GATE: General Arabic Text Embedding for Enhanced Semantic Textual Similarity with Matryoshka Representation Learning and Hybrid Loss Training

Anonymous ACL submission

Abstract

Semantic textual similarity (STS) is central to retrieval, clustering, and semantic understanding, yet Arabic remains underserved by high-quality sentence-embedding models and evaluation resources. We introduce **GATE**, a family of Arabic sentence-embedding models that combine Matryoshka Representation Learning with hybrid training over STS ranking and NLI classification objectives. We evaluate GATE across diverse Arabic tasks, including STS, NLI, pair classification, dialectal semantic relatedness, sentence classification, clustering, and retrieval, against strong multilingual baselines such as BGE-M3, Arctic-Embed-v2, and Qwen3-Embedding models. Despite its compact 135M-parameter scale, GATE achieves highly competitive results and surpasses larger multilingual models on several Arabic benchmarks. Notably, it obtains the strongest performance on native Arabic entailment and leads key dialectal and cross-lingual classification settings. A human audit further confirms that the translated training data is linguistically adequate and preserves the intended semantic relations with high reliability. With more than 250K HuggingFace downloads, GATE demonstrates both strong benchmark performance and practical adoption as an Arabic sentence-embedding family.

1 Introduction

Text embeddings are a foundational component of modern NLP systems, enabling applications such as clustering, information retrieval, and semantic similarity estimation (Reimers, 2019; Gao et al., 2023, 2021). Most high-performing embedding models rely on contrastive learning, which optimizes representations by drawing semantically related texts closer while separating unrelated ones (Gao et al., 2021; He et al., 2020; Radford et al., 2021).

In embedding models designed for retrieval and semantic-similarity tasks, the dominant recipe follows a two-stage process: large-scale weakly supervised pre-training followed by fine-tuning on curated text pairs or **triplets** (an anchor sentence, a semantically related "positive" sentence, and an unrelated "negative" sentence used to define a contrastive objective) (Li et al., 2023; Wang et al., 2022). These approaches predominantly rely on InfoNCE-style objectives with in-batch negatives (He et al., 2020), which are effective for retrieval but less suited for sentence-level semantic tasks. Recent studies show that Semantic Textual Similarity (STS) in particular benefits from objectives that explicitly model graded semantic relations rather than purely contrastive signals (Huang et al., 2024).

Reimers and Gurevych (Reimers, 2019) showed that fine-tuning on Natural Language Inference (NLI) data reshapes the embedding space by explicitly encoding entailment and contradiction, yielding substantial STS gains. Such semantic-rich supervision is especially important for Arabic. Arabic exhibits templatic morphology where words derive from root-pattern combinations (Habash, 2010; Antoun et al., 2020), giving high surface-form sparsity and weak lexical overlap between semantically related sentences. Standard written Arabic omits **diacritics** (short-vowel marks that disambiguate otherwise identical surface forms), producing orthographic ambiguity (Abdul-Mageed et al., 2020), and word order is comparatively flexible. Embedding models trained primarily on Indo-European data over-rely on surface lexical cues that are less informative in Arabic; NLI and STS supervision push embeddings toward deeper semantic relations.

To address these challenges, we introduce **GATE**, an Arabic sentence-embedding family that integrates Matryoshka Representation Learning (MRL) (Kusupati et al., 2022) with a hybrid multi-task objective. GATE jointly optimizes STS ranking and NLI classification on Arabic triplet data with hard negatives, overcoming the limitations of

single-loss, InfoNCE-only training (Gutmann and Hyvärinen, 2010), and produces scalable embeddings that retain strong performance across multiple output dimensions.

2 Related Work

Semantic Textual Similarity (Cer et al., 2017) measures graded semantic equivalence between sentence pairs and underpins applications such as machine translation, summarization, and question answering (Pathak et al., 2019; Liu et al., 2022; Wu et al., 2021). Subsequent NLI-based supervision (Reimers, 2019) reshapes embedding spaces around entailment and contradiction and remains a primary recipe for sentence-level alignment; most such work targets English or multilingual settings and uses fixed-dimensional embeddings, leaving Arabic-specific modeling underexplored.

Matryoshka Representation Learning (MRL) (Kusupati et al., 2022) learns hierarchical embeddings that remain effective when truncated, and is now adopted by recent production systems including OpenAI's third-generation embeddings (OpenAI, 2024) and EmbeddingGemma (Vera et al., 2025). Large embedding LLMs such as *E5-Mistral-7B-Instruct* (Wang et al., 2023) and *Udever-Bloom-1B1* (Zhang et al., 2023) achieve strong general-purpose performance but are computationally heavy and not Arabic-optimized.

