Swan and ArabicMTEB: Dialect-Aware, Arabic-Centric, Cross-Lingual, and Cross-Cultural Embedding Models and Benchmarks

Gagan Bhatia [ξ] **El Moatez Billah Nagoudi** [ξ,γ] **Abdellah El Mekki** [ω] **Fakhraddin Alwajih** [ξ]
Muhammad Abdul-Mageed[ξ,ω,λ]
[ξ] The University of British Columbia, [ω] MBZUAI, [γ] PSU, [λ] Invertible AI
{gagan30@student.,muhammad.mageed@}ubc.ca

Abstract

We introduce **Swan**, a family of embedding models centered on Arabic, designed for both small-scale and large-scale applications. Swan comprises two variants: Swan-Small, built on ARBERTv2, and Swan-Large, based on ArMistral, a pretrained Arabic large language model. To evaluate our models, we propose a comprehensive benchmark suite, dubbed ArabicMTEB, that assesses cross-lingual, multi-dialectal, multi-domain, and multi-cultural Arabic text embedding performance. ArabicMTEBcovers eight diverse tasks sourced from 94 datasets. Swan-Large achieves state-of-the-art results, outperforming Multilingual-E5-large in most Arabic tasks, while Swan-Small consistently surpasses Multilingual-E5-base. Our extensive evaluations show that Swan models are both dialectally and culturally aware, achieving strong performance across diverse Arabic domains while maintaining significant cost efficiency. This work significantly advances the field of Arabic language modelling and provides valuable resources for future research and applications in Arabic NLP. Our models and benchmark are available at our GitHub page: https://github.com/UBC-NLP/swan.

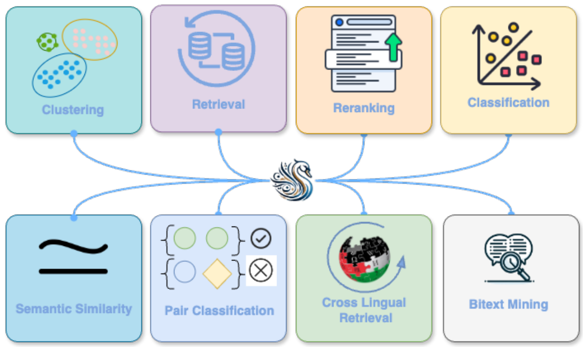

Figure 1: Overview of our ArabicMTEB benchmark tasks, covering *clustering, retrieval, reranking, classification, semantic similarity, pair classification, cross-lingual retrieval,* and *bitext mining*.

1 Introduction

NLP has seen rapid advancements in recent years, driven by groundbreaking developments in deep learning and the emergence of sophisticated distributed text representations such as word and sentence embeddings (Devlin et al., 2018; Reimers and Gurevych, 2019). These embeddings, which transform text into dense vectors, enable effective semantic understanding and are pivotal for enhancing performance across many downstream applications, including text classification, semantic search, and machine translation. Moreover, text embeddings have become fundamental to the success of large language models (LLMs) (Touvron et al., 2023; Jiang et al., 2023; Gemma-Team et al., 2024), which are increasingly integrated into a variety of real-world systems and tools. One of the most promising applications of these embeddings is in the realm of Retrieval-Augmented Generation (RAG) (Shao et al., 2023; rag, 2023), where LLMs are augmented with information retrieval capabilities. In RAG-based systems, lightweight embedding models retrieve relevant information from large corpora, fed as context to models like ChatGPT (OpenAI, 2023) or GPT-4 (OpenAI et al., 2024). This synergy between embeddings and LLMs has demonstrated significant improvements in both general-purpose tasks such as question answering (Lin et al., 2023; rag, 2023) and domain-specific applications (Bhatia et al., 2024; Shi et al., 2023; Lin et al., 2023).

Despite these advances, the predominant focus of current embedding models has been on English and Chinese, which limits their applicability to other languages. This is particularly true for Arabic, a collection of languages, language varieties, and diverse dialects with rich morphology (Abdul-Mageed et al., 2023a, 2024a), making it challenging to develop effective language representations (Nagoudi et al., 2022; Huang et al.,

Findings of the Association for Computational Linguistics: NAACL 2025, pages 4669–4685
April 29 - May 4, 2025

2024). Existing multilingual models often fail to capture these complexities, leading to a suboptimal performance on Arabic NLP tasks (Abdul-Mageed et al., 2020a; Elmadany et al., 2022). Addressing this limitation requires the development of Arabic-specific embedding models that are sensitive to the linguistic and cultural nuances of Arabic.

