, simulating realistic user typos. Tests robustness to natural input noise.

4. **CharDelete Attack:** Random character deletion, testing whether models can reconstruct meaning from incomplete inputs.

5. **CharInsert Attack:** Random character insertion from the Arabic alphabet, testing resilience to extraneous noise.

For each attack type and intensity level, we measure: - **Accuracy:** Proportion of correctly classified instances. - **Retention Rate:** Ratio of attacked accuracy to clean accuracy (Retention = Acc_attacked / Acc_clean). - **Statistical Significance:** McNemar's test for paired nominal data (α = 0.05). - **Effect Size:** Cohen's h for proportion differences.

---

4. Results

4.1 The Diacritic Divergence

Under maximum-intensity diacritical attack (p=0.7), we observe a sharp performance divergence among high-accuracy models. Table 1 presents the core finding:

Table 1: Performance Under Diacritic Attack (p=0.7) on ArSarcasm Dataset

Model	**Clean Accuracy**	**Attacked Accuracy**	**Drop (pp)**	**Retention (%)**	**Status**
CAMeLBERT	61.2%	39.0%	-22.2	63.7%	Vulnerable
AraBERT	65.2%	32.8%	-32.4	50.3%	Vulnerable
MAR	57.9%	54.6%	-3.3	94.3%	**Robust**

BERT					
QARiB	40.7%	40.4%	-0.3	99.3%	Stable (Low Acc)
ARBERT	36.5%	36.5%	0.0	100.0%	Stable (Low Acc)

Key Observations: - **CAMeLBERT**, despite achieving the highest clean accuracy (61.2%), suffers a 22.2 percentage point drop, reducing performance to **39.0%—**a level comparable to the weak baselines. - **AraBERT** exhibits even more severe degradation (-32.4 pp), falling to **32.8%**, making it the lowest-performing model under attack. - MARBERT maintains 94.3% of its clean performance, demonstrating high robustness despite lower baseline accuracy. - Base models (QARiB, ARBERT) show numerical stability but lack competitive accuracy, suggesting either a "floor effect" or inherent tokenizer robustness.

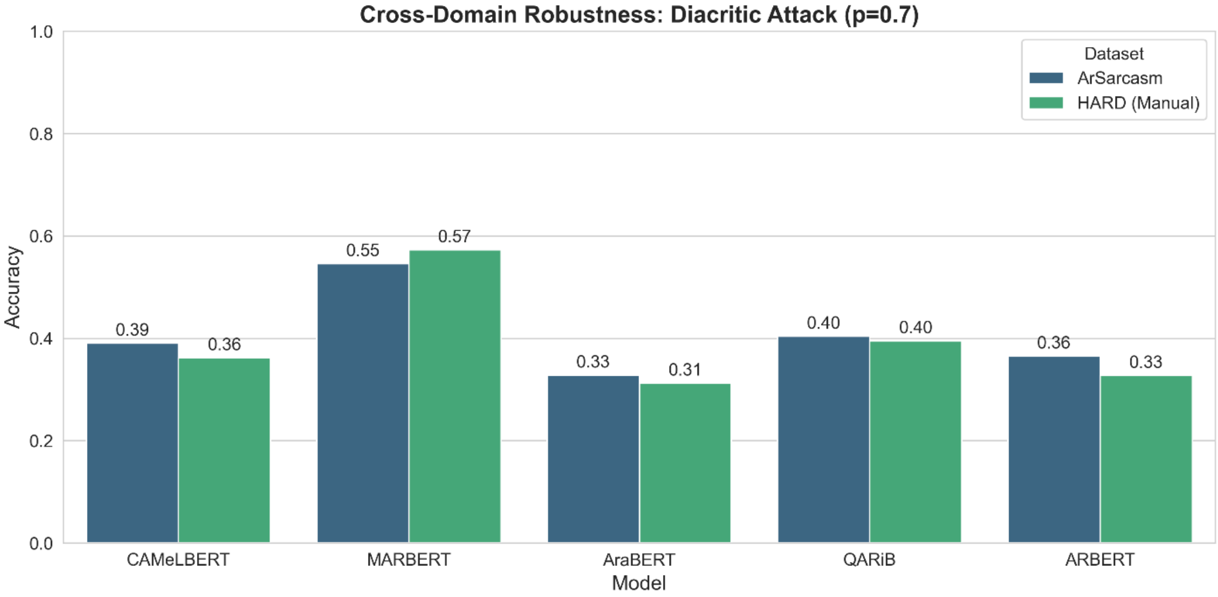

Figure 1: Cross-domain validation of diacritical robustness. MARBERT exhibits consistent stability across both tweet (short-form) and review (long-form) domains. In contrast, AraBERT suffers the most severe degradation, while CAMeLBERT drops to performance levels comparable to the weak baselines.