In Arabic NLP, *AraBERT* (Antoun et al., 2020) and *MARBERT* (Abdul-Mageed et al., 2020) adapt pretrained transformers to MSA and dialectal content. Multilingual encoders such as *LaBSE* (Feng et al., 2020) and *Multilingual-E5* (Wang et al., 2024) extend coverage across languages but struggle with fine-grained Arabic semantics. In contrast, GATE integrates semantic-rich NLI supervision with Matryoshka-based multi-dimensional training, explicitly targeting fine-grained Arabic semantic similarity while maintaining efficiency.

3 GATE Framework

The GATE framework combines Matryoshka representation learning with a multi-task hybrid training approach to enhance Arabic text embeddings, trained on Arabic-adapted SNLI (Bowman et al., 2015) and MultiNLI (Kim et al., 2019) resources.

3.1 Dataset

We use Arabic-adapted subsets derived from SNLI (Bowman et al., 2015) and MultiNLI (Kim et al., 2019): these datasets provide large-scale supervision for entailment, contradiction, and semantic similarity at scales not yet matched by native Arabic NLI resources, necessitating adaptation from these established multilingual benchmarks.

We use three subsets: a Triplet split (571K train / 6.58K test) for contrastive training with NLI contradictions as hard negatives, an STS split (8.63K / 1.68K) for the similarity-ranking objective, and a Pair Classification split (981K / 19.7K) for the entailment-based classification objective. Subset sizes are reported in Appendix A.

Translation Pipeline. To adapt SNLI and MultiNLI to Arabic, we employed a neural machine translation (NMT) pipeline based on OpenNMT (Klein et al., 2017), with CTranslate2 for efficient inference and SentencePiece tokenization.

3.1.1 Translation Quality Audit

Because our training data is machine-translated, we conducted a structured quality audit following recent precedent in multilingual NLI dataset construction (Heredia et al., 2024; Tas et al., 2025; Artetxe et al., 2024). We drew 250 triplets uniformly at random from the public *Arabic-NLi-Triplet* dataset and expanded each into two directed pairs — (anchor, positive) with expected entailment, and (anchor, negative) with expected contradiction — yielding 500 distinct pair items. Each was independently annotated by three bilingual English–Arabic annotators on two dimensions: **Arabic adequacy** (1–4 scale) and binary **relation preservation**. Full annotation guidelines are provided in Appendix D.

Findings. The audit confirms that the translated data is of high Arabic surface quality (adequacy 3.70–3.94 of 4) and that the training signal is reliable: under majority agreement, **84.2%** of all pairs preserve the expected NLI relation, rising to **90.4%** for anchor–positive entailments. Anchor–negative contradictions are weaker (78.0% majority preservation), reflecting a known property of NLI-derived training data where contradiction examples in SNLI/MultiNLI are not always strict logical contradictions (Gururangan et al., 2018). The downstream effect of this calibrated signal is measured directly in Section 4.3, where models trained on this data lead native-Arabic NLI evaluation.

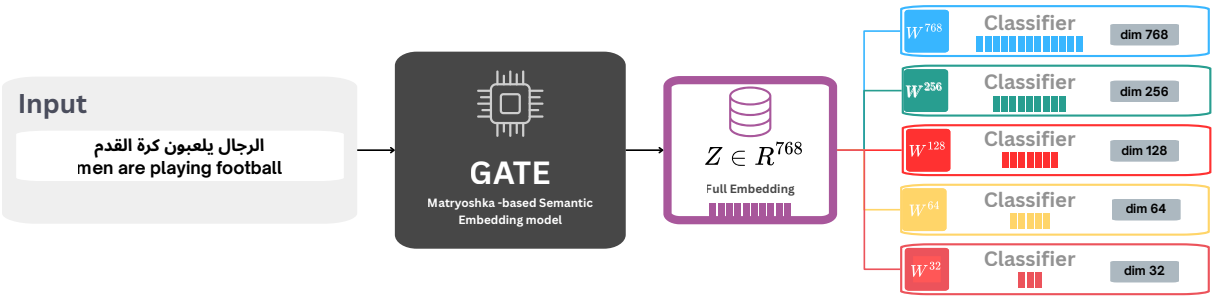

Figure 1: Matryoshka embedding architecture with variable-dimension representations and weight tying.

Quantity	Value	Agr.
Adequacy sent. 1 (mean, 1–4)	3.84 / 3.70 / 3.91	0.41
Adequacy sent. 2 (mean, 1–4)	3.89 / 3.81 / 3.94	0.36
Relation preserved, unanimous	61.80%	0.23
anchor–positive	**78.00%**	0.27
anchor–negative	45.60%	0.16
Relation preserved, majority	**84.20%**	–
anchor–positive	**90.40%**	–
anchor–negative	78.00%	–

Table 1: Translation quality audit on 500 pairs × 3 annotators (1,500 decisions). Adequacy means are reported per annotator (Ann. 1 / Ann. 2 / Ann. 3); agreement is mean pairwise quadratic-weighted κ for adequacy and Fleiss' κ for binary relation preservation. Unanimous: all three annotators agree; majority: at least two of three.

