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) underpins retrieval, clustering, and semantic understanding, yet remains underexplored for Arabic due to limited high-quality datasets and sentence embedding models. We introduce **GATE**, a General Arabic Text Embedding framework that achieves state-of-the-art performance on Arabic STS benchmarks within MTEB. GATE integrates Matryoshka Representation Learning with a novel hybrid training strategy that combines STS ranking and NLI classification using Arabic triplet data. This design yields embeddings that capture fine-grained Arabic semantics while remaining robust across multiple embedding dimensions. Despite using significantly fewer parameters, GATE outperforms larger multilingual and proprietary models by 20-25% on Arabic STS. Beyond accuracy, GATE enables efficient deployment through variable-dimension embeddings, offering strong performance retention at reduced dimensionalities suitable for resource-constrained settings. Our results demonstrate that combining semantic-rich supervision with multi-dimensional representations provides a practical and effective solution for Arabic sentence embeddings.

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).

Despite their success, current embedding pipelines typically follow a two-stage process: large-scale weakly supervised pre-training followed by fine-tuning on curated text pairs (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).

Prior work has demonstrated that unsupervised pre-training objectives do not necessarily yield sentence embeddings aligned with human semantic judgments. 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, leading to substantial gains on STS benchmarks. Such semantic-rich supervision is especially important for Arabic, where rich morphology, flexible word order, and the absence of diacritics amplify semantic ambiguity.

Arabic exhibits linguistic properties that differ substantially from Indo-European languages such as English, posing unique challenges for sentence embedding models. It is characterized by rich templatic morphology, where words are derived from root-pattern combinations, leading to high surface-form sparsity and weak lexical overlap between semantically related sentences (Habash, 2010; Antoun et al., 2020). Additionally, standard written Arabic typically omits diacritics, resulting in orthographic ambiguity where identical surface forms may correspond to multiple meanings depending on context (Abdul-Mageed et al., 2020). Arabic also allows relatively flexible word order compared to the more rigid syntactic structures of English, further complicating sentence-level semantic alignment. As a consequence, embedding models trained primarily on Indo-European languages tend to over-rely on surface-level lexical cues, which are less informative in Arabic. These characteris-

Work	Embedding Size	Primary Language Focus	Hybrid Loss	Multi-Dimensional	Semantic-Rich Fine-Tuning
OpenAI Text-Embedding v3 (OpenAI, 2023)	1536 / 3072	Multilingual	✗	✓	✗
E5-Mistral-7B-Instruct (Wang et al., 2023)	4096	English-Focused	✗	✗	✗
Udever-Bloom-1B1 (Zhang et al., 2023)	1536	Multilingual	✗	✗	✗
EmbeddingGemma (Vera et al., 2025)	768,512,256,128	Multilingual	✗	✓	✗
AraBERT (Antoun et al., 2020)	768	Arabic-Specific	✗	✗	✗
MARBERT (Abdul-Mageed et al., 2020)	768	Arabic-Specific	✗	✗	✗
LaBSE (Feng et al., 2020)	768	Multilingual	✗	✗	✗
Multilingual E5 (Wang et al., 2024)	384	Multilingual	✗	✗	✗
GATE Models (Proposed)	**768, 512, 256, 128, 64**	**Semantic Arabic-Specific**	✓	✓	✓

Table 1: Comparison of GATE with recent Arabic, multilingual, and lightweight embedding models by training strategy and representation design.

tics motivate the use of semantic-rich supervision, such as Natural Language Inference and STS objectives, to encourage embeddings to encode deeper semantic relations beyond lexical similarity.

To address these challenges, we introduce **GATE**, a General Arabic Text Embedding model that integrates Matryoshka Representation Learning (MRL) (Kusupati et al., 2022) with a multi-task hybrid training strategy. GATE jointly optimizes semantic ranking for Semantic Textual Similarity (STS) and classification for Natural Language Inference (NLI) using Arabic-specific datasets and hard negatives, overcoming the limitations of single-loss, InfoNCE-only training (Gutmann and Hyvärinen, 2010). Through semantic-rich Arabic fine-tuning and multi-dimensional representation learning, GATE produces scalable embeddings that retain strong performance across multiple output dimensions while capturing fine-grained Arabic semantics.