In this work, we introduce Swan, a family of dialect-aware, Arabic-centric, cross-lingual, and cross-cultural embedding models designed to bridge this gap and push the boundaries of Arabic NLP. Our contributions are as follows:**(1)** We introduce Swan, a cutting-edge family of Arabic embedding models. This includes two variants: Swan-Small, based on ARBERTv2 (Elmadany et al., 2022), and Swan-Large, built upon ArMistral, a further pretrained Arabic language model. **(2)** We present ArabicMTEB, a comprehensive evaluation benchmark for Arabic text. ArabicMTEB is designed to assess cross-lingual, multi-dialectal, multi-domain, and multi-cultural performance, spanning eight tasks and 94 datasets. Figure 1 provides an overview of ArabicMTEB. **(3)** Our larger model, Swan-Large, showcases state-of-the-art text embedding capabilities, surpassing Multilingual-E5-large (Wang et al., 2024b) on most Arabic tasks. Similarly, our smaller, Swan-Small, consistently outperforms Multilingual-E5-base (Wang et al., 2024b) on most Arabic tasks. **(4)** Through rigorous benchmarking, we demonstrate that Swan models are not only dialectally and culturally aware, but also excel across diverse Arabic domains while maintaining a significantly lower monetary cost.

The rest of the paper is organized as follows: In Section 2, we review related work with a particular emphasis on Arabic text embedding models, their applications and challenges. We present our approach to model training of Swan models in Section 3. Section 4 outlines how we built our benchmark dataset, ArabicMTEB. Section 5 is about our experiments and model analysis. We conclude in Section 6.

2 Related Work

In recent years, there have been remarkable advancements in text embedding models, with a shift towards developing universal embeddings for diverse tasks and domains. Despite this, specialized models and benchmarks for languages like Arabic remain underexplored.

Multilingual Text Embedding Models. With the need for language-agnostic embeddings growing, multilingual models such as LASER (Artetxe and Schwenk, 2019) and LaBSE (Feng et al., 2022) were developed using BiLSTM and Transformer encoders, respectively. Building on this, the Multilingual-E5 (Wang et al., 2024c) series extends the E5 architecture to support diverse languages using multilingual text pairs and synthetic data. GRIT (Muennighoff et al., 2024) further unifies generative and embedding tasks within a single model. Newer models such as ColBERT-XM (Louis et al., 2024) and Gecko (Lee et al., 2024) refine multilingual embeddings through modular and distilled architectures.

Arabic-Specific Models. Despite progress in Arabic NLP, existing models are not optimized for Arabic text embedding and retrieval. Efforts like ARBERT (Abdul-Mageed et al., 2021a) and AraMus (Alghamdi et al., 2023) have focused on encoding and generation but are not tailored for sentence-level embeddings. While language-agnostic models such as LASER and Multilingual-E5 include Arabic in their training data, they may not fully capture its linguistic intricacies and diversity. To address this, Nacar and Koubaa (2024) introduced models and training datasets to improve semantic similarity performance for Arabic.

Text Embedding Benchmarks. Most text embedding evaluations rely on a narrow set of datasets, limiting their generalisation ability. To address this, the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023) introduced eight task categories with 58 datasets and 112 languages. However, it remains predominantly focused on English. Similar benchmarks have been developed for other languages, such as C-MTEB (Xiao et al., 2023) for Chinese. For Arabic, evaluations have primarily centred on Semantic Text Similarity (STS) tasks (Nacar and Koubaa, 2024). However, excelling in STS does not guarantee optimal performance in tasks like clustering or reranking (Muennighoff et al., 2023). Existing Arabic benchmarks like ORCA (Elmadany et al., 2023) and ALUE (Seelawi et al., 2021) focus on natural language understanding (NLU), while Dolphin (Nagoudi et al., 2023a) targets natural language generation (NLG). This work is the first comprehensive benchmark for evaluating Arabic text embeddings across multiple tasks.

Benchmark	Lang	Tasks	Datasets	Tasks	CRTR	Ar Cul/Dom
MTEB (Muennighoff et al., 2022)	English	*RTR, STS, PairCLF, CLF, RRK, CLR, SUM*	56	7	×	×
C-MTEB (Xiao et al., 2023)	Chinese	*RTR, STS, PairCLF, CLF, RRK, CLR*	35	6	×	×
De-MTEB (Sturua et al., 2024)	German	*RTR, STS, PairCLF, CLF, RRK, CLR*	17	6	×	×
F-MTEB (Ciancone et al., 2024)	French	*RTR, STS, PairCLF, CLF, RRK, CLR, BTM*	17	7	×	×
Es-MTEB(Mohr et al., 2024)	Spanish	*RTR, STS, PairCLF, CLF, RRK, CLR*	17	6	×	×
Polish (Poświata et al., 2024)	Polish	*RTR, STS, PairCLF, CLF, CLR*	26	5	×	×
Ru-MTEB (Poświata et al., 2024)	Russian	*RTR, STS, PairCLF, CLF, RRK, CLR*	23	6	×	×
	Danish				×	×
Scand. (Enevoldsen et al., 2024)	Norweg.	*RTR, CLF, BTM, CLR*	26	4	×	×
	Swedish				×	×
`ArabicMTEB` (Ours)	Arabic	*RTR, STS, PairCLF, CLF, RRK, CLR, BTM, CRTR*	94	8	✓	✓