3.2 Proposed Arabic Matryoshka Models

We propose a family of Matryoshka-based sentence embedding models designed for Arabic semantic similarity and natural language inference. These models share a common training framework but differ in their underlying pretrained backbones, enabling a controlled comparison across Arabic-specific and multilingual representations.

At the core of the framework is *GATE-AraBERT-v1*, a multi-task embedding model derived from *Arabic-Triplet-Matryoshka-V2* and fine-tuned using hybrid STS and NLI supervision. Additional variants include *MARBERT-all-nli-triplet-Matryoshka*, which targets both Modern Standard Arabic and dialectal content, and *Arabic-LaBSE-Matryoshka*, which adapts cross-lingual representations for Arabic. We also include *Arabic-all-nli-triplet-Matryoshka* and *E5-all-nli-triplet-Matryoshka* as multilingual baselines trained under the same triplet-based Matryoshka framework.

This unified model family allows us to isolate the impact of semantic-rich supervision and Matryoshka training across different pretrained encoders while maintaining a consistent optimization setup.

3.2.1 Matryoshka-Based Triplet Training

Matryoshka Representation Learning (MRL) (Kusupati et al., 2022) produces sentence embeddings that remain semantically meaningful when truncated to a prefix of lower dimension as shown in Figure 1. Given an encoder F producing $\mathbf{z} \in \mathbb{R}^d$, MRL requires that every prefix sub-vector $\mathbf{z}_{1:m}$ for $m \in M = \{768, 512, 256, 128, 64\}$ remain a usable representation. Following the Efficient MRL formulation, we use a single classifier matrix with prefix-truncated views $\mathbf{W}^{(m)} = \mathbf{W}_{1:m}$, so a single forward pass yields all dimensional views without re-encoding.

We train on the public *Arabic-NLi-Triplet* dataset (557,850 triplets) using the MultipleNegativesRankingLoss, equivalent to a softmax-based contrastive loss with in-batch and labeled negatives:

$$\mathcal{L}_{\text{triplet}} = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\exp\big(s(x_i, x_i^+)/\tau\big)}{\sum_{j=1}^{n} \exp\big(s(x_i, x_j^+)/\tau\big)}, \tag{1}$$

where $s(\cdot,\cdot)$ is cosine similarity and τ is a temperature. Contradiction pairs from NLI serve as hard negatives. The triplet loss is wrapped by MatryoshkaLoss, which applies the same contrastive objective at every $m \in M$ over the shared encoder. Training uses an NVIDIA A100, batch size 128, and maximum sequence length 512.

3.2.2 Hybrid Loss Training Approach

GATE-AraBERT-v1 augments the contrastive Matryoshka objective with explicit NLI classification supervision. Mini-batches are sampled alternately from STS or NLI data, and the corresponding loss is applied. The STS ranking loss L_{sts} enforces correct relative ordering of pairs via a log-sum-exp formulation that implicitly enforces a margin without manual tuning. The NLI classification loss L_{cls} is a softmax over embedding cosine similarities

Model	Dim	# Params.	STS17	STS22	STS22-v2	Avg.
Arabic-Triplet-Matryoshka-V2	768	135M	**85.31**	**60.70**	**63.96**	**69.99**
GATE-AraBERT-v1	768	135M	82.78	59.75	63.09	**68.54**
AraBERTv02	768	135M	54.53	46.86	49.95	50.45
MARBERT-all-nli-triplet-Matryoshka	768	163M	**82.18**	**58.08**	**61.32**	**67.19**
MARBERTv2	768	163M	60.98	49.92	53.75	54.88
Arabic-LaBSE-Matryoshka	768	471M	**82.46**	57.25	60.58	**66.76**
LaBSE	768	471M	69.07	**57.66**	**60.98**	62.57
E5-all-nli-triplet-Matryoshka	384	278M	**80.37**	56.34	59.64	**65.45**
multilingual-e5-small	384	278M	74.62	**58.13**	**61.40**	64.72
Arabic-all-nli-triplet-Matryoshka	768	135M	**82.40**	51.38	54.45	**62.74**
paraphrase-multilingual-mpnet-base-v2	768	135M	79.09	**52.18**	**55.37**	62.21

Table 2: Performance comparison of Matryoshka-based models and their base counterparts on Arabic STS tasks from MTEB.

with label-based (rather than in-batch) negatives, encouraging explicit semantic separation between entailment, neutral, and contradiction:

$$L_{\text{cls}} = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\exp(s(x_i, y_i^+)/\tau)}{\sum_{y\in\{y_i^+\}\cup\{y_{i,j}^-\}}\exp(s(x_i, y)/\tau)}. \tag{2}$$

At each step, the active objective is $L = L_{\text{cls}}$ on NLI batches and $L = L_{\text{sts}}$ on STS batches. This alternating scheme avoids manual loss weighting and preserves task specialization.