2 Related Work

Semantic Textual Similarity (STS) (Cer et al., 2017) is a core NLP task that 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). As a result, STS has become a primary benchmark for evaluating the quality of sentence embedding models.

Recent advances in representation learning have shown that sentence embeddings derived solely from masked language modeling objectives are poorly aligned with human judgments of semantic similarity. Reimers and Gurevych (Reimers, 2019) demonstrated that fine-tuning on Natural Language Inference (NLI) data reshapes the embedding space by explicitly encoding entailment and contradiction relations, leading to substantial gains on STS benchmarks. Subsequent work has adopted semantic-rich supervision using NLI and STS objectives, often with ranking or triplet-based losses, to improve sentence-level semantic alignment. However, most of these approaches focus on English or multilingual settings and rely on fixed-dimensional embeddings, leaving Arabic-specific semantic modeling relatively underexplored.

Matryoshka Representation Learning (MRL) (Kusupati et al., 2022) addresses efficiency and scalability by learning hierarchical embeddings that remain effective when truncated to lower dimensions. Recent embedding systems, including OpenAI's third-generation models (OpenAI, 2024), demonstrate the practical value of multi-dimensional representations for deployment. At the same time, large language models 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 expensive and not optimized for Arabic-specific semantic distinctions.

In Arabic NLP, models such as *AraBERT* (Antoun et al., 2020) and *MARBERT* (Abdul-Mageed et al., 2020) improve language understanding by adapting pretrained transformers to Modern Standard Arabic and dialectal content. Multilingual models, including *LaBSE* and *Multilingual E5*, extend coverage across languages but continue to struggle with fine-grained Arabic semantics in STS tasks (Wang et al., 2024). More recently, lightweight embedding models such as *EmbeddingGemma* (Vera et al., 2025) demonstrate strong performance–efficiency trade-offs for general-purpose multilingual embedding, yet do not explicitly integrate semantic-rich Arabic supervision or hybrid STS–NLI objectives.

Table 1 summarizes these approaches along key dimensions. In contrast to prior work, GATE integrates semantic-rich NLI supervision with Matryoshka-based multi-dimensional training, explicitly targeting fine-grained Arabic semantic similarity while maintaining efficiency and scalability.

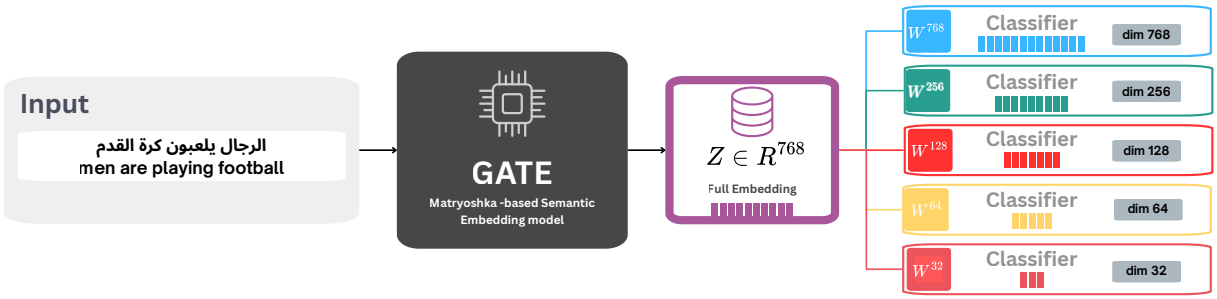

Figure 1: Matryoshka embedding architecture with variable-dimension representations and weight tying.

3 GATE Framework

The GATE framework focuses on Matryoshka representation learning and a multi-task hybrid training approach to enhance Arabic text embeddings. Utilizing the Arabic versions of the Stanford Natural Language Inference (SNLI) and Multi Natural Language Inference (MultiNLI) datasets refines embeddings for optimal performance across various NLP tasks.

3.1 Dataset

Our study utilizes Arabic-adapted subsets derived from the Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) and MultiNLI (Kim et al., 2019) datasets, which are widely used benchmarks for learning sentence-level semantic relations. These datasets provide large-scale, high-quality supervision for entailment, contradiction, and semantic similarity, making them particularly suitable for training embedding models that require fine-grained semantic alignment.