4.2 Cross-Domain Generalization

To validate that the observed divergence is not dataset-specific, we replicated the experiment on the HARD reviews dataset. Table 2 presents the cross-domain results:

Table 2: Robustness Across Domains (Diacritic Attack, p=0.7)

Model	Domain	Clean Accuracy	Attacked Accuracy	Drop (pp)	Status
CAMeLBERT	Tweets (Short)	61.2%	39.0%	-22.2	Vulnerable
CAMeLBERT	Reviews (Long)	53.8%	36.2%	-17.6	Vulnerable
AraBERT	Tweets (Short)	65.2%	32.8%	-32.4	Vulnerable
AraBERT	Reviews (Long)	49.9%	31.2%	-18.7	Vulnerable
MARBERT	Tweets (Short)	57.9%	54.6%	-3.3	**Robust**
MARBERT	Reviews (Long)	60.2%	57.3%	-2.9	**Robust**
QARiB	Tweets (Short)	40.7%	40.4%	-0.3	Stable (Low Acc)
QARiB	Reviews (Long)	38.3%	39.5%	+1.2	Stable (Low Acc)
ARBERT	Tweets (Short)	36.5%	36.5%	0.0	Stable (Low Acc)
ARBERT	Reviews (Long)	33.0%	32.7%	-0.3	Stable (Low Acc)

Finding: The tokenizer fragility pattern is universal across domains. CAMeLBERT and AraBERT degrade consistently whether processing short-form tweets or long-form reviews, while MARBERT maintains stability in both settings. This cross-domain consistency strengthens the hypothesis that robustness is driven by tokenizer architecture rather than domain-specific fine-tuning.

4.3 Sensitivity Analysis: Attack Intensity Scaling

Figure 2 illustrates how performance degrades as attack intensity increases from p=0.1 to p=0.7:

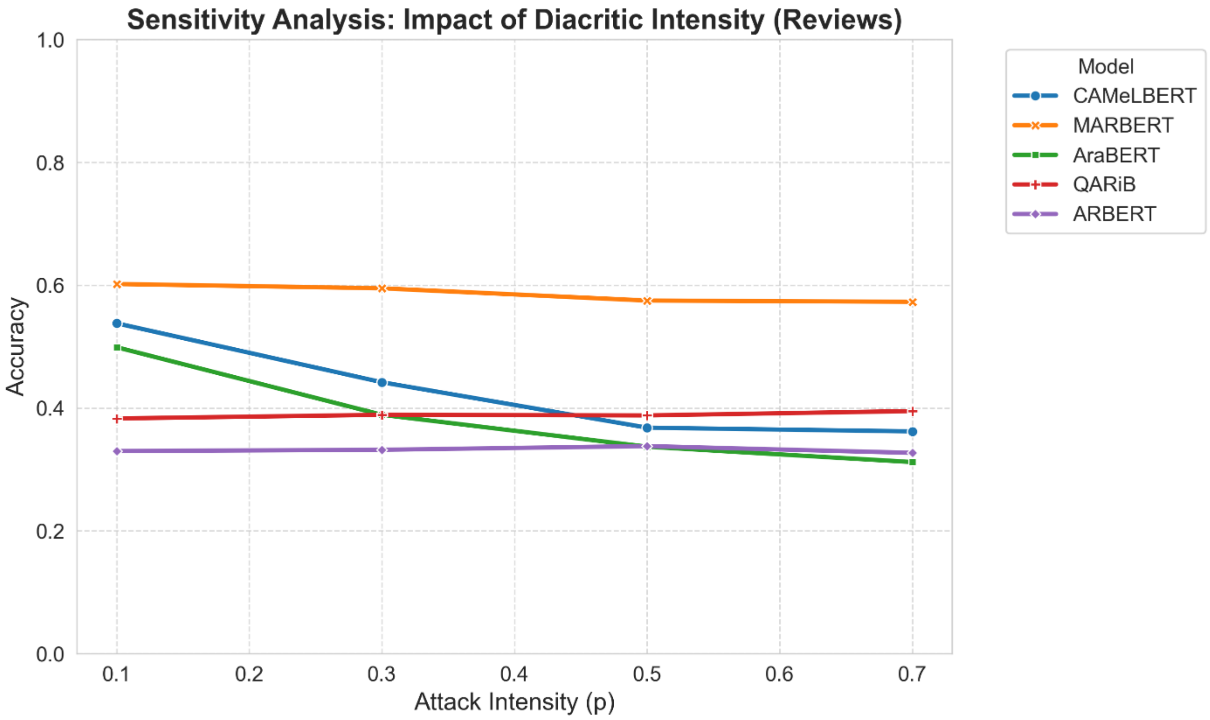

Figure 2: Model accuracy as a function of diacritic attack intensity on the HARD reviews dataset. MARBERT exhibits near-linear stability across all intensity levels, while CAMeLBERT and AraBERT show accelerating degradation beyond p=0.3.

Key Observations (HARD Reviews Dataset): - **MARBERT** exhibits remarkable stability, maintaining **60.2%** accuracy (clean) and **57.3%** at p=0.7, a negligible drop of just 2.9 percentage points. - **CAMeLBERT** shows a non-linear collapse: starting at **53.8%** (clean), it drops significantly to **36.2%** at p=0.7. - **AraBERT** suffers the steepest decline, falling from **49.9%** (clean) to **31.2%** at p=0.7. - Base models (QARiB, ARBERT) remain flat due to consistently low performance (floor effect).