4 Results and Discussion

4.1 Performance on Arabic STS Benchmarks

We evaluate GATE models and their base counterparts on Arabic Semantic Textual Similarity tasks from the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2022), including STS17 and STS22-v2. Table 2 reports the results.

Matryoshka-based models consistently outperform their corresponding base models across all benchmarks. *Arabic-Triplet-Matryoshka-V2* achieves the best overall performance (69.99 average), with particularly strong results on STS17 (85.31), while *GATE-AraBERT-V1* follows closely with an average score of 68.54.

The GATE-v1 vs. ATM-V2 trade-off. The 1.45-point STS gap between *GATE-AraBERT-v1* (68.54) and *Arabic-Triplet-Matryoshka-V2* (69.99) is a deliberate generalization–specialization trade-off introduced by hybrid loss training. ATM-V2 is optimized purely with the contrastive Matryoshka objective and therefore peaks on STS-style ranking; GATE-v1 augments this with explicit NLI classification supervision, sacrificing fractional STS performance in exchange for two measurable downstream gains documented later in the paper: (i) it wins the XNLI Arabic pair-classification task (87.46 vs. 85.42, Table 4); (ii) it achieves the highest average across the five Arabic sentence-classification tasks (Table 5); and (iii) error analysis shows it produces substantially more conservative similarity scores in the low-similarity regime, reducing the lexical-overlap false-positive bias common to contrastive-only models. Practitioners requiring peak STS performance should prefer ATM-V2; practitioners requiring robust behavior across NLI/classification/STS jointly should prefer GATE-v1.

Other Matryoshka variants, including *MARBERT-all-nli-triplet-Matryoshka* and *Arabic-LaBSE-Matryoshka*, also achieve strong results, substantially exceeding their base counterparts. In contrast, non-Matryoshka base models perform significantly worse, highlighting the importance of semantic-rich supervision and multi-dimensional training for Arabic STS.

When does MRL help? An asymmetry across backbones. The effect of MRL+NLI supervision is strongly conditioned by the pretraining regime of the underlying encoder. Measuring the average gain Δ_{avg} of each MRL-enhanced model over its base across STS17, STS22, and STS22-v2, Arabic-specific backbones gain substantially ($\Delta_{\text{AraBERTv02}} = +19.5$, $\Delta_{\text{MARBERTv2}} = +12.3$, positive on all three benchmarks), while multilingual backbones gain markedly less and inconsistently ($\Delta_{\text{LaBSE}} = +4.2$, $\Delta_{\text{mE5-small}} = +0.7$, $\Delta_{\text{mpnet-v2}} = +0.5$; all three improve on STS17 but regress on STS22 / STS22-v2). The mechanistic reading: AraBERTv02 and MARBERTv2 are pretrained on

Arabic text under MLM only, so MRL+NLI introduces sentence-level semantic geometry that was largely absent. Multilingual encoders already incorporate cross-lingual or paraphrase-style supervision during pretraining (LaBSE uses translation ranking (Feng et al., 2020); multilingual-E5 uses contrastive multilingual pairs (Wang et al., 2024); mpnet-v2 distills English paraphrase supervision), leaving less marginal value for a second round of single-language contrastive training.

Impact of loss selection. Table 3 summarizes the effect of loss selection on Arabic STS performance. The baseline cross-entropy loss (L_{CE}) performs poorly (50.45 average), confirming its limitations for fine-grained semantic similarity. Matryoshka loss (L_{MRL}) yields the strongest STS performance: *Arabic-Triplet-Matryoshka-V2* achieves a 69.99 average and 85.31 on STS17. The hybrid loss ($L_{sts} + L_{cls}$), used by GATE-AraBERT-v1, achieves slightly lower but still competitive STS performance (68.54 average) in exchange for the downstream generalization documented in §4.3, including the XNLI-ar and dialectal classification wins.

Loss	STS17	STS22	STS22-v2	Avg.
L_{CE}	54.53	46.86	49.95	50.45
L_{MRL}	**85.31**	**60.70**	**63.96**	**69.99**
$L_{sts} + L_{cls}$	82.78	59.75	63.09	68.54

Table 3: Effect of loss function on Arabic STS. L_{CE}: AraBERTv02 cross-entropy fine-tune. L_{MRL}: pure-Matryoshka contrastive (ATM-V2). L_{sts}+L_{cls}: hybrid (GATE-AraBERT-v1).