Table 1: Comparison of various text embedding benchmarks proposed in the literature across the different covered task clusters. **RTR**: retrieval, **STS**: semantic textual similarity, **PairCLF**: pair classification, **CLF**: classification, **CLR**: clustering, **RRK**: reranking, **BTM**: bitext mining, **CRTR**: cross-lingual retrieval.

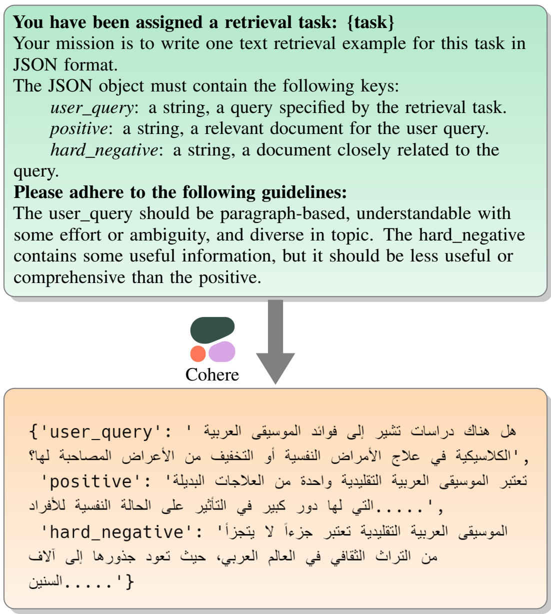

Figure 2: Methodology to generate our synthetic data.

Family	Language	Type	Dataset	Level	Size
Monoling	Ar	Human	ORCA-MSA	Sent	378K
			ORCA-DIA		122K
			MMARCO-ar		8.1M
		Synthetic	Synth-MSA	Parag	100K
			Synth-DIA		15K
			Synth-DOM		20K
Crossling	Ar to 15 lg	Human	MMARCO	Sent	3M
	Ar to 6 lg		XOR-TyDi		20.5K
Multiling	11 lg	Human	Mr-Tydi	Sent	49K
	16 lg		Miracl		343K
Total					12.5M

Table 2: The diverse datasets employed for training our Arabic embedding models. In the synthetic dataset, we have three datasets: the MSA dataset, the dialectal dataset (Egyptian and Moroccan), and domain-based focusing on medical, financial, legal and news domains.

3 Swan

3.1 Training Data

To train Swan, we develop the most extensive training corpus for Arabic embedding models, leveraging a unique assembly of datasets to ensure comprehensive linguistic coverage and diversity. Our training data covers paragraph-based and sentence-based datasets curated from multiple sources. Table 2 shows an overview of our training datasets.

MSA Datasets. We focus on two sources: **(i) Human-generated data:** Composed from ORCA (Elmadany et al., 2023) and mMARCO (Bonifacio et al., 2021). ORCA is a compilation of labelled datasets with tasks such as semantic text similarity (STS), sentence classification, text classification, natural language inference (NLI), and question answering. We use all the training sets from ORCA, encompassing 60 different datasets. mMARCO-ar is the translated version of MARCO, which is a human-generated dataset (Bajaj et al., 2018). Both of these datasets are cleaned up and de-duplicated using Polydedupe (Bhatia, 2023),[1] which is further described in Appendix C. **(ii) Synthetically-generated data:** To augment our MSA training data for retrieval tasks, we use Command R+ (Cohere For AI, 2024) to generate high-quality synthetic data.[2] The generation methodology is inspired by Wang et al. (2024a), and we employ the procedure shown in Figure 2 to generate our synthetic dataset. We generate $100k$ in general MSA data and $5k$ in instances for specific domains such as finance, news, medicine, and legal for a total of $120k$ MSA instances.

