

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

Omer Nacar ^{1*} Anis Koubaa ² Serry Sibae ³ Yasser Al-Habashi ³

Adel Ammar ³ Wadii Boulila ³

¹Tuwaiq Academy ²Alfaisal University ³Prince Sultan University

*o.najar@tuwaiq.edu.sa

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 compact family of Arabic sentence embedding models and a controlled study of Arabic-specific backbones, contrastive Matryoshka Representation Learning, dimensional truncation, and hybrid STS-NLI supervision. We evaluate the resulting models across Arabic semantic similarity, NLI, pair classification, sentence classification, clustering, retrieval, reranking, and dialectal semantic relatedness, against strong multilingual baselines including BGE-M3, Arctic-Embed-v2, and Qwen3-Embedding. Despite their 135M-parameter scale, GATE-family models are competitive with substantially larger multilingual encoders and achieve strong results on several native-Arabic semantic tasks. A three-annotator audit further finds high Arabic surface adequacy and 84.2% majority relation preservation for the translated supervision. The released checkpoints, data resources, and experimental artifacts support reproducible Arabic embedding research.

1 Introduction

Text embeddings are a foundational component of modern NLP systems, supporting applications such as semantic search, clustering, retrieval, and similarity estimation (Reimers and Gurevych, 2019; Gao et al., 2023, 2021). Most high-performing embedding models use contrastive learning to bring semantically related texts closer in the embedding space while separating unrelated texts (Gao et al., 2021; He et al., 2020; Radford et al., 2021).

A common training recipe combines large-scale weak supervision with fine-tuning on curated sentence pairs or triplets, where each triplet contains an anchor, a semantically related positive, and an unrelated negative example (Li et al., 2023; Wang

et al., 2022). Such methods often use InfoNCE-style objectives and in-batch negatives (He et al., 2020), which are highly effective for retrieval. However, sentence-level semantic understanding also requires modeling graded similarity and explicit inferential relations. In particular, Semantic Textual Similarity (STS) benefits from supervision beyond a purely binary contrastive signal (Huang et al., 2024).

This challenge is especially relevant for Arabic. Its rich morphology, orthographic ambiguity caused in part by omitted diacritics, flexible word order, and substantial dialectal variation make surface lexical overlap an imperfect indicator of meaning (Habash, 2010; Antoun et al., 2020; Abdul-Mageed et al., 2021). Although multilingual embedding models offer broad coverage, their training data and objectives may not adequately capture the fine-grained semantic distinctions needed for native Arabic and dialect-sensitive applications. NLI supervision is particularly useful in this setting because it explicitly distinguishes entailment, neutrality, and contradiction rather than relying only on lexical similarity (Reimers and Gurevych, 2019).

In this work, we present **GATE**, a compact family of Arabic sentence-embedding models. Rather than proposing a new encoder architecture or loss function, GATE provides a controlled empirical study of practical design choices for Arabic embeddings: Arabic-specific versus multilingual backbones, contrastive Matryoshka Representation Learning (MRL) (Kusupati et al., 2022), embedding-dimensional truncation, and joint STS-NLI supervision. The family includes pure Matryoshka contrastive models, such as ATM-V2, and a hybrid STS-NLI variant, GATE-v1, allowing us to examine the trade-off between STS specialization and broader semantic-task generalization.

We evaluate these choices across seven Arabic task categories, covering semantic similarity, pair classification, sentence classification, retrieval,

reranking, and dialect-aware semantic relatedness. This scope does not claim complete coverage of every Arabic embedding benchmark; rather, it assesses generalization across a diverse set of Arabic semantic tasks.

Our contributions are:

- We release **GATE**, a compact family of Arabic sentence-embedding checkpoints built around complementary contrastive MRL and hybrid STS–NLI training configurations.
- We provide controlled analyses of the effects of backbone choice, Matryoshka dimensional truncation, and hybrid STS–NLI supervision on Arabic embedding quality.
- We evaluate the resulting models across seven Arabic task categories, including native-Arabic NLI and dialect-sensitive semantic relatedness, and compare them with contemporary multilingual embedding models of substantially larger scale.