Dimensional robustness. Matryoshka Representation Learning enables strong performance retention under dimensional reduction: for *Arabic-Triplet-Matryoshka-V2*, average STS performance drops only 0.56 points from full 768-dim (69.99) to 64-dim (69.43), confirming that MRL allows aggressive compression without substantial accuracy loss (per-dimension breakdown in Appendix B).

4.2 Comparison with Large Embedding Models

We compare GATE against large-scale embedding models including *e5-mistral-7b-instruct* (Wang et al., 2023), *udever-bloom-1b1* (Zhang et al., 2023), and the OpenAI text-embedding APIs (OpenAI, 2023, 2024) (Figure 2). At 135M parameters, both ATM-V2 and GATE-AraBERT-v1 match or exceed billion-parameter models on Arabic STS, demonstrating that carefully optimized, Arabic-specialized embeddings can rival much larger general-purpose LLM embeddings.

4.3 Beyond STS: Broader Arabic Evaluation

A central criticism of prior Arabic embedding work is that evaluation is concentrated on STS, leaving downstream task generalization unverified. We therefore evaluate GATE checkpoints alongside six recent contemporary embedding baselines — BGE-M3 (Chen et al., 2024), Arctic-Embed-L-v2.0 (Merrick et al., 2024), and Qwen3-Embedding 0.6B and 4B (Zhang et al., 2025) — on seven Arabic task categories: pair-classification (including the *native-Arabic* ArEntail benchmark), dialect-aware semantic relatedness, sentence-level classification, clustering, retrieval, and reranking. All evaluations use Arabic-only or Arabic-touching subsets (e.g. STS17 `ar-ar`, XNLI `ar`, SemRel-2024 `arb_Arab`).

Native Arabic NLI is the strongest validation. The most important downstream evaluation for our argument is **ArEntail** (Obeidat et al., 2024), a natively-annotated Arabic NLI benchmark that does not depend on translation. We report average precision (AP) on the binary pair-classification setup using cosine similarity, with 95% CIs from 1,000 bootstrap resamples of the test set. Table 4 shows that **Arabic-Triplet-Matryoshka-V2 (135M) achieves the highest point estimate** (91.55, 95% CI [89.56, 93.34]); its lower bound exceeds the point estimate of BGE-M3 (89.08, 568M), and both GATE checkpoints clearly outperform Qwen3-Embedding-4B (86.88) — a model with 30× more parameters. The top three CIs overlap, so we do not claim strict 95% significance over BGE-M3; rather, the result establishes that Arabic-specialized GATE encoders are competitive with, and typically above, much larger multilingual encoders on a native-Arabic semantic-understanding task.

Competitive monolingual retrieval without retrieval-specific training. Although GATE is trained only on contrastive NLI triplets and STS pairs, it transfers to Arabic retrieval out-of-the-box. Arabic-Triplet-Matryoshka-V2 attains 60.9 nDCG@10 on MIRACL-ar and 67.0 MAP on MIRACL reranking, and GATE-AraBERT-v1 reaches 68.8 on Namaa-MrTyDi reranking — the best score among GATE-derived models on that task.

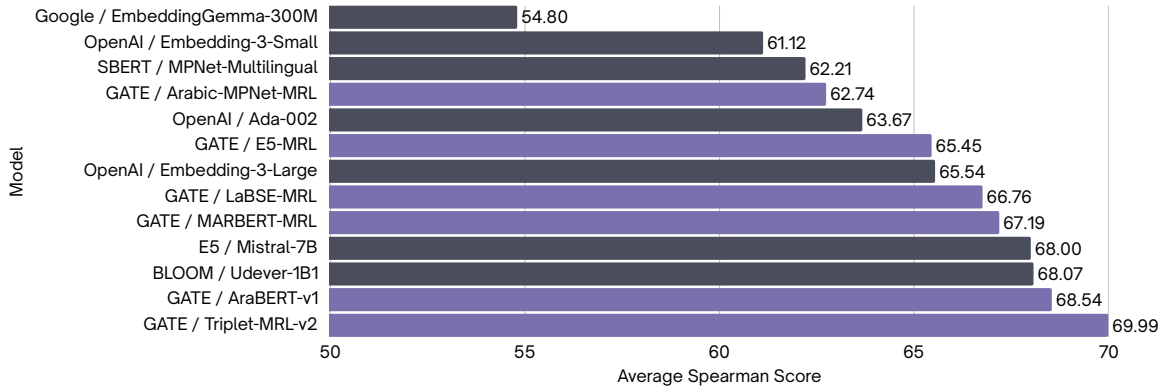

Figure 2: Performance comparison between Matryoshka-based GATE models and larger embedding models on MTEB Arabic benchmarks.