Dialectal Arabic Datasets. Similar to the MSA

[1] https://github.com/gagan3012/PolyDeDupe

[2] We performed various in-house evaluations comparing multiple models. Command R+ was chosen as it is open-source and efficient in generating Arabic varieties (standard and dialectal).

datasets, we focus on two sources: **(i) Human-generated data:** We use publicly available dialectal Arabic data, which primarily covers Gulf, Egyptian, Moroccan, and Levantine varieties of Arabic (Elmadany et al., 2022; Nagoudi et al., 2023b; Alwajih et al., 2024; Abdul-Mageed et al., 2020b, 2021c, 2022, 2023b, 2018, 2020c; Keleg et al., 2023; Keleg and Magdy, 2023; Zaidan and Callison-Burch, 2014; Bouamor et al., 2018). The total number of samples is $122K$. **(ii) Synthetically-generated data:** As most human-generated dialectal data comes from noisy environments such as social media, it often results in short texts of low quality. Thus, we use Command-R+ to generate paragraph-based synthetic data for Egyptian and Moroccan dialects to improve the performance of our models on dialectal Arabic. We generated $15k$ dialectal instances using the same methodology as our synthetic MSA datasets described above.

Cross-Lingual & Multilingual Datasets. To adapt our model for cross-lingual and multilingual scenarios, we incorporate the mMARCO dataset, which provides translations of the MS MARCO dataset into 15 languages (Bonifacio et al., 2021). To ensure that documents correspond accurately to their queries in different languages, we utilize specific IDs. We create $100k$ samples for each cross-lingual pair and shuffle the IDs to prevent repetition, thus guaranteeing that unique data samples are used for each language. We utilize the MIRACL (Zhang et al., 2022), XOR-TyDI (Asai et al., 2021), and Mr.TyDi (Zhang et al., 2021) datasets as our crosslingual and multilingual resources.

3.2 Training Strategy

For `Swan`, we consider two models: `Swan-Small` and `Swan-Large`. The choice of training two models with different sizes is driven by the need to balance performance and computational efficiency. `Swan-Small` is designed to cater to scenarios where lower computational resources are available or when a lightweight model is preferred for deployment on edge devices. In contrast, `Swan-Large` is intended for settings where achieving SoTA performance is paramount, leveraging a larger parameter size to better capture the nuances of Arabic.

Data Preprocessing. We incorporate human-generated and synthetic datasets into our training pipeline to ensure robust performance across various dialects and cultural contexts. We first train on MSA datasets, followedby fine-tuning on dialectal datasets. This two-step approach ensures that both MSA and dialectal varieties are well represented, promoting better generalization across the full spectrum of Arabic varieties. Our dataset is constructed with a query format, including positive and negative samples.

`Swan-Small`. Built upon ARBERTv2 (Abdul-Mageed et al., 2021b), a powerful BERT-based model for the Arabic language. Here, our model is trained using the InfoNCE loss (van den Oord et al., 2019), which maximizes the similarity between related sentences while minimizing the similarity between unrelated sentences. The model is trained for five epochs on the entire dataset with a learning rate of $5e^{-6}$ and a batch size of 128, incorporating 15 hard negatives. `Swan-Small` has 164M parameters and a dimension size of 768.

`Swan-Large`. `Swan-Large` is based on ArMistral-7B, an in-house further pretrained version of Mistral-7B (Jiang et al., 2023)[3]. To train `Swan-Large`, we use LoRA (Hu et al., 2021) for parameter efficient training and InfoNCE loss for optimization. We train the model for three epochs on the entire dataset with a learning rate of $5e^{-6}$ and a batch size of 128, incorporating seven hard negatives. `Swan-Large` has $7.2B$ parameters and a dimension size of $4,096$.

3.3 Training Methodology

Given a relevant query-document pair (q^+, d^+), we modify the query by appending an instructional template to it. This process transforms the original query q^+ into a new form q^+_{inst} as defined below:

$$q^+_{\text{inst}} = \text{Instruction: }\{\text{task\_instruction}\}\text{ Query:}\{q^+\}$$

Here, "*{task_instruction}*" refers to a one-sentence description of the embedding task taken from Table 12, which outlines the instructions for different tasks. Using a pretrained LLM, we append an *[EOS]* token at the end of both the modified query and the document. These are then fed into the LLM to extract embeddings $\mathbf{h}_{q^+_{\text{inst}}}$ and $\mathbf{h}_{d^+}$ from the vector at the last [EOS] layer. Again, training of the embedding model is conducted using the InfoNCE loss function, which is widely recognized for its effectiveness in learning high-quality embeddings. The objective is minimized using the following formulation:

[3]Further details about ArMistral can be found in Appendix A.

$$\min\left(-\log\frac{\phi(q^{+}_{\text{inst}},d^{+})}{\phi(q^{+}_{\text{inst}},d^{+})+\sum_{n_i\in\mathbb{N}}\phi(q^{+}_{\text{inst}},n_i)}\right)$$

In the equation above, $\mathbb{N}$ denotes the set of negative samples, and $\phi(q,d)$ is the similarity scoring function between a query q and a document d.