We select SNLI and MultiNLI for two primary reasons. First, they offer diverse sentence pairs with explicitly annotated semantic relations, which have been shown to be effective for training sentence embeddings across languages. Second, comparable Arabic NLI resources remain limited in scale and coverage, necessitating adaptation from well-established multilingual benchmarks.

Subset	Columns	Training	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 2: Datasets used in training and evaluation.

As shown in Table 2, the Triplet subset is used for contrastive training, the STS subset for semantic similarity supervision, and the Pair Classification subset for entailment-based classification within the hybrid loss framework.

Translation Pipeline and Quality Validation. 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, combined with SentencePiece tokenization. Translation quality was validated through manual inspection of randomly sampled sentence pairs across all subsets, focusing on semantic consistency rather than surface-level fluency. Particular attention was paid to preserving entailment and contradiction relations, as errors in these cases would directly impact training signals. Samples with clear semantic drift or mistranslation were discarded during dataset construction.

While automatic translation may introduce noise, prior work has shown that NLI-based embedding training is robust to moderate translation imperfections. Moreover, the scale and semantic diversity of the translated datasets mitigate the impact of individual translation errors. Empirically, the strong performance of GATE on Arabic STS benchmarks suggests that the resulting training data provides reliable semantic supervision.

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) enables the learning of multi-granular sentence embeddings that remain semantically meaningful when truncated to lower dimensions. Instead of training a single fixed-size representation, MRL enforces semantic consistency across a hierarchy of embedding sizes, making it well-suited for resource-constrained and large-scale NLP applications, including Arabic NLP.

Formally, given an input sentence x, a neural encoder $F(\cdot;\theta_F)$ produces a high-dimensional embedding $\mathbf{z} \in \mathbb{R}^d$. MRL enforces that every prefix sub-vector $\mathbf{z}_{1:m} \in \mathbb{R}^m$, for $m \in M$, can independently represent the input. The set M is defined as a sequence of progressively reduced dimensions, in our case $M = \{768, 512, 256, 128, 64\}$.

Given a labeled dataset $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, MRL optimizes a classification loss at each dimension m. The overall Matryoshka loss is defined as:

$$\mathcal{L}_{\text{MRL}} = \sum_{m \in M} c_m \, \mathcal{L}_{\text{CE}}\big(\mathbf{W}^{(m)} \mathbf{z}_{1:m}, y\big), \quad (1)$$

where $\mathcal{L}_{\text{CE}}$ denotes the softmax cross-entropy loss, $\mathbf{W}^{(m)} \in \mathbb{R}^{L \times m}$ is the classifier for dimension m, and c_m is a fixed weighting coefficient. In our experiments, all c_m are set uniformly ($c_m = 1$), giving equal importance to all embedding dimensions. To reduce memory overhead, we employ weight tying across classifiers by setting $\mathbf{W}^{(m)} = \mathbf{W}_{1:m}$, following the Efficient MRL formulation proposed by Kusupati et al. (2022).

Variable-Dimension Embeddings and Weight Tying. In the proposed framework, a single forward pass through the encoder produces a maximum-dimensional sentence embedding $\mathbf{z} \in \mathbb{R}^d$. Lower-dimensional embeddings are obtained by prefix truncation, $\mathbf{z}_{1:m} \in \mathbb{R}^m$, where $m \in M$ and $m < d$. Importantly, no re-encoding or additional projection layers are used; all embedding dimensions share the same encoder parameters. This design ensures that variable-dimension representations are computationally efficient and semantically consistent across dimensions (Kusupati et al., 2022).

To avoid parameter duplication, a single classifier matrix $\mathbf{W} \in \mathbb{R}^{L \times d}$ is learned. For each dimension m, the corresponding classifier is obtained by truncation, $\mathbf{W}^{(m)} = \mathbf{W}_{1:m}$. During training, gradients from losses computed at different dimensions jointly update the shared encoder and classifier weights, encouraging early embedding dimensions to encode the most salient semantic information. This mechanism acts as a form of regularization, improving robustness and enabling high-quality embeddings even at low dimensionalities (Kusupati et al., 2022; Reimers, 2019).