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 and Gurevych, 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., 2024a) and *Udever-Bloom-1B1* (Zhang et al., 2025a) 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., 2021) adapt

pretrained transformers to MSA and dialectal content. Multilingual encoders such as *LaBSE* (Feng et al., 2022) and *Multilingual-E5* (Wang et al., 2024b) 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.

Recent Arabic-centric embedding work includes *Swan* and *ArabicMTEB* (Bhatia et al., 2025). *Swan* introduces compact and large Arabic-focused embedding models, while *ArabicMTEB* provides a broad evaluation suite spanning eight task categories and 94 datasets. This work is complementary: GATE focuses on controlled analysis of Arabic backbone choice, Matryoshka training, dimensional truncation, and hybrid STS–NLI supervision.

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 (Williams et al., 2018) resources.

3.1 Dataset

We use Arabic-adapted subsets derived from SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018). 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 Table 1.

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.

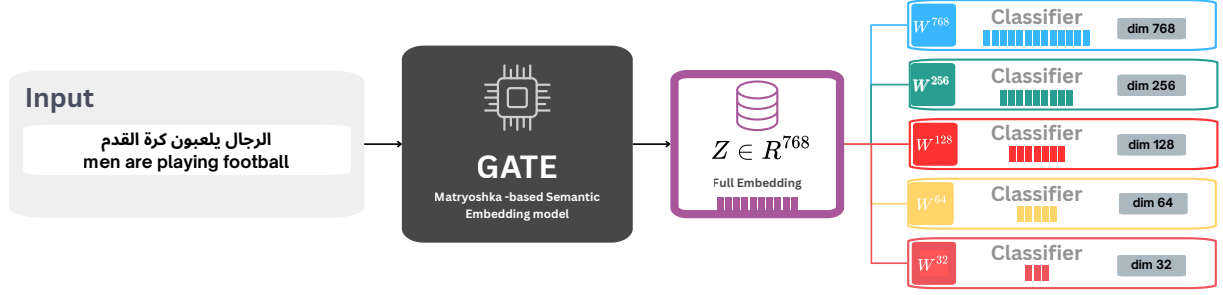


Figure 1: Matryoshka embedding architecture with variable-dimension representations and weight tying.

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

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 2: Translation quality audit on 500 pairs $\times$ 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.

3.1.1 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 C.

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.

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 Arabic semantic-supervision framework but differ in backbone, training objective, and intended use. Table 3 summarizes the family and defines the short names used throughout the paper.

We use **ATM-V2** for *Arabic-Triplet-Matryoshka-V2* and **GATE-v1** for *GATE-AraBERT-v1* throughout the paper. The remaining variants are named by their backbone, followed by **-MRL**. This structure isolates the effect of backbone choice and hybrid STS-NLI supervision under an otherwise consistent Arabic training recipe.

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

Short name	Backbone	Params.	Training objective	Role in this study
ATM-V2	AraBERTv02	135M	Pure Matryoshka contrastive training on Arabic NLI triplets.	Primary STS-oriented model and direct comparison for GATE-v1.
GATE-v1	ATM-V2 / AraBERTv02	135M	Hybrid Matryoshka training with STS ranking and NLI classification.	Main GATE model; trades a small amount of STS performance for stronger NLI and sentence-classification generalization.
MARBERT-MRL	MARBERTv2	163M	Pure Matryoshka contrastive training on the same Arabic triplet recipe.	Controlled Arabic dialect-aware backbone comparison.
LaBSE-MRL	LaBSE	471M	Pure Matryoshka contrastive training on the same Arabic triplet recipe.	Controlled cross-lingual backbone comparison.
mE5-MRL	multilingual-E5-small	278M	Pure Matryoshka contrastive training on the same Arabic triplet recipe.	Controlled multilingual retrieval-backbone comparison.
MPNet-MRL	paraphrase-multilingual-mpnet-base-v2	135M	Pure Matryoshka contrastive training on the same Arabic triplet recipe.	Controlled multilingual paraphrase-backbone comparison.