3.4 Inclusion of Hard-Negatives

To enhance the model's performance, it is crucial to use negative documents closely aligned with the query's context (Karpukhin et al., 2020; Khondaker et al., 2022). This method allows us to observe the impact of introducing more challenging or "hard" negatives into the training process. We only generate hard negatives for the Arabic subset of our training data from Section 3.1. We found that using 15 hard negatives for `Swan-Small` yields the best performance, whereas for our bigger model, `Swan-Large`, the model overfits a more significant number of hard negatives, and 7 gives us the best performance.

Impact of Hard Negatives. Hard negatives in contrastive learning are examples that closely resemble correct instances but are ultimately incorrect. Their inclusion encourages the model to learn finer-grained distinctions, improving its ability to differentiate between similar but distinct classes. The process involves converting all documents into a vector form within the embedding space. Subsequently, these document embeddings are compared using the cosine similarity score to establish their relevance to the query. Once all documents are scored, they are sorted by their similarity to the query. The top-ranked document is typically the positive example, while the rest are potential negatives. Our experiments assess the impact of varying the hard negatives used while training our models, `Swan-Large` and `Swan-Small`. We train each model with different quantities of hard negatives. Namely, we experiment with using hard negatives values from the set {*1, 3, 7, 15, 31*} per training instance. `Swan-Small` achieves its highest performance of 56.25 with 15 hard negatives. The model exhibits a general upward trend as the number of hard negatives increases, peaking at 15 before slightly declining at 31. This pattern suggests that while additional hard negatives initially enhance learning by introducing valuable challenges, excessive complexity may lead to diminishing returns, ultimately hindering further improvement. `Swan-Large` achieves its peak performance of 60.42 when trained with seven hard negatives, suggesting an optimal balance that enhances learning without overloading the model. Notably, increasing the number of hard negatives beyond this point does not lead to further gains, indicating a threshold where additional complexity ceases to improve learning outcomes.

Model (HN)	1	3	7	15	31
`Swan-Small`	48.84	52.19	<u>54.13</u>	**56.25**	51.93
`Swan-Large`	59.48	59.35	**60.42**	59.44	<u>59.83</u>

Table 3: Impact of number of Hard Negatives (HN).

Task	Datasets	Langs	Dialects	Metric
RTR	36	1	4	nDCG@10
CRTR	12	7	0	nDCG@10
CLF	18	1	6	AP
BTM	11	5	8	F1
RRK	5	2	0	MAP
STS	5	1	3	Spearman Corr
CLR	4	1	0	v-measure
PairCLF	3	1	0	AP
Total	**94**	**9**	**11**	

Table 4: Overview of our Tasks in `ArabicMTEB`. *Total represents the unique languages.

4 ArabicMTEB Benchmark

In this section, we present `ArabicMTEB`, a comprehensive Arabic-centric text embedding benchmark designed to evaluate text embeddings across a wide range of tasks and scenarios. `ArabicMTEB` addresses the limitations of existing benchmarks that either exclude Arabic or lack coverage of diverse Arabic language varieties, dialects, and cultural nuances. Our benchmark includes 94 datasets spanning 8 distinct tasks, as summarized in Table 4. Further details about the datasets used in benchmark can be found in Appendix B. `ArabicMTEB` was developed to provide comprehensive coverage of Arabic text embedding capabilities, ensuring the inclusion of MSA and other varieties. It offers diverse task types, such as retrieval, classification, and semantic similarity, to evaluate embeddings holistically across different scenarios by incorporating novel domain-specific, dialectal, and country-level culturally aware datasets, `ArabicMTEB` represents a more applicable and realistic assessment of Arabic text embeddings.

4.1 Task Categories

ArabicMTEB categorizes evaluation datasets into the following key task categories, with each type providing a unique perspective on the capabilities of text embeddings. The corresponding metadata for each task, covering the considered number of datasets, number of languages, number of dialects, and evaluation metric, is presented in Table 4.

Arabic Text Retrieval. This task uses Arabic queries to retrieve Top-k relevant documents from a large Arabic corpus. ArabicMTEB includes 35 retrieval datasets such as XPDA (Shen et al., 2023) and Dolphin's long-form QA datasets (Nagoudi et al., 2023b). Including these datasets helps evaluate complex information retrieval scenarios in Arabic.

Bitext Mining. This task identifies sentence-level translations between different languages and dialects. ArabicMTEB includes 12 datasets spanning various language pairs like Arabic to French and English. This task is crucial for understanding text embeddings' cross-lingual and dialectal translation capabilities.

Cross-Lingual Retrieval. This task uses Arabic queries to retrieve documents in other languages, such as English, German, Spanish, and Chinese. ArabicMTEB employs the mMarco Dev set (Bonifacio et al., 2021) and includes 11 language pairs.