Model	Params	AP	95% CI
Arabic-Triplet-Mat.-V2	135M	**91.55**	[89.56, 93.34]
BGE-M3	568M	89.08	[86.70, 91.25]
GATE-AraBERT-v1	135M	88.44	[85.89, 90.71]
Qwen3-Embedding-4B	4B	86.88	[84.19, 89.44]
Arctic-Embed-L-v2.0	568M	78.36	[74.57, 82.04]
Qwen3-Embedding-0.6B	595M	77.10	[73.32, 81.21]
XNLI-ar (translated; subset of models)			
Arabic-Triplet-Mat.-V2	135M	85.42	—
GATE-AraBERT-v1	135M	**87.46**	—
BGE-M3	568M	77.86	—

Table 4: Native-Arabic ArEntail Pair-Classification Results.

Model	AJGT	Hotel	Restaur.	Tweet	Store
Arabic-Triplet-Matryoshka-V2	87.43	49.73	75.36	**49.69**	37.02
GATE-AraBERT-v1	**88.27**	54.04	77.55	45.43	**38.78**
BGE-M3	82.84	**56.61**	**79.00**	45.31	37.65
Arctic-Embed-L-v2.0	82.04	52.36	67.32	47.30	32.89
Qwen3-Embedding-4B	76.38	52.87	69.22	45.86	33.06
Qwen3-Embedding-0.6B	73.91	49.50	64.28	45.59	32.69

Table 5: Sentence-level classification accuracy on five Arabic benchmarks.

Retrieval-tuned baselines remain stronger on absolute retrieval metrics, as expected, so GATE and retrieval-tuned models such as BGE-M3 are complementary: GATE supplies a compact Arabic-specialized semantic backbone for monolingual applications, and can be combined with retrieval-tuned rerankers when deployment demands optimized retrieval (full numbers in Appendix C).

Sentence-level classification: GATE wins. On Arabic sentence-level classification tasks as shown in Table 5, GATE-AraBERT-v1 achieves the highest average score among models with full coverage, outperforming BGE-M3 on 4 of 5 tasks. The gain is largest on dialectal data (AJGT Jordanian, +5.4 over BGE-M3), suggesting that the Arabic-specific backbone retains an advantage on dialect-sensitive content.

Dialect robustness on SemRel-2024. The SemRel-2024 Arabic benchmark (Ousidhoum et al., 2024) provides natively annotated semantic-relatedness scores across three Arabic varieties: Modern Standard Arabic (MSA), Algerian (`arq`), and Moroccan (`ary`). Table 6 shows that the dialectal picture is markedly different from MSA-only evaluation. Qwen3-Embedding-4B leads on MSA (64.60) but *collapses on Algerian* (24.00) and Moroccan (45.76). The four Arabic-specialized or Arabic-strong models (ATM-V2, GATE-v1, BGE-M3, Arctic-Embed-v2) all cluster tightly on Algerian (47–49) and are the only models robust across dialectal Arabic. GATE-AraBERT-v1 achieves the highest Algerian score among all evaluated models (48.84), suggesting that the MARBERT-style dialectal exposure inherited by Arabic-specific backbones persists through hybrid-supervised fine-tuning.

4.4 Error Analysis

We probe per-pair behavior across three representative similarity regimes (low / moderate / high; full sentence pairs and per-model cosine predictions in Appendix E). Three patterns emerge. **(i) GATE checkpoints uniquely suppress the lexical-overlap false-positive bias.** On the low-similarity

Model	MSA	Alg.	Mor.
Arabic-Triplet-Matryoshka-V2	61.38	47.68	41.94
GATE-AraBERT-v1	47.86	**48.84**	24.80
BGE-M3	59.10	47.25	30.28
Arctic-Embed-L-v2.0	55.10	47.60	45.30
Qwen3-Embedding-0.6B	58.95	26.63	42.27
Qwen3-Embedding-4B	**64.60**	24.00	**45.76**

Table 6: Per-dialect Spearman correlation on SemRel-2024 Arabic (arb / arq / ary).

pair (gold 0.10), the four contemporary multilingual baselines all predict cosine 0.50–0.66, driven by a shared lexical token (

رجل, "a man"). ATM-V2 (0.033) and GATE-AraBERT-v1 (0.040) are an order of magnitude lower, matching the gold annotation almost exactly. Semantic-rich Arabic supervision, not scale, is what suppresses this bias. **(ii) GATE models are more conservative on paraphrastic pairs.** In the moderate and high regimes, GATE and ATM-V2 predict 0.53–0.74 while baselines push up to 0.91, the expected trade-off of the same conservativeness that wins (i). **(iii) Hybrid loss restores some paraphrase sensitivity.** GATE-AraBERT-v1 (hybrid) is higher than ATM-V2 (pure contrastive) on the moderate pair (0.645 vs. 0.533) while remaining near-equal on the low-similarity pair, recovering paraphrase sensitivity without losing false-positive suppression.