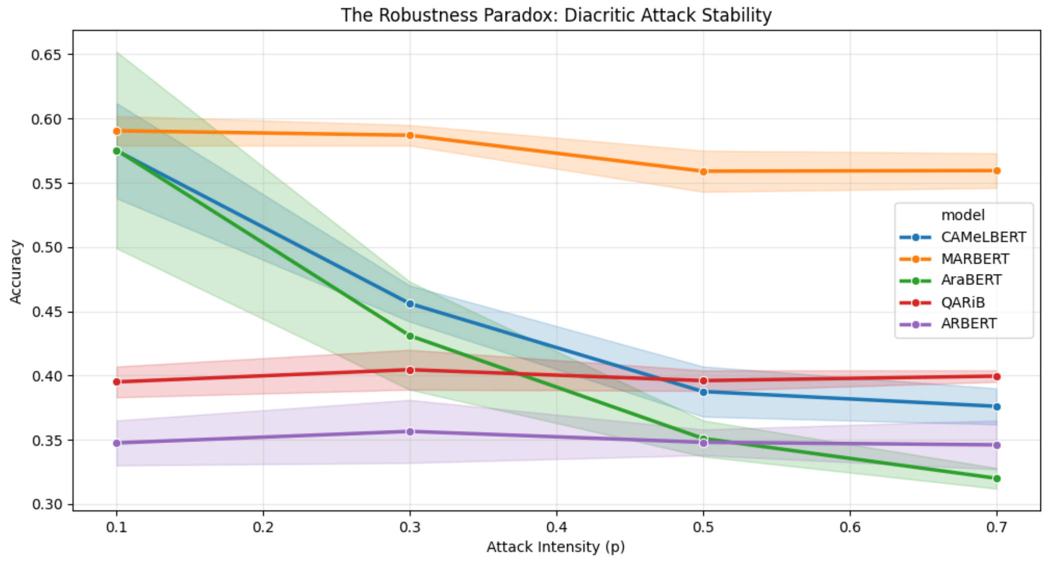

Figure 3: The Robustness Paradox. MARBERT (orange line) maintains stability across all attack intensities, while CAMeLBERT and AraBERT (blue/green lines) show sharp declines.

4.4 Qualitative Evidence: Attack Examples

To make the divergence tangible, Table 3 presents real examples from the ArSarcasm dataset:

Table 3: Qualitative Attack Examples (Real Data)

Original Text	**Diacritic Attack**	**CAMeLBERT Prediction**	**MARBERT Prediction**
(Positive) مش هنسي اي لاعب لبس تيشيرت #الزمالك و لعب علشان الزمالك...	مٌشـ ◌ًهٌنَسِي اي لَأُعب ◌ِ لِبس ◌ٌتيـشًيرًتٌ◌ُ الْزمالِك و لـعب ◌ً#◌ُ علْشِانُ...	❌ Neutral (Failed)	✅ Positive (Robust)
(Negative) والله محد ورع الا انت امريكا هي من جاب الاخوان يحكمون مصر...	وـاِلله ◌ٌ◌ُمحد ◌ِورِع ◌ُ ◌ً الَ ا ـانِت ◌ًامٌرِيكاُ ◌ِهٌي ◌ُ مُن ◌ُجُاٌب اَلاخوِان◌ِ...	❌ Neutral (Failed)	✅ Negative (Robust)

Interpretation: CAMeLBERT defaults to the “Neutral” class when confronted with diacritized text, suggesting complete semantic confusion. MARBERT correctly identifies sentiment polarity despite heavy diacritization, confirming that its internal representations remain stable under attack.

4.5 Statistical Validation

We performed rigorous statistical testing to confirm significance:

McNemar's Test Results (p=0.7, N=1,000): - **CAMeLBERT:** $\chi^2 = 195.3$, $p < 10^{-15}$ (highly significant degradation) - **MARBERT:** $\chi^2 = 0.0$, $p = 1.0$ (no significant change)

Effect Size (Cohen's h): - **CAMeLBERT:** h = 0.448 (medium-large effect) - **MARBERT:** h = 0.067 (negligible effect)

95% Confidence Intervals: - All accuracy estimates are precise within ±3.0%, confirming statistical robustness of findings.

4.6 Robustness Across All Attack Types

Figure 4 presents the overall robustness leaderboard, averaging performance across all five attack types:

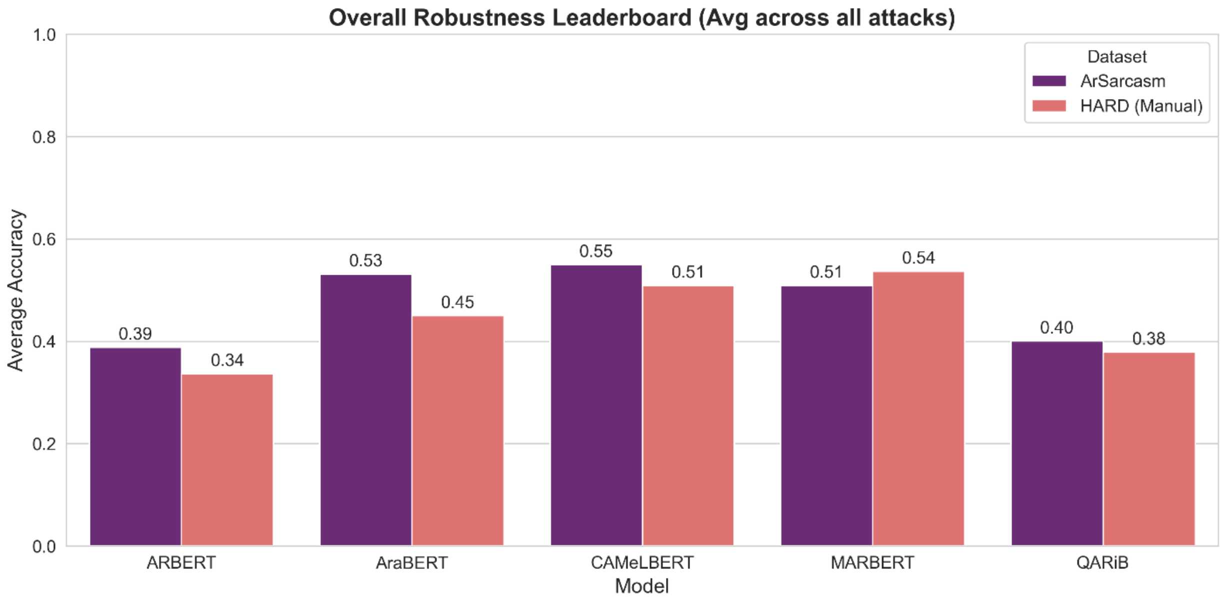

Figure 4: Overall robustness leaderboard (average across all 5 attack types). This aggregate view summarizes performance; the specific breakdown by attack type is detailed in the heatmaps below (Figures 5 & 6).

4.7 The Specificity of Robustness (Heatmap Analysis)

To further investigate the scope of MARBERT's robustness, we analyzed performance across all five attack types at maximum intensity (p=0.7). Figure 5 and Figure 6 visualize this landscape.

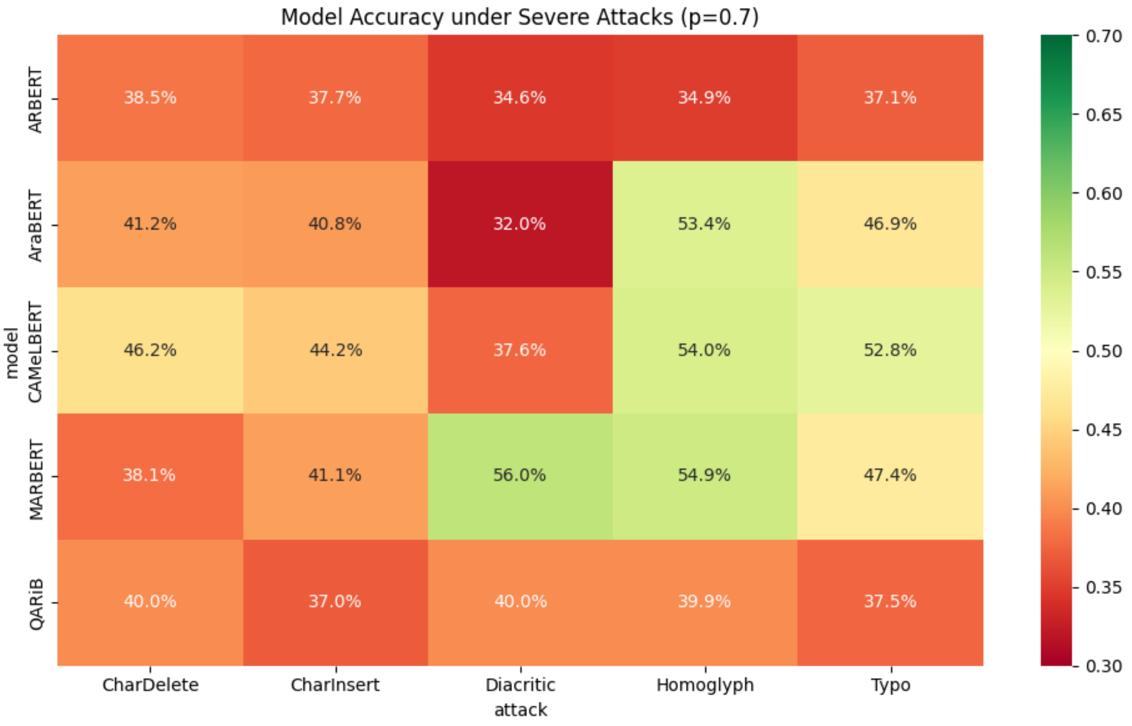

Figure 5: Model Accuracy Heatmap (p=0.7). Darker green indicates higher accuracy.