Re-Ranking. This task reorders candidate documents for a query based on embedding similarity scores. ArabicMTEB features five re-ranking datasets such as MIRACL (Zhang et al., 2023), enabling the evaluation of embeddings' ability to refine search results.

Semantic Textual Similarity (STS). STS measures the correlation between the embeddings of two sentences, assessing their semantic similarity. ArabicMTEB includes five STS datasets like STS17 and STS22 (Cer et al., 2017), along with two synthetic datasets generated using GPT-4 (Details of the creation of these datasets can be found in Appendix G).

Classification. Classification predicts labels for input texts using text embeddings. ArabicMTEB comprises 18 multi-domain datasets, from ORCA (Elmadany et al., 2022). This task evaluates models' ability to categorize Arabic text accurately, making it a valuable benchmark for downstream tasks such as sentiment analysis.

Pair Classification. This task predicts the relationship between two sentences based on their embeddings. ArabicMTEB includes three datasets, such as XNLI (Conneau et al., 2018).

Clustering. Clustering groups sentences into clusters based on embedding similarity, evaluating unsupervised learning performance. ArabicMTEB includes four clustering datasets, such as Arabic News Articles and stance detection datasets from (Baly et al., 2018).

4.2 Dialectal ArabicMTEB

Dialectal ArabicMTEB is a specialized fork of the original ArabicMTEB, focusing exclusively on Arabic dialectal datasets. This extension addresses the unique challenges posed by the significant variations in Arabic dialects across different regions, which have been underrepresented in NLP research. While research on dialectal Arabic text embedding has been limited, dialectal ArabicMTEB fills this gap by providing a comprehensive collection of 19 datasets specifically curated to evaluate embeddings' performance on diverse Arabic dialects. These datasets span multiple tasks, offering a robust framework for assessing model performance across various dialectal contexts: **(1) Bitext Mining.** Eight datasets covering dialects such as Algerian, Egyptian, Jordanian, Lebanese, Moroccan, Saudi, and Yemeni (Nagoudi et al., 2022; Bouamor et al., 2014). **(2) Retrieval.** Five datasets focusing on dialects from Algeria, Egypt, Morocco, and the Gulf regions (Nagoudi et al., 2023b). **(3) Classification:** Five datasets for binary, regional, and country-level dialect identification (Abdul-Mageed et al., 2021d, 2024b; Elmadany et al., 2022; Abdul-Mageed et al., 2021b; Ahmed et al., 2024). **(4) STS.** A novel synthetic dataset for Egyptian text similarity generated using Command-R+.

4.3 Domain-Specific ArabicMTEB

Arabic Text retrieval tasks are currently trending in real-world applications. They are utilized across multiple fields, including healthcare, finance, and legal sectors. Having specialized evaluation datasets is crucial for building text embeddings tailored to these domains. To meet this need, we introduce domain-specific ArabicMTEB, a specialized fork of the broader ArabicMTEB benchmark. Domain-specific ArabicMTEB focuses on the news, finance, legal, medical, and general knowledge domains, offering a closer approximation to real-world scenarios. The creation of this benchmark involves collecting Arabic documents from these specialized sources and from Arabic Wikipedia. We

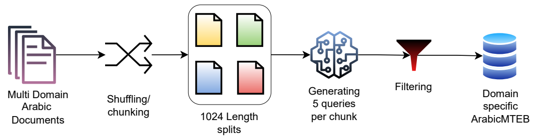

Figure 3: Generation pipeline for our domain specific `ArabicMTEB`.

then segment and chunk the documents into texts of $1,024$ tokens each. Subsequently, we randomly select chunks and employ GPT4-Turbo (OpenAI et al., 2024) to generate five different styles of queries for each chunk. We filter out duplicate and repeated queries using GPT3.5 (OpenAI et al., 2024) to ensure a high-quality evaluation dataset. Our evaluation data creation pipeline is visualized in Figure 2. The resulting benchmark, which we call `ArabicMTEB-Lite`, contains $10k$ queries and $100k$ documents spanning the domains described above.

4.4 Cultural ArabicMTEB

To show that our models are culturally aware, we have introduced Cultural `ArabicMTEB`, a collection of datasets from 20 different Arab countries where we focus on specific cultural aspects like their geography, history, etc. To construct the Cultural `ArabicMTEB`, we use Arabic Wikipedia as our primary data source. For each included Arab country, we extract articles related to that country from its corresponding Wikipedia portal. The portal covers multiple categories (e.g., geography, economy, history) with subcategories (e.g., local movies, food items). This process resulted in 5K to 55K articles per country. Next, we generate retrieval questions and passages for each country. For this, we use GPT-4o-mini to develop, for each passage (from an article), a corresponding question whose specific answer is available within the passage itself. Following the same methodology, but applied to Egyptian and Moroccan dialectal versions of Wikipedia, we generate dialectal queries and their corresponding passages using Command-R+. Cultural `ArabicMTEB` contains $1k$ queries and an average of $15k$ documents from various countries as described above.