Triplet-Based Contrastive Training. To learn semantically meaningful sentence representations, we train our Matryoshka models using a triplet-based contrastive objective derived from Natural Language Inference (NLI) data. Each training instance consists of a triplet (x, x^+, x^-), where x is an anchor sentence, x^+ is a semantically related sentence (entailment), and x^- is a semantically dissimilar sentence (contradiction). Contradiction pairs from NLI are explicitly used as **hard negatives**, providing a stronger training signal than randomly sampled negatives.

Rather than using a margin-based hinge triplet loss, we adopt the *MultipleNegativesRankingLoss*, which is equivalent to a softmax-based contrastive loss with in-batch and labeled negatives. For a batch of n anchor-positive pairs, the loss is defined as:

$$\mathcal{L}_{\text{triplet}} = -\frac{1}{n} \sum_{i=1}^{n} \log \frac{\exp\big(s(x_i, x_i^+)/\tau\big)}{\sum_{j=1}^{n} \exp\big(s(x_i, x_j^+)/\tau\big)}, \quad (2)$$

where $s(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature parameter. All non-matching positive examples within the batch implicitly serve as negatives, while contradiction pairs further strengthen negative separation. This formulation provides a smooth alternative to margin-based triplet losses and has been shown to produce high-quality sentence embeddings.

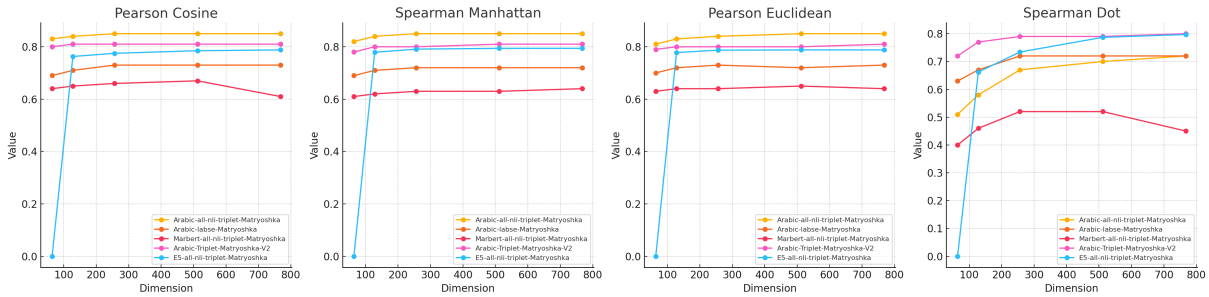

Figure 2: Results of correlation-based similarity metrics across embedding dimensions.

Matryoshka-Aware Optimization. To support multi-dimensional representations, the triplet loss is wrapped by *MatryoshkaLoss*, which applies the contrastive objective independently at each embedding dimension $m \in M$. This ensures that semantic alignment between anchor, positive, and negative examples is preserved across all truncated representations.

In practice, we train our models on the *arabic-nli-triplet* dataset (558K triplets) using embeddings of sizes $[768, 512, 256, 128, 64]$. Training is performed on an NVIDIA A100 GPU with a batch size of 128 and a maximum sequence length of 512 tokens. All experiments are implemented using the *SentenceTransformerTrainer*, which manages optimization, evaluation, and checkpointing across embedding dimensions.

3.2.2 Hybrid Loss Training Approach

To address the limitations of single-objective embedding training, we adopt a multi-task hybrid training strategy that optimizes sentence embeddings for both semantic similarity and semantic discrimination. Training proceeds via alternating optimization over multiple datasets, where each mini-batch is associated with a task-specific loss. This design avoids manual loss weighting while preserving task specialization.