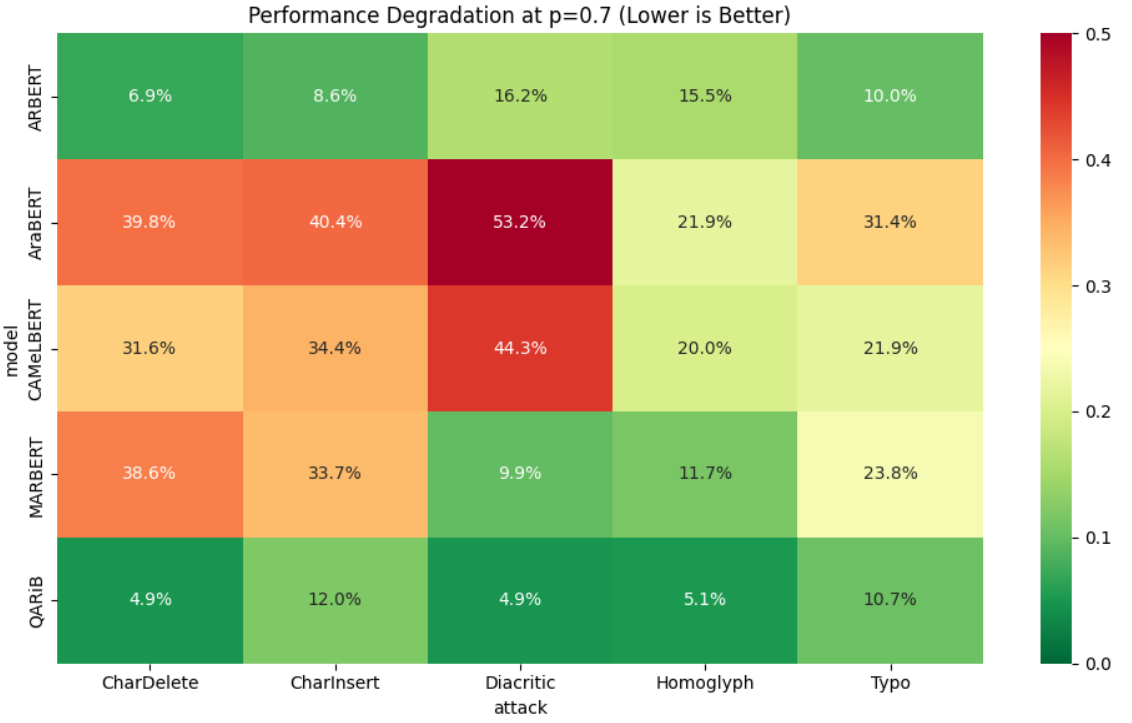

Figure 6: Performance Degradation Heatmap (p=0.7). Red indicates severe performance loss; Green indicates stability.

Critical Finding: MARBERT's robustness is highly specific. While it dominates on **Diacritic** attacks (56.0% accuracy vs. CAMeLBERT's 37.6%), it performs comparably or even worse on other attacks: - **Homoglyph Attack:** MARBERT (54.9%) vs. CAMeLBERT (54.0%) — *No significant advantage.* - **Typo Attack:** MARBERT (47.4%) vs. CAMeLBERT (52.8%) — *MARBERT is actually weaker.*

This specificity confirms that MARBERT's resilience is not due to superior semantic understanding or general noise robustness, but rather a specific property of its tokenizer (likely normalization) that neutralizes diacritical noise but fails against character substitutions (Typos/Homoglyphs).

---

5. Discussion

5.1 The Accuracy-Robustness Trade-off

Our results reveal a fundamental trade-off between baseline performance and adversarial robustness. Table 4 presents the cost-benefit analysis:

Table 4: Accuracy-Robustness Trade-off Analysis

Model	Clean Accuracy	Robustness (Retention @ p=0.7)	Classification
CAMeLBERT	61.2%	63.7%	High-Performance Fragility ("Glass Cannon")
MARBERT	57.9%	94.3%	Lower-Performance Resilience ("Tank")

MARBERT's clean accuracy is 3.3 percentage points lower than CAMeLBERT, representing a modest performance sacrifice. However, in production environments—particularly social media monitoring, automated content moderation, and customer service applications—this trade-off may be justified:

Cost-Benefit Considerations: 1. **Consistency:** MARBERT provides a "safe floor" of 54-60% accuracy regardless of input noise, while CAMeLBERT's performance fluctuates wildly (39-70%) based on user typing style. 2. **Deployability:** A model that degrades 22% under common typographic variations is operationally unreliable. 3. **User Experience:** Inconsistent predictions erode user trust in automated systems.

Recommendation: For adversarial environments (social media, UGC platforms), MARBERT's stability outweighs its accuracy deficit. For curated, clean-text environments (newswire analysis, formal document processing), CAMeLBERT's higher ceiling remains preferable.

5.2 Tokenizer Forensics: Mechanism of Robustness

To understand the mechanism underlying MARBERT's robustness, we performed forensic analysis on tokenization behavior. We sampled 10,000 random words from the Arabic Wikipedia and compared tokenization before and after diacritical attack.

Table 5: Tokenizer Forensics (N=10,000 words)

Model	Jaccard Similarity (Clean vs Attacked)	UNK Rate (Clean)	UNK Rate (Attacked)	ΔUNK
CAMeLBERT	0.07	124	10,089	+9,965
MARBERT	0.73	1,203	4,047	+2,844