5 Evaluation

We evaluate the performance of our models, `Swan-Small` and `Swan-Large`, across the multiple proposed `ArabicMTEB` benchmarks and compare them with existing SoTA models, including MARBERT (Abdul-Mageed et al., 2020a), ARBERTv2 (Elmadany et al., 2022), CamelBERT (Inoue et al., 2021), multilingual E5 models (Wang et al., 2024b), and Arabic-triplet-Matryoshka-V2 (ATM-V2) (Nacar and Koubaa, 2024). Our evaluation results encompasses overall `ArabicMTEB` (Table 5), dialectal `ArabicMTEB` (Table 6), domain-specific `ArabicMTEB` tasks (Table 7), and cultural `ArabicMTEB` (Table 8). In these tables, the tasks will be referred to as **RTR**: Retrieval, **STS**: Semantic Textual Similarity, **PairCLF**: Pair Classification, **CLF**: Classification, **CLR**: Clustering, **RRK**: Reranking, and **BTM**: BiText Mining.

`ArabicMTEB` Results. Table 5 presents the overall results of our models on the `ArabicMTEB` benchmark. Our models demonstrate top-tier performance across a variety of NLP tasks. `Swan-Small` achieves an average score of 57.33, surpassing its main competitors, Me5-base (55.29) and Me5-small (55.06), by a significant margin. This model performs exceptionally well in retrieval (58.42), classification (57.34), and pair classification (74.93), outperforming ATM-V2, which only scores 45.24 on average. Similarly, `Swan-Large` sets a new state-of-the-art performance with an average score of 62.45, beating Me5-large (61.65) and even the massive e5-mistral-7b model (59.00). The model excels particularly in retrieval (65.63), classification (54.89), and bitext mining (71.24), indicating its robustness across both cross-lingual and Arabic-centric tasks. These results validate our training strategy of using diverse training data covering multiple languages, where `Swan-Large` outperforms its counterparts by more than five points in cross-lingual tasks such as bitext mining.

Dialectal `ArabicMTEB` Results. Table 6 shows the dialectal `ArabicMTEBresults`. `Swan-Small` scores an average of 63.41, considerably higher than Me5-small (45.27) and AlcLaM (30.44), showing strong performance across retrieval (63.16) and classification (54.52). `Swan-Large` achieves an impressive average score of 70.45, leading all tasks and outperforming the e5-mistral-7b model, which scores 60.81. The standout result is in bitext mining, which achieves 72.10, showcasing a substantial 14-point improvement over AlcLaM (59.38). Our models' significant advantage in dialectal retrieval and bitext mining is their unique training with a combination of synthetic and human-generated dialectal datasets, which is absent in many competi-

Model	Size	Dim.	RTR	STS	PairCLF	CLF	RRK	CLR	BTM	Avg.
Arabert-v2-base	160M	768	8.62	39.77	66.30	55.77	60.03	41.74	0.70	38.99
CamelBERT	163M	768	9.21	47.69	67.43	55.66	60.20	39.89	1.85	40.28
ARBERTv2	164M	768	15.12	47.88	68.87	56.85	62.21	39.25	1.99	41.74
ATM-V2	135M	768	37.45	55.90	70.12	46.42	61.45	32.35	12.98	45.24
text2vec	118M	384	27.69	**59.37**	71.41	47.94	57.76	37.26	38.32	48.54
LaBSE	471M	768	34.98	54.15	70.60	49.57	62.17	41.42	33.28	49.45
Me5-small	118M	384	55.14	56.73	73.97	50.85	67.92	42.37	38.47	55.06
Me5-base	278M	768	56.91	57.99	74.30	52.30	**69.07**	**42.56**	33.90	55.29
`Swan-Small`	164M	768	**58.42**	59.34	**74.93**	**57.34**	68.43	40.43	**42.45**	**57.33**
e5-mistral-7b	7110M	4096	56.34	57.02	70.24	53.21	66.24	39.44	70.5	59.00
Me5-large	560M	1024	64.01	**59.45**	75.06	53.43	**70.79**	**42.49**	66.33	61.65
`Swan-Large`	7230M	4096	**65.63**	59.10	**75.62**	**54.89**	69.42	41.24	**71.24**	**62.45**

Table 5: Overall `ArabicMTEB` results.