NLI Classification Objective. For natural language inference, we formulate the task as a pairwise classification problem over premise–hypothesis pairs labeled as entailment, neutral, or contradiction. We employ a Softmax-based classification loss operating directly on sentence embeddings. Given a premise x_i, its correct hypothesis y_i^+, and a set of hypotheses with incorrect labels $\{y_{i,j}^-\}_{j=1}^k$, the loss is defined as:

$$L_{\text{cls}} = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\exp\big(s(x_i, y_i^+)/\tau\big)}{\sum\limits_{y \in \{y_i^+\} \cup \{y_{i,j}^-\}} \exp(s(x_i, y)/\tau)} \tag{3}$$

where $s(\cdot,\cdot)$ denotes cosine similarity between sentence embeddings and τ is a temperature scaling parameter. Negatives are label-based rather than in-batch, encouraging explicit semantic separation.

STS Similarity Ranking Objective. For Semantic Textual Similarity (STS), we adopt a cosine similarity ranking loss that enforces correct relative ordering between sentence pairs. Given two sentence pairs (x_i, x_j) and (x_m, x_n) with gold similarity scores such that $s(x_i, x_j) > s(x_m, x_n)$, the loss is defined as:

$$L_{\text{sts}} = \log\left(1 + \sum \exp\left(\frac{\cos(x_m, x_n) - \cos(x_i, x_j)}{\tau}\right)\right) \tag{4}$$

where the summation ranges over all misordered sentence pairs and $\cos(\cdot)$ denotes cosine similarity.

Implicit Margin Property. Although Equation 4 does not include an explicit margin hyperparameter, the log-sum-exp formulation implicitly enforces a margin by penalizing cases where a less similar pair attains higher cosine similarity than a more similar one. This smooth ranking objective avoids manual margin tuning while maintaining strong separation in the embedding space.

Training Procedure. During training, each mini-batch is sampled from a single dataset and optimized using its corresponding loss. Formally, the objective at each training step is:

$$L = \begin{cases} L_{\text{cls}}, & \text{NLI batch}, \\ L_{\text{sts}}, & \text{STS batch}. \end{cases} \tag{5}$$

This alternating optimization enables the model to jointly learn semantic discrimination from NLI

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

supervision and fine-grained semantic alignment from STS supervision, improving generalization across downstream tasks.

4 Results and Discussion

4.1 Robustness Across Embedding Dims

We evaluate the robustness of Matryoshka embeddings across multiple dimensionalities (768, 512, 256, 128, and 64) using correlation-based similarity metrics. Following standard STS practice, we report both Pearson correlation, which measures linear agreement with ground-truth similarity scores, and Spearman correlation, which captures rank consistency between sentence pairs. Figure 2 summarizes these results across cosine, Manhattan, Euclidean, and dot-product similarity functions.

Overall, higher-dimensional embeddings (768, 512) yield the strongest performance, while lower-dimensional variants (128, 64) exhibit moderate degradation, most notably for dot-product similarity. Across metrics, *Arabic-Triplet-Matryoshka-V2* and *Arabic-all-nli-triplet-Matryoshka* demonstrate consistently high correlations, maintaining strong performance even under dimensional truncation. Other Matryoshka variants show similar trends, confirming that MRL enables a favorable trade-off between embedding efficiency and semantic accuracy.

4.2 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, STS22, and STS22-v2. Table 3 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 slightly lower performance of *GATE-AraBERT-V1* compared to *Arabic-Triplet-Matryoshka-V2* reflects the expected trade-off introduced by hybrid loss training, which prioritizes generalization across both STS and NLI tasks.

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. Overall, these results confirm that combining Matryoshka Representation Learning with semantic-rich objectives leads to robust and efficient Arabic sentence embeddings.

Impact of Semantic-Rich Fine-Tuning. Models trained with semantic-rich supervision (NLI and STS objectives) consistently outperform their base counterparts across Arabic STS benchmarks. For instance, *Arabic-Triplet-Matryoshka-V2* improves over its base model by more than 19 points on STS17 and over 13 points on average, demonstrating that contrastive pretraining alone is insufficient for fine-grained Arabic semantics. *GATE-AraBERT-v1* similarly achieves strong gains while remaining competitive with substantially larger multilingual models, supporting the effectiveness of explicit entailment and similarity supervision for