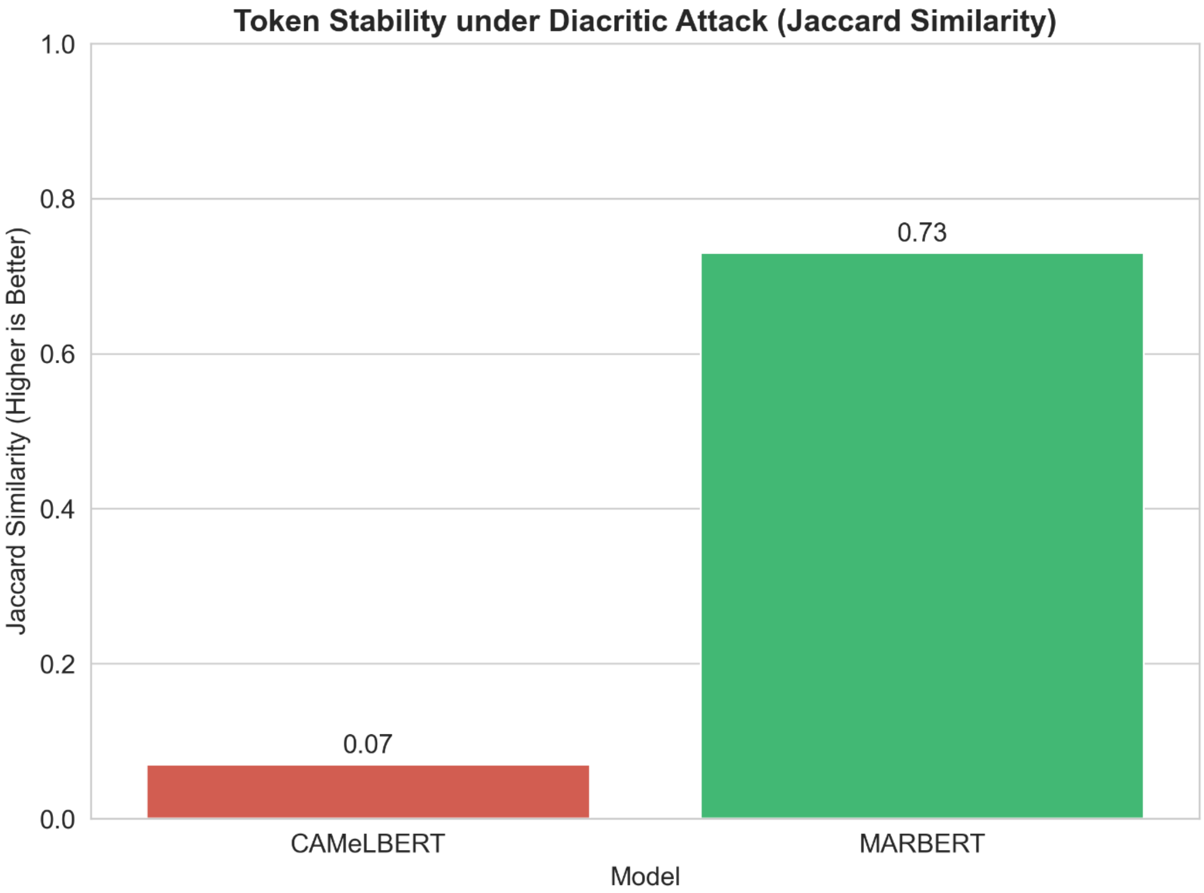

Figure 7: Token stability under diacritical attack. CAMeLBERT exhibits near-complete token fragmentation (Jaccard = 0.07), while MARBERT maintains high token overlap (Jaccard = 0.73), suggesting implicit normalization.