Table 3: GATE model family and controlled comparison variants. ATM-V2 is optimized for peak Arabic STS performance, whereas GATE-v1 adds hybrid STS–NLI supervision for broader semantic-task generalization.

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 using the `MultipleNegativesRankingLoss`, equivalent to a softmax-based contrastive loss with in-batch and labeled negatives:

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

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-v1 is trained jointly on Arabic semantic textual similarity (STS) and three-way natural language inference (NLI). The STS objective uses `CosineSimilarityLoss` to regress the cosine similarity between two sentence embeddings toward the corresponding gold similarity score. The NLI

objective uses `SoftmaxLoss` to classify sentence pairs as entailment, neutral, or contradiction.

We use round-robin multi-dataset batching: STS and NLI mini-batches are interleaved in equal proportion until one dataset is exhausted. At each training step, only the loss associated with the active dataset is applied. This design avoids manually weighting heterogeneous objectives while allowing the encoder to learn both graded semantic similarity and explicit inferential relations.

Unlike ATM-V2, which is optimized solely with contrastive Matryoshka training on Arabic NLI triplets, GATE-v1 combines graded similarity regression with supervised NLI classification. The comparison therefore isolates the effect of adding explicit pair-classification supervision to an Arabic sentence embedding model, and examines the trade-off between STS specialization and broader semantic-task generalization.

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., 2023), includ-

Model	Dim	# Params.	STS17	STS22	STS22-v2	Avg.
ATM-V2	768	135M	85.31	60.70	63.96	69.99
GATE-v1	768	135M	82.78	59.75	63.09	68.54
AraBERTv02	768	135M	54.53	46.86	49.95	50.45
MARBERT-MRL	768	163M	82.18	58.08	61.32	67.19
MARBERTv2	768	163M	60.98	49.92	53.75	54.88
LaBSE-MRL	768	471M	82.46	57.25	60.58	66.76
LaBSE	768	471M	69.07	57.66	60.98	62.57
mE5-MRL	384	278M	80.37	56.34	59.64	65.45
multilingual-e5-small	384	278M	74.62	58.13	61.40	64.72
MPNet-MRL	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 4: Performance comparison of Matryoshka-based models and their base counterparts on Arabic STS tasks from MTEB.

ing STS17 and STS22-v2. Table 4 reports the results.

Matryoshka-based models consistently outperform their corresponding base models across all benchmarks. *ATM-V2* achieves the best overall performance (69.99 average), with particularly strong results on STS17 (85.31), while *GATE-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-v1* (68.54) and *ATM-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 7); (ii) it achieves the highest average across the five Arabic sentence-classification tasks as shown in Table 9; 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-MRL* and *LaBSE-MRL*, 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., 2022); multilingual-E5 uses contrastive multilingual pairs (Wang et al., 2024b); mpnet-v2 distills English paraphrase supervision), leaving less marginal value for a second round of single-language contrastive training.

Impact of loss selection. Table 5 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: *ATM-V2* achieves a 69.99 average and 85.31 on STS17. The hybrid loss ($L_{sts} + L_{cls}$), used by *GATE-v1*, achieves slightly lower but still competitive STS performance (68.54 average) in exchange