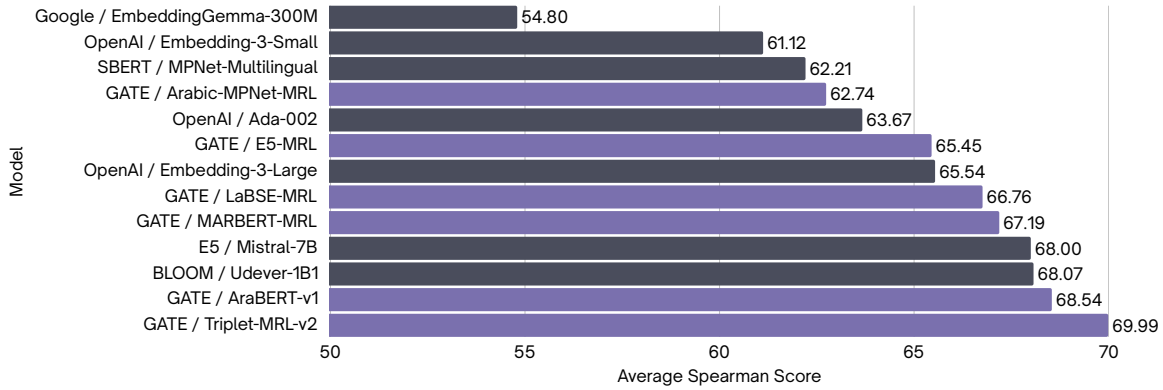

Figure 3: Performance comparison between Matryoshka models and larger models on MTEB Arabic benchmarks.

Arabic sentence embeddings.

Loss	STS17	STS22	STS22-v2	Average
L_{CE}	54.53	46.86	49.95	50.45
L_{MRL}	**85.31**	**60.7**	**63.96**	**69.99**
$L_{sts} + L_{cls}$	82.78	59.75	63.09	68.54

Table 4: Effect of Matryoshka and hybrid loss functions on Arabic STS benchmarks

Table 4 illustrates the impact 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. In contrast, Matryoshka loss (L_{MRL}) yields the strongest results, with *Arabic-Triplet-Matryoshka-V2* achieving 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 performance (68.54 average), reflecting a trade-off between task generalization and STS-specific optimization.

Beyond accuracy, Matryoshka Representation Learning enables strong performance retention under dimensional reduction, as shown in Table 5. Even at 64 dimensions, performance degradation remains minimal, highlighting MRL's suitability for efficient deployment in resource-constrained settings.

As shown in Table 5, results demonstrate that the model maintains robust performance across all dimensions. At the full 768-dimensional embedding, the model achieves an average score of 69.99, with 85.31 on STS17. Even when reduced to 512 and 256 dimensions, the performance remains nearly unchanged, with average scores of 69.92 and 69.86, respectively. Even at the lowest dimension of 64, the model still delivers a strong average score of 69.43, confirming that MRL allows for significant compression without substantial loss in accuracy.

4.3 Comparison of GATE Models with LLMs

We compare GATE models against large-scale embedding models, including *e5-mistral-7b-instruct*, *udever-bloom-1b1*, and OpenAI embedding APIs. As shown in Figure 3, Matryoshka-based models achieve competitive or superior Arabic STS performance despite using significantly fewer parameters. In particular, *Arabic-Triplet-Matryoshka-V2* and *GATE-AraBERT-v1*, with only 135M parameters, outperform billion-parameter models on Arabic STS benchmarks. These results highlight the efficiency of the Matryoshka framework, demonstrating that carefully optimized, smaller models can match or exceed the performance of much larger LLMs for fine-grained semantic similarity tasks.

4.4 Error Analysis

We conduct a targeted error analysis to examine how Arabic-trained Matryoshka models behave across different semantic similarity regimes. Rather than focusing solely on aggregate correlation scores, we analyze model predictions for representative sentence pairs drawn from low, moderate, and high similarity ranges. These categories correspond to distinct regions of the STS annotation scale and allow us to qualitatively assess systematic prediction biases.

Across low-similarity cases (Table 6), models tend to overestimate semantic similarity, frequently assigning scores in the 0.3–0.5 range despite near-zero ground truth labels. This pattern indicates a false-positive bias driven by lexical overlap, where

Evaluation Dim.	STS17	STS22	STS22-v2	Average
768	85.31	**60.7**	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	63.8	69.43

Table 5: Impact of embedding dimensions on the performance of *Arabic-Triplet-Matryoshka-V2*.