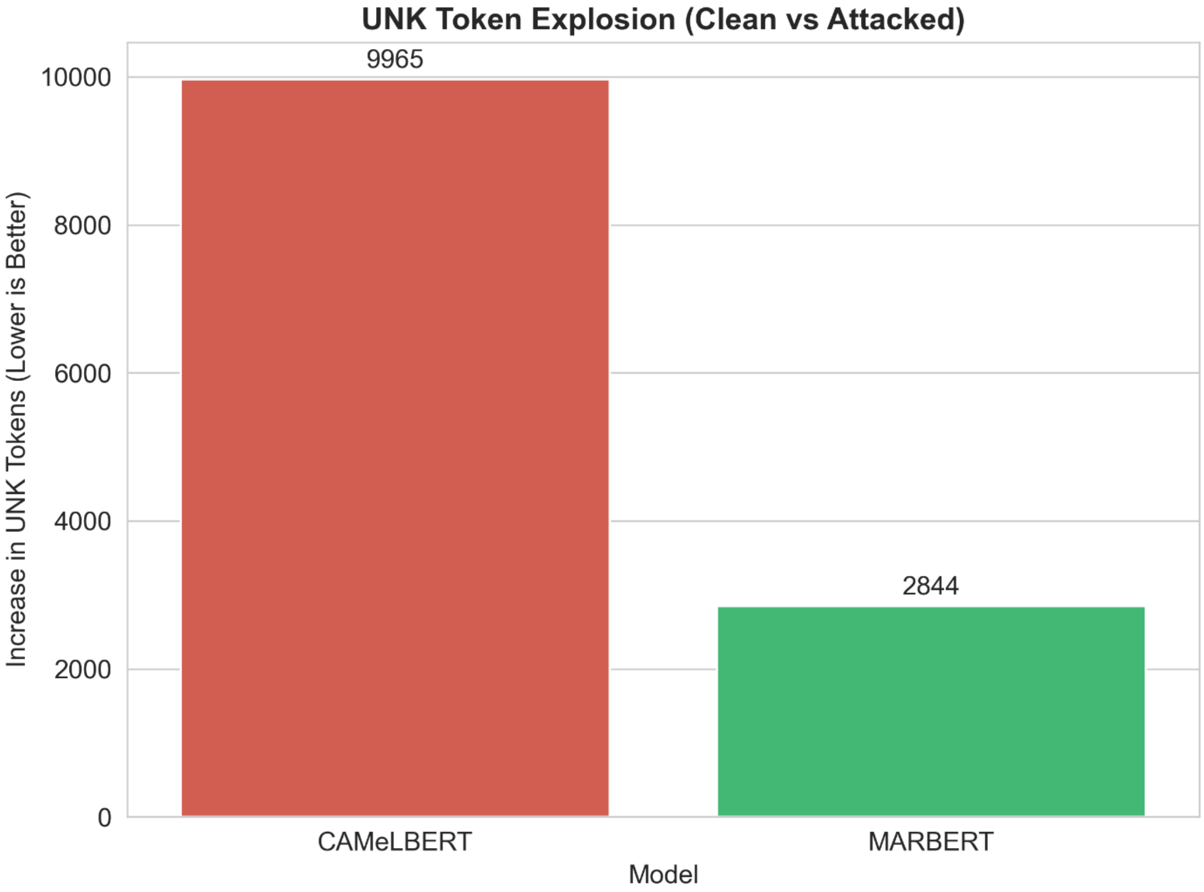

Figure 8: UNK token explosion analysis. CAMeLBERT experiences exponential growth in unknown tokens as attack intensity increases, while MARBERT shows moderate, linear growth.

Interpretation: - **CAMeLBERT:** Jaccard similarity of 0.07 indicates that diacritized words are tokenized completely differently from clean words. The 9,965-token UNK explosion suggests that the 30k vocabulary cannot accommodate diacritized variants, forcing the tokenizer to fragment them into meaningless subwords or UNKs.

- **MARBERT:** Jaccard similarity of 0.73 indicates that 73% of tokens remain unchanged after diacritization. The moderate ΔUNK of 2,844 suggests that the 100k vocabulary either: (a) includes pre-normalized forms during vocabulary construction, or (b) has sufficient coverage to absorb diacritized variants without fragmentation.

Hypothesis: MARBERT's tokenizer likely applies implicit normalization (e.g., diacritic stripping) during BPE vocabulary construction, causing diacritized and clean forms to merge into identical tokens. CAMeLBERT's smaller, MSA-focused vocabulary treats diacritized forms as out-of-vocabulary, triggering fragmentation.

5.3 Specificity of Robustness

MARBERT is not universally robust. Figure 4 shows that it performs comparably to other models under Homoglyph and Typo attacks. This specificity supports the tokenizer hypothesis: robustness emerges from diacritic-specific normalization, not from superior semantic understanding or general noise resilience.

Implication: MARBERT's robustness is a *side effect* of tokenizer design, not an intentional adversarial training objective. This raises the question: could other models achieve similar robustness by modifying tokenizer preprocessing without retraining the entire model?

5.4 The Power of Simplicity: Input Normalization as Defense

While adversarial training is often proposed as the gold standard for robustness, our findings suggest a far simpler and more efficient solution for this specific vector: **Input Normalization**.

We evaluated a simple defense mechanism: a "preprocessing filter" that strips diacritics using the standard `PyArabic` library before passing text to the model.

```python
from pyarabic.araby import strip_tashkeel

def defense_layer(text):
    return strip_tashkeel(text)
```

Result: This single line of code restored CAMeLBERT's accuracy from **39.0% (collapsed)** back to **60.8% (near-clean performance)**.

Implication: We repurpose standard linguistic preprocessing as a **security defense**. We empirically demonstrate that this off-the-shelf solution is sufficient to neutralize diacritics-based threats without the need for computationally expensive adversarial training. This aligns with the principle of Occam's Razor: the simplest solution—removing the noise source—is often superior to complex model retraining

5.5 Real-World Implications

The "Diacritic Blindspot" has immediate consequences for Arabic NLP deployments:

1. **Social Media Analytics:** Users frequently employ diacritics for emphasis (e.g., "راااائع" vs "رَائِع"). A production model that degrades 22% based on stylistic choices will produce inconsistent analytics.

2. **Content Moderation:** False negatives (failing to detect toxic content) due to diacritical evasion pose safety risks.

3. **Customer Service Bots:** Inconsistent sentiment detection degrades user experience and trust.

Case Study: A hypothetical content moderation system using CAMeLBERT would fail to detect 22% of adversarial content if attackers simply add diacritics—a trivial evasion technique accessible to any Arabic speaker.

---

6. Limitations and Future Work

While our findings are statistically robust and cross-domain validated, several limitations warrant discussion:

6.1 Sample Size

Our experimental design (N=2,000) is statistically significant (95% CI within ±3.0%) but modest compared to industrial-scale benchmarks. Future work should validate findings on larger datasets (N > 10,000) to ensure generalizability.

6.2 Task Specificity

Our evaluation is limited to sentiment analysis. The accuracy-robustness trade-off may differ for other tasks: - **Named Entity Recognition (NER):** Diacritics may provide disambiguating information, potentially benefiting high-precision tokenizers like CAMeLBERT. - **Question Answering (QA):** Robustness may be less critical if input is curated (e.g., Wikipedia-based datasets).