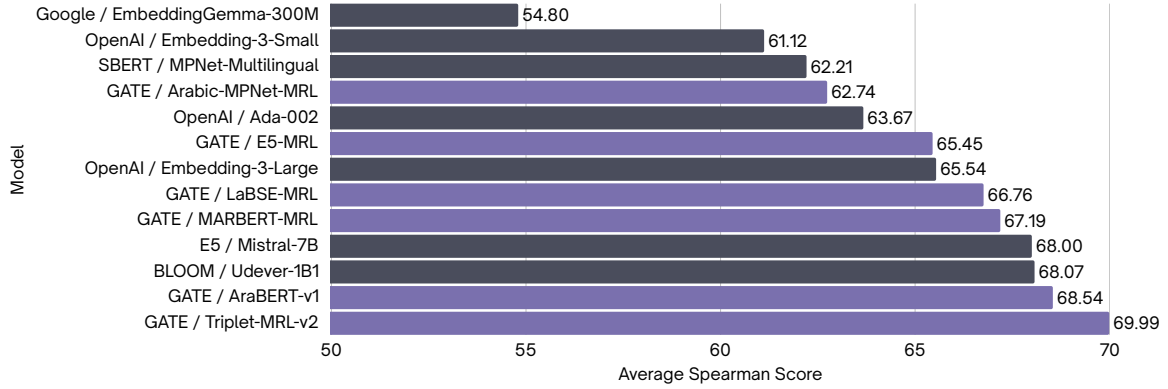


Figure 2: Performance comparison between Matryoshka-based GATE models and larger embedding models on MTEB Arabic benchmarks.

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 5: 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-v1).

Dimensional robustness. Table 6 shows that ATM-V2 retains nearly all of its STS performance after dimensional truncation. Its average score declines by only 0.56 points, from 69.99 at 768 dimensions to 69.43 at 64 dimensions. This small reduction shows that Matryoshka training produces compact prefix embeddings that remain effective for Arabic semantic similarity, enabling substantially lower storage and retrieval cost when full-dimensional vectors are unnecessary.

4.2 Comparison with Large Embedding Models

We compare GATE against large-scale embedding models including *e5-mistral-7b-instruct* (Wang et al., 2024a), *udever-bloom-1b1* (Zhang et al., 2025a), and the OpenAI text-embedding APIs (OpenAI, 2023, 2024) (Figure 2). At 135M parameters, both ATM-V2 and GATE-v1 match or exceed billion-parameter models on Arabic STS, demonstrating that carefully optimized,

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 6: ATM-V2 performance at different Matryoshka embedding dimensions.

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 including 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., 2025b) 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., 2025), 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 7 shows that **ATM-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 at 86.88%, which is 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.

Model	Params	AP	95% CI
ATM-V2	135M	91.55	[89.56, 93.34]
BGE-M3	568M	89.08	[86.70, 91.25]
GATE-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]
<i>XNLI-ar (translated; subset of models)</i>			
Arabic-Triplet-Mat.-V2	135M	85.42	—
GATE-v1	135M	87.46	—
BGE-M3	568M	77.86	—

Table 7: Arabic ArEntail Pair-Classification Results.

Competitive monolingual retrieval without retrieval-specific training. Table 8 evaluates transfer to Arabic retrieval without retrieval-specific training. ATM-V2 achieves 60.91 nDCG@10 and 67.02 MAP on MIRACL-ar retrieval and reranking, respectively, while GATE-v1 reaches 64.60 MAP. Retrieval-tuned baselines remain stronger in absolute terms: BGE-M3 achieves 78.33 nDCG@10 and 79.04 MAP. These results position GATE as a compact Arabic semantic backbone for general-purpose use, while retrieval-tuned models remain preferable when maximizing retrieval performance is the primary goal.

Sentence-level classification: GATE wins. On Arabic sentence-level classification tasks as shown in Table 9, GATE-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.

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

Table 8: MIRACL-ar retrieval and reranking performance.

Model	AJGT	Hotel	Restaur.	Tweet	Store
ATM-V2	87.43	49.73	75.36	49.69	37.02
GATE-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 9: Sentence-level classification accuracy

Dialect robustness on SemRel-2024. SemRel-2024 Arabic (Ousidhoum et al., 2024) provides native semantic-relatedness annotations for MSA, Algerian (arq), and Moroccan (ary). As shown in Table 10, rankings change substantially across varieties. Qwen3-Embedding-4B leads on MSA (64.60) but drops to 24.00 on Algerian and 45.76 on Moroccan. In contrast, ATM-V2, GATE-v1, BGE-M3, and Arctic-Embed-v2 achieve similar Algerian scores (47-49). GATE-v1 obtains the highest Algerian score is 48.84 %, suggesting that dialectal exposure in Arabic specific backbones can persist after hybrid STS-NLI fine-tuning.