Model	Score	Sentence1	Sentence2
Ground Truth	0.1	رجل يعزف على الجيتار (A man playing the guitar)	رجل يقود سيارة (A man driving a car)
Arabic-all-nli-triplet-Matryoshka	0.48		
Arabic-Triplet-Matryoshka-V2	0.48		
GATE-AraBert-v1	0.04		
Arabic-labse-Matryoshka	0.32		
Marbert-all-nli-triplet-Matryoshka	0.38		

Table 6: Model scores for a no similarity sample.

Model	Score	Sentence1	Sentence2
Ground Truth	0.72	الرجال يلعبون كرة القدم (men are playing football)	الأولاد يلعبون كرة القدم (boys are playing football)
Arabic-all-nli-triplet-Matryoshka	0.685		
Arabic-Triplet-Matryoshka-V2	0.661		
GATE-AraBert-v1	0.81		
Arabic-labse-Matryoshka	0.835		
Marbert-all-nli-triplet-Matryoshka	0.836		

Table 7: Model scores for a moderate similarity sample.

Model	Score	Sentence1	Sentence2
Ground Truth	1	رجل يقوم بخدعة بالبطاقات (A man doing a card trick)	رجل يقوم بخدعة ورق (A man performing a card trick)
Arabic-all-nli-triplet-Matryoshka	0.91		
Arabic-Triplet-Matryoshka-V2	0.87		
Arabic-labse-Matryoshka	0.84		
Marbert-all-nli-triplet-Matryoshka	0.85		
GATE-AraBert-v1	0.73		

Table 8: Model scores for a high similarity sample.

shared tokens (e.g., common subjects) are incorrectly interpreted as semantic relatedness. Among the evaluated models, *GATE-AraBERT-V1* exhibits the strongest suppression of this effect, producing a score of 0.04, suggesting that hybrid loss training improves discrimination between lexically similar but semantically unrelated pairs.

For moderate similarity examples (Table 7), predicted scores across models cluster more tightly around the ground truth, typically ranging between 0.66 and 0.83. This reduced variance indicates greater stability when semantic overlap is present but not exact. In this regime, Matryoshka-based models consistently maintain relative ordering, with *Marbert-all-nli-triplet-Matryoshka* and *Arabic-labse-Matryoshka* producing the closest alignment to gold scores.

In high-similarity cases (Table 8), all models achieve strong agreement with ground truth annotations, with scores exceeding 0.84. This suggests that near-paraphrase detection is well captured by all evaluated architectures. However, *GATE-AraBERT-V1* produces slightly more conservative estimates, which we attribute to the influence of classification supervision in hybrid loss training. While this may marginally reduce peak similarity scores, it improves robustness against false positives in low-similarity scenarios.

Overall, the error analysis reveals a consistent trade-off between sensitivity and conservativeness across models. Hybrid loss training reduces overestimation in low-similarity cases, while Matryoshka-only objectives maximize alignment in high-similarity regimes. These trends complement the aggregate correlation results and provide insight into model behavior beyond global metrics.

5 Conclusion

In this work, we introduced GATE, a General Arabic Text Embedding model leveraging MRL and hybrid loss training to enhance STS tasks. Evaluations on MTEB benchmarks confirmed strong performance retention across reduced dimensions and improved generalization over larger models. GATE fills key gaps in Arabic NLP by optimizing embeddings for fine-grained Arabic semantics. Future work will extend Arabic NLP benchmarks, diversify datasets, and explore multilingual generalization for broader real-world impact.

6 Limitations

This work presents certain limitations. The lack of comprehensive Arabic NLP benchmarks restricts a broader evaluation beyond STS tasks. Additionally, error analysis reveals a tendency to overestimate similarity in unrelated sentence pairs, often due to shared lexical elements, leading to false positives. Enhancing negative pair handling could further refine model accuracy. While our approach is optimized for Arabic, the methodology holds the

potential for multilingual adaptation, expanding its applicability.

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*.

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*.

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*.

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.

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.

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

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

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

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*.

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*.