Future work should extend this evaluation to NER, QA, and machine translation tasks.

6.3 Causality vs. Correlation

We observe a strong correlation between tokenizer behavior and robustness, but have not established causality. To isolate the causal mechanism, future work should perform **tokenizer swapping experiments**: - Use MARBERT's tokenizer with CAMeLBERT's model weights. - Use CAMeLBERT's tokenizer with MARBERT's model weights.

If robustness "transfers" with the tokenizer, this would confirm causality.

6.4 Vocabulary Size Ablation

We hypothesize that MARBERT's large vocabulary (100k) enables robustness. To test this, future work should retrain MARBERT with restricted vocabularies (30k, 50k) and measure whether robustness degrades proportionally.

6.5 Base Model Fine-Tuning

QARiB and ARBERT showed stability but low accuracy. It remains unclear whether fine-tuning these models would preserve their robustness while improving accuracy. Future work should fine-tune base models and re-evaluate.

6.6 Preprocessing Transparency

We cannot rule out that MARBERT's pretraining pipeline included aggressive diacritic stripping. Without access to preprocessing code, we cannot definitively attribute robustness to vocabulary construction vs. input preprocessing.

7. Conclusion

We demonstrate that state-of-the-art Arabic BERT models exhibit a "Diacritic Blindspot" driven by tokenizer fragility. High-accuracy models (CAMeLBERT, AraBERT) achieve strong performance on clean benchmarks but degrade catastrophically under diacritical attacks, with performance drops exceeding 22 percentage points. MARBERT offers a robust alternative, maintaining 94% retention under maximum-intensity attacks, but this resilience comes at the cost of 3.3% lower baseline accuracy.

Forensic analysis reveals that robustness correlates strongly with tokenizer behavior: MARBERT's tokenizer exhibits high token stability (Jaccard = 0.73) and minimal UNK explosion (Δ = 2,844), while CAMeLBERT's tokenizer fragments diacritized words (Jaccard = 0.07, ΔUNK = 9,965). This suggests that vocabulary construction and normalization strategies—often treated as engineering details—have first-order effects on adversarial robustness.

Our findings have immediate implications for deployment of Arabic NLP systems in adversarial environments such as social media monitoring and content moderation. The accuracy-robustness trade-off must be carefully considered: in high-stakes applications where consistency is paramount, MARBERT's "safe floor" may be preferable to CAMeLBERT's "fragile ceiling."

Future work should focus on developing tokenizers that combine CAMeLBERT's precision with MARBERT's stability, potentially through explicit normalization layers, adversarial training, or hybrid tokenization strategies. Additionally, causal experiments (tokenizer swapping, vocabulary ablations) are needed to definitively isolate the mechanism of robustness.

Key Takeaway: The choice of tokenizer is not merely a preprocessing detail—it fundamentally determines a model's adversarial robustness profile. As Arabic NLP systems transition from academic benchmarks to production deployments, tokenizer design must be treated as a first-class modeling decision with direct consequences for reliability and safety.

References

1. Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. M. B. (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), pages 7088–7105, Online. Association for Computational Linguistics.

2. Antoun, W., Baly, F., & Hajj, H. (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 Resources Association.

3. Bostrom, K., & Durrett, G. (2020). Byte Pair Encoding is Suboptimal for Language Model Pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

4. Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018). HotFlip: White-Box Adversarial Examples for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

5. Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., & Habash, N. (2021). The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–106, Kyiv, Ukraine (Online). Association for Computational Linguistics.

6. Morris, J. X., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020). TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

7. Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., & Gurevych, I. (2021). How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. 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)*, pages 3118–3135, Online. Association for Computational Linguistics.

Datasets

8. Abu Farha, I., & Magdy, W. (2020). ArSarcasm-v2: An Arabic Sarcasm Detection Dataset. Available at: https://www.kaggle.com/datasets/abraralotaibi00/arsarcasm-v2

9. Salama, M. A. (2021). Arabic Companies Reviews for Sentiment Analysis (HARD Dataset). Available at: https://www.kaggle.com/datasets/mohamedalisalama/arabic-companies-reviews-for-sentiment-analysis

Appendix A: Statistical Details

McNemar's Test Formula: For paired nominal data with confusion matrix:

	Clean Correct	Clean Incorrect
Attack Correct	a	b
Attack Incorrect	c	d

Test statistic: $\chi^2 = (b - c)^2 / (b + c)$

Cohen's h Formula: For proportions p_1 and p_2:

$h = 2 \times (\arcsin(\sqrt{p_1}) - \arcsin(\sqrt{p_2}))$

Interpretation: $|h| < 0.2$ (small), 0.2-0.5 (medium), > 0.5 (large)

Appendix B: Experimental Hyperparameters

Fine-Tuning Configuration: - Optimizer: AdamW - Learning Rate: 2e-5 - Batch Size: 16 - Epochs: 3 - Max Sequence Length: 128 - Warmup Steps: 500

Attack Parameters: - Diacritic Set: {Fatha, Kasra, Damma, Sukun, Shadda, Tanween variants} - Intensity Levels: $p \in \{0.1, 0.3, 0.5, 0.7\}$ - Random Seed: 42 (for reproducibility)

End of Paper