Model	MSA	Alg.	Mor.
ATM-V2	61.38	47.68	41.94
GATE-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 10: Per-dialect Spearman correlation on SemRel-2024 Arabic

4.4 Error Analysis

We examine model behavior on three representative Arabic sentence pairs from low-, moderate-, and high-similarity regimes. Tables 11, 12, and 13 report the corresponding cosine similarities for the GATE-family models and multilingual baselines.

رجل يقود سيارة / رجل يعزف على الجيتار “a man playing the guitar” / “a man driving a car” Gold similarity: 0.10	
Model	Cosine
ATM-V2	0.033
GATE-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 11: Low-similarity pair with lexical overlap but unrelated meanings.

الأولاد يلعبون كرة القدم / الرجال يلعبون كرة القدم “men playing football” / “boys playing football” Gold similarity: 0.72	
Model	Cosine
ATM-V2	0.533
GATE-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 12: Moderate-similarity pair with related but non-identical meanings.

Suppressing lexical-overlap false positives.

The low-similarity pair in Table 11 shares the token رجل (“man”) but describes unrelated actions. All multilingual baselines assign a high similarity of 0.496–0.657, whereas ATM-V2 and GATE-v1 assign 0.033 and 0.040, respectively, much closer to the gold score of 0.10. This suggests that Arabic semantic supervision helps reduce lexical-overlap false positives.

Conservativeness on related and paraphrastic pairs. Tables 12 and 13 show the corresponding trade-off. ATM-V2 and GATE-v1 are more conservative than the multilingual baselines on both pairs. However, GATE-v1 assigns a higher score than ATM-V2 to the moderate pair (0.645 versus 0.533), indicating that hybrid STS–NLI training recovers some sensitivity to related meanings while retaining the low-similarity behavior above.

4.5 Availability and Reproducibility

All artifacts required to reproduce this work are publicly available. The *Arabic-NLI-Triplet* training resource, including the triplet, STS, and pair-classification subsets, is available on Hugging

رجل يقوم بخدعة ورق / رجل يقوم بخدعة بالبطاقات “a man doing a card trick” “a man performing a card trick” Gold similarity: 1.00	
Model	Cosine
ATM-V2	0.714
GATE-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 13: High-similarity near-paraphrase pair.

Face.¹ The released GATE-family checkpoints, including ATM-V2 and GATE-v1, are available through the model collection.² Source code, evaluation scripts, configuration files, and commands for reproducing the reported experiments are available at <https://github.com/omarnj-lab/GATE-Experiments>.

For each experiment, we release the model identifier, training configuration, data split, and evaluation command. Training uses an NVIDIA A100 GPU, maximum sequence length 512, batch size 128, and the loss formulations described in Equations 1. Evaluation uses the released checkpoint revisions and the benchmark configurations documented in the repository. The public artifacts are distributed under the licenses specified in their respective repositories.

5 Conclusion

We presented **GATE**, a compact family of Arabic sentence embeddings and a controlled study of Arabic backbones, Matryoshka training, dimensional truncation, and hybrid STS–NLI supervision. Across seven Arabic task categories, semantic-rich Arabic supervision enables compact models to remain competitive with substantially larger multilingual encoders, with particular strength on native-Arabic semantic tasks. The results also reveal a practical trade-off: pure Matryoshka training favors STS performance, while hybrid STS–NLI supervision improves broader classification generalization. Dialect-aware evaluation further highlights that strong MSA averages do not necessarily imply

¹<https://huggingface.co/collections/Omarnj-lab-arabic-nli-and-semantic-similarity-datasets>

²<https://huggingface.co/collections/Omarnj-lab-arabic-matryoshka-and-gate-embedding-models>

robustness across Arabic varieties. Together with publicly released artifacts, GATE provides a practical foundation for Arabic semantic applications.