4.5 Public Release and Deployment Evidence

The GATE family is publicly released on HuggingFace under an open licence. As of submission, *GATE-AraBERT-v1* and *Arabic-Triplet-Matryoshka-V2* reaches 250K cumulative downloads across a 12-month deployment window), providing real-world evidence that the released checkpoints remain useful relative to newer multilingual encoders.

5 Conclusion

We introduced **GATE**, a family of compact Arabic sentence embeddings that combine Matryoshka Representation Learning with a hybrid STS–NLI training objective. Through a systematic evaluation across seven Arabic task categories, we show that semantic-rich supervision on an Arabic-specific backbone yields embeddings that are competitive with — and on native-Arabic semantic understanding, favorable over — contemporary multilingual models one to two orders of magnitude larger. A three-annotator audit confirms that the machine-translated training signal is reliable, and the released checkpoints have seen substantial real-world adoption. Together, the empirical findings, the open-released artifacts, and the dialect-aware evaluation position GATE as a practical and well-validated foundation for fine-grained Arabic semantic understanding.

6 Limitations

GATE is evaluated on Arabic, and the hybrid loss is applied to GATE-AraBERT-v1 as a controlled ablation against the pure-Matryoshka ATM-V2. The training data is machine-translated NLI; the audit in Section 3.1.1 confirms 84.2% relation preservation under majority agreement (90.4% for entailment positives), with the weaker negative-pair signal reflecting a known property of NLI-derived contradictions (Gururangan et al., 2018). Natural directions for future work include incorporating natively-authored Arabic hard negatives, extending the hybrid loss across additional backbones, and applying the same methodology to other morphologically rich languages.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: Deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2024. Translation errors significantly impact low-resource languages in cross-lingual learning. In *Proceedings of EACL 2024 (Short Papers)*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity

text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT 2018*.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Maite Heredia, Julen Etxaniz, Muitze Zulaika, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2024. XNLIeu: a dataset for cross-lingual NLI in Basque. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Junqin Huang, Zhongjie Hu, Zihao Jing, Mengya Gao, and Yichao Wu. 2024. Piccolo2: General text embedding with multi-task hybrid loss training. *arXiv preprint arXiv:2405.06932*.

Seonhoon Kim, Inho Kang, and Nojun Kwak. 2019. Semantic sentence matching with densely-connected recurrent and co-attentive information. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6586–6593.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. 2022. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2022. Key phrase aware transformer for abstractive summarization. *Information Processing & Management*, 59(3):102913.

Luke Merrick, Puxuan Yu, Daniel Campos Nguyen, Iz Beltagy Le Scao, Aditya Phadke, et al. 2024. Arctic-Embed 2.0: Multilingual retrieval without compromise. *arXiv preprint arXiv:2412.04506*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Rasha Obeidat, Mais Al-Harbi, Mahmoud Al-Ayyoub, and Yousef Al-Hashedi. 2024. ArEntail: Manually-curated arabic natural language inference dataset from news headlines. In *Language Resources and Evaluation*.

OpenAI. 2023. Openai embeddings documentation. https://platform.openai.com/docs/guides/embeddings.

OpenAI. 2024. New embedding models and api updates. Accessed: 2024-08-28.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Meriem Beloucif, Christine De Kock, et al. 2024. SemRel2024: A collection of semantic textual relatedness datasets for 13 languages. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Amarnath Pathak, Partha Pakray, and Jereemi Bentham. 2019. English–mizo machine translation using neural and statistical approaches. *Neural Computing and Applications*, 31(11):7615–7631.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Mert Tas, Levent Albayrak, et al. 2025. TurkMedNLI: a Turkish medical natural language inference dataset through translation and human annotation. *PeerJ Computer Science*.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. 2025. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

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

Yongliang Wu, Shuliang Zhao, and Ruiqiang Guo. 2021. A novel community answer matching approach based on phrase fusion heterogeneous information network. *Information Processing & Management*, 58(1):102408.

Xin Zhang, Zehan Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. 2023. Language models are universal embedders. *arXiv preprint arXiv:2310.08232*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Dataset Subset Sizes

Subset	Columns	Train	Test
STS	text, text pair, score	8.63K	1.68K
Triplet	text, text triplet	571K	6.58K
Pair Classification	text, text pair, label	981K	19.7K

Table 7: Number of sentence pairs (STS, Pair Classification) or triplets (Triplet) per subset, in training and evaluation splits.