6 Limitations

GATE is evaluated on Arabic, and the hybrid loss is applied to GATE-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.

Acknowledgments

We gratefully acknowledge Tuwaiq Academy for its valuable support of this work, including research resources, computational infrastructure, and financial support.

References

- Muhammad Abdul-Mageed, AbdelRahim Elmadany, et al. 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.
- 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.
- 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)*.
- Gagan Bhatia, El Moatez Billah Nagoudi, Abdellah El Mekki, Fakhraddin Alwajih, and Muhammad Abdul-Mageed. 2025. *Swan and ArabicMTEB: Dialect-aware, Arabic-centric, cross-lingual, and cross-cultural embedding models and benchmarks*. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4669–4685, Albuquerque, New Mexico. Association for Computational Linguistics.
- Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 632–642.
- Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 1–14.
- 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*, 4(5).
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Ari-vazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 878–891.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6894–6910.
- 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*.
- 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 *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 9726–9735. IEEE.
- Maite Heredia, Julen Etxaniz, Maitze Zulaika, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2024. XN-LIeu: 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*.

- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, system demonstrations*, pages 67–72.
- 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. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.
- Rasha Obeidat, Yara Al-Harabsheh, Mahmoud Al-Ayyoub, and Maram Gharaibeh. 2025. Arentail: manually-curated arabic natural language inference dataset from news headlines: R. obeidat et al. *Language Resources and Evaluation*, 59(1):509–535.
- 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 Benthani. 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.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.
- 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. 2024a. Improving text embeddings with large language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 11897–11916.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.
- Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. *Preprint*, arXiv:1704.05426.
- 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. 2025a. Language models are universal embedders. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 252–265.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025b. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Short name	Full checkpoint name	Description
GATE-v1	GATE-AraBERT-v1	Main hybrid GATE model, trained jointly on Arabic STS similarity regression and three-way NLI classification with round-robin multi-dataset batching.
ATM-V2	Arabic-Triplet-Matryoshka-V2	AraBERTv02-based model trained with pure Matryoshka contrastive learning; the primary STS-oriented checkpoint in the GATE family.
MARBERT-MRL	MARBERT-all-nli-triplet-Matryoshka	MARBERTv2 backbone trained with the pure Matryoshka triplet recipe.
LaBSE-MRL	Arabic-LaBSE-Matryoshka	LaBSE backbone trained with the pure Matryoshka triplet recipe.
mE5-MRL	E5-all-nli-triplet-Matryoshka	multilingual-E5-small backbone trained with the pure Matryoshka triplet recipe.
MPNet-MRL	Arabic-all-nli-triplet-Matryoshka	paraphrase-multilingual-mpnet-base-v2 backbone trained with the pure Matryoshka triplet recipe.

Table 14: Glossary of GATE-family checkpoints and controlled backbone variants.

Term	Meaning	Role in this paper
MRL	Matryoshka Representation Learning.	Learns embeddings whose prefix dimensions remain useful after truncation.
STS	Semantic Textual Similarity.	Measures graded semantic similarity between sentence pairs.
NLI	Natural Language Inference.	Distinguishes entailment, neutral, and contradiction relations.
Arabic-NLi-Triplet	Public Arabic resource derived from translated SNLI and MultiNLI data.	Source of Arabic triplet, STS, and pair-classification supervision.
ArEntail	Natively annotated Arabic NLI benchmark.	Primary native-Arabic pair-classification evaluation.
XNLI-ar	Arabic subset of the Cross-Lingual Natural Language Inference benchmark.	Additional translated Arabic NLI evaluation.
SemRel-2024	Arabic semantic-relatedness benchmark covering MSA, Algerian, and Moroccan varieties.	Dialect-aware semantic-relatedness evaluation.
MTEB	Massive Text Embedding Benchmark.	Source of Arabic STS and downstream embedding evaluations.
MIRACL-ar	Arabic subset of the MIRACL retrieval benchmark.	Monolingual Arabic retrieval and reranking evaluation.

Table 15: Glossary of training resources, benchmarks, and task terminology.

A Glossary of Models, Datasets, & Tasks

This appendix defines the abbreviated model names, datasets, and evaluation terms used throughout the paper. Table 14 maps each short model name to its released checkpoint, while Table 15 defines the training resources and benchmarks.

B Reproducibility Details

This section documents the shared configuration for the pure Matryoshka runs, including ATM-V2 and the controlled backbone variants. Each training example contains an anchor, an entailment-positive sentence, and a contradiction-derived hard negative. We train with MatryoshkaLoss wrapped around MultipleNegativesRankingLoss and evaluate Arabic STS performance at each Matryoshka dimension. Table 16 summarizes the configuration; the public repository provides the complete scripts,

checkpoint configurations, and evaluation commands.

C Translation-Audit Annotation Protocol

This appendix documents the annotation protocol used for the translation audit reported in Section 3.1.1. Three bilingual English–Arabic annotators independently assessed 500 sentence pairs sampled from the *Arabic-NLi-Triplet* resource. Each item received two judgments: Arabic adequacy for each sentence and preservation of the expected NLI relation.

Pair Types and Expected Relations

Each item is a directed slice of an NLI triplet. As summarized in Table 17, anchor–positive pairs are expected to express entailment, whereas anchor–negative pairs are expected to express contradiction.

Setting	Pure Matryoshka configuration
Training framework	sentence-transformers
Training objective	MatryoshkaLoss wrapping MultipleNegativesRankingLoss
Training data	Arabic NLI triplets with entailment positives and contradiction-derived hard negatives
Matryoshka dimensions	768, 512, 256, 128, 64
Epochs	3
Train / evaluation batch	128 / 128 per device
Warm-up ratio	0.10
Precision	fp16
Batch sampler	NO_DUPLICATES
Development evaluation	Arabic STS cosine similarity at each Matryoshka dimension
Evaluation/checkpointing	Every 100 training steps
Checkpoint retention	Two most recent checkpoints
Hardware	NVIDIA A100 GPU

Table 16: Shared reproducibility settings for pure Matryoshka training.

Pair type	Expected relation
<i>Anchor-positive</i>	Entailment: the positive sentence should follow from, or be semantically supported by, the anchor.
<i>Anchor-negative</i>	Contradiction: the negative sentence should be incompatible with the anchor.

Table 17: Pair types evaluated in the translation audit.

Arabic Adequacy Assessment

Annotators rated each Arabic sentence independently using the four-point scale in Table 18. Ratings focus on meaning preservation, grammaticality, and comprehensibility; acceptable dialectal variation or differences in word order should not be penalized.

Score	Definition
4	Fully adequate: fluent, grammatical, and semantically coherent Arabic; minor stylistic awkwardness is acceptable.
3	Mostly adequate: the intended meaning is clear, with only minor distortion or awkward phrasing.
2	Partially adequate: substantial ambiguity, meaning distortion, or grammatical errors; the meaning remains partly recoverable.
1	Inadequate: broken, nonsensical, or clearly mistranslated Arabic.

Table 18: Four-point Arabic adequacy rating scale.

Relation-Preservation Assessment

Annotators judged whether the expected relation still held when considering only the Arabic pair.

They recorded *yes* when the relation was preserved and *no* otherwise. For a *no* judgment, annotators entered the corrected relation—entailment, neutral, or contradiction—in the `annotator_relation` field. A blank `annotator_relation` field was interpreted as agreement with the expected relation. All annotations were completed independently, without consultation between annotators.

Aggregation and Reporting

We report per-annotator mean adequacy scores for both sentence positions, mean pairwise quadratic-weighted Cohen’s κ for adequacy, unanimous and majority relation-preservation rates, and Fleiss’ κ for the binary relation-preservation judgment.