B Matryoshka Dimensional Sweep

Dim.	STS17	STS22	STS22-v2	Avg.
768	85.31	**60.70**	63.96	**69.99**
512	85.17	60.62	**63.98**	69.92
256	**85.39**	60.41	63.77	69.86
128	84.67	60.27	63.62	69.52
64	84.04	60.44	63.80	69.43

Table 8: Matryoshka dimensional sweep for *Arabic-Triplet-Matryoshka-V2*. Average STS performance drops only 0.56 points from 768 to 64 dimensions.

C Retrieval and Reranking Details

Model	nDCG@10	MAP (Rerank)
Arabic-Triplet-Matryoshka-V2	60.91	67.02
GATE-AraBERT-v1	56.11	64.60
BGE-M3	**78.33**	**79.04**
Arctic-Embed-L-v2.0	75.66	76.85

Table 9: MIRACL-ar retrieval (nDCG@10) and reranking (MAP).

D Translation-Audit Annotation Guidelines

This appendix reproduces the guidelines given to the three bilingual English–Arabic annotators for the translation-quality audit described in Section 3.1.1.

Task

You are auditing 500 Arabic sentence pairs derived from the *Arabic-NLi-Triplet* training data. Each pair is one directed slice of an NLI triplet:

- *anchor–positive*: the anchor and the positive sentence. Expected relation: **entailment**.
- *anchor–negative*: the anchor and the negative sentence. Expected relation: **contradiction**.

Your two judgments per pair are (1) the Arabic quality of each sentence and (2) whether the expected relation still holds in the translated Arabic pair.

Adequacy (1–4 scale)

Rate each Arabic sentence independently:

- **4 – Fully adequate**: fluent, grammatical, and semantically coherent Arabic prose. Minor stylistic awkwardness is acceptable.
- **3 – Mostly adequate**: core meaning is clear, with minor distortion or stylistic awkwardness.
- **2 – Partially adequate**: significant meaning ambiguity or grammatical errors; meaning partially recoverable.
- **1 – Inadequate**: grammatically broken, nonsensical, or a clearly broken mistranslation.

Score on meaning preservation, not on dialect or word-order naturalness.

Relation preservation (yes/no)

Given only the two Arabic sentences, decide whether the expected relation (entailment for anchor–positive, contradiction for anchor–negative) still holds for a competent Arabic reader. If *no*, indicate the corrected relation (entailment, neutral, or contradiction) in the `annotator_relation` column. An empty `annotator_relation` cell is treated as agreement with the expected relation. Annotators work independently; do not consult one another during annotation.

Aggregation

We report:

- Per-annotator mean adequacy (sentence 1, sentence 2).
- Mean pairwise quadratic-weighted Cohen's κ on adequacy.
- Unanimous and majority relation preservation rates.
- Fleiss' κ on the binary relation-preservation judgment.

E Error Analysis: Per-Pair Cosine Predictions

Tables 10, 11, and 12 report the full per-model cosine similarity on the three representative sentence pairs analyzed in the main paper (low, moderate, and high gold similarity regimes).

رجل يقود سيارة / رجل يعزف على الجيتار
"a man playing the guitar" / "a man driving a car"
Gold: 0.10

Model	Cosine
Arabic-Triplet-Mat.-V2	**0.033**
GATE-AraBERT-v1	**0.040**
BGE-M3	0.657
Arctic-Embed-L-v2.0	0.529
Qwen3-Embedding-0.6B	0.496
Qwen3-Embedding-4B	0.619

Table 10: Low-similarity regime (lexically overlapping, semantically unrelated).

الأولاد يلعبون كرة القدم / الرجال يلعبون كرة القدم
"men playing football" / "boys playing football"
Gold: 0.72

Model	Cosine
Arabic-Triplet-Mat.-V2	0.533
GATE-AraBERT-v1	0.645
BGE-M3	**0.866**
Arctic-Embed-L-v2.0	0.825
Qwen3-Embedding-0.6B	0.696
Qwen3-Embedding-4B	0.778

Table 11: Moderate-similarity regime.

رجل يقوم بخدعة ورق / رجل يقوم بخدعة بالبطاقات
"a man doing a card trick" / "a man performing a card trick"
Gold: 1.00

Model	Cosine
Arabic-Triplet-Mat.-V2	0.714
GATE-AraBERT-v1	0.735
BGE-M3	0.849
Arctic-Embed-L-v2.0	0.785
Qwen3-Embedding-0.6B	0.796
Qwen3-Embedding-4B	**0.911**

Table 12: High-similarity regime (near-paraphrase).