Model	RTR	STS	CLF	BTM	Avg.
Arabert-v2-b	8.67	41.64	47.97	0.99	24.82
MARBERT	5.45	50.06	53.46	2.34	27.83
ARBERTv2	7.52	49.36	54.31	2.51	28.43
CamelBERT	6.92	59.48	50.69	2.65	29.93
AlcLaM	8.56	50.90	**54.74**	7.54	30.44
ATM-V2	36.23	74.13	34.39	11.67	39.10
Me5-base	61.60	74.84	34.87	3.30	43.65
Me5-small	57.61	76.35	34.78	12.35	45.27
Me5-large	66.88	77.02	35.47	51.08	57.61
e5-mistral-7b	72.35	77.37	35.91	57.62	60.81
`Swan-Small`	63.16	76.57	54.52	59.38	63.41
`Swan-Large`	**77.03**	**79.22**	53.46	**72.10**	**70.45**

Table 6: Dialectal `ArabicMTEB` results.

Model	News	Legal	Med	Fin	Wiki	Avg	Cost ↓
`Swan-Large`	**90.42**	**89.96**	**81.64**	57.34	**93.10**	**82.49**	0.75$
Openai-3-lg	88.1	89.68	80.24	**61.46**	91.52	82.20	9.88$
Cohere-v3.0	85.23	86.52	63.27	42.80	90.96	73.76	7.54$
`Swan-Small`	81.55	78.86	70.97	42.48	80.46	70.86	**0.44$**
Openai-3-small	71.42	85.23	71.50	32.90	82.20	68.65	3.75$
Cohere-light-v3.0	70.32	86.83	67.68	22.68	90.34	67.57	2.55$
Openai-ada-002	65.34	81.83	71.76	39.62	76.79	67.07	1.66$

Table 7: Domain-specific `ArabicMTEB` results.

Model	MSA-Cult	Egy-DIA	Mor-DIA	Avg.
`Swan-Large`	**82.19**	**83.55**	**65.35**	**77.03**
Cohere-v3.0	81.86	82.90	65.23	76.66
OpenAI-3-large	81.49	78.45	64.90	74.95
Cohere-light-v3.0	80.75	64.82	56.84	67.47
Me5-large	78.65	61.34	60.66	66.88
OpenAI-3-Small	74.55	65.89	54.13	64.86
`Swan-Small`	75.56	60.35	53.56	63.16
Me5-base	74.56	56.34	53.91	61.60
Me5-small	73.81	53.56	45.45	57.61
ATM-V2	63.78	23.45	21.45	36.23
ARBERTv2	9.34	8.55	4.67	7.52
MARBERT	2.73	0.44	0.19	1.12

Table 8: Cultural `ArabicMTEB` results.

tive models.

Domain-Specific `ArabicMTEB` Results. As seen from Table 7, `Swan-Small` performs exceptionally well, with an average score of 70.86, surpassing OpenAI's text-embedding-3-small model (68.65) and Cohere-light-v3.0 (67.57). Its best performance is in the legal domain, where it scores 78.86. `Swan-Large` sets a new standard in domain-specific tasks, scoring 82.49 on average, surpassing OpenAI's text-embedding-3-large (82.20) and Cohere's multilingual model (73.76). The model excels particularly in the news domain (90.42), medical (81.64), and Wikipedia (93.10), indicating its superior generalization across varied Arabic domains. Moreover, the cost-effectiveness of our models is evident: using `Swan-Large` costs only 0.75 for 10k documents compared to 9.88 for OpenAI's model, making it a more efficient solution for large-scale deployments.

Cultural ArabicMTEB Results. Cultural `ArabicMTEB` is designed to capture culturally sensitive aspects of the Arabic language, such as regional dialects, local idiomatic expressions, and culturally specific knowledge. We generated queries from country-specific Wikipedia articles, including questions about local cuisine, traditional practices, and historical events, which challenge the models to capture more than just linguistic information. For example, `Swan-Large` achieved the highest performance on tasks related to Egyptian cultural queries, outperforming other models on retrieval tasks by 1.5%. However, we observed slightly lower performance on Moroccan dialect queries, where cultural nuances (such as regional vocabulary) presented a more significant challenge.

Synthetic Data Analysis. We systematically an-

Model	ArRTR	DOM-RTR	DIA-RTR	STS	PairCLF	CLF	RRK	CLK	BTM	Avg.
Swan-Small	15.12	8.46	7.52	37.88	62.87	56.85	62.21	39.25	1.99	32.46
+ Arabic	28.39	39.34	15.23	41.49	70.25	51.89	68.57	39.12	18.74	41.45
+ Synthetic-MSA	31.07	40.45	53.45	**55.78**	74.23	54.27	**68.88**	39.43	18.19	48.42
+ Synthetic-DOM	**32.01**	**49.02**	49.34	52.90	**75.45**	54.43	67.45	**40.56**	17.35	**48.72**
+ Synthetic-DIA	31.20	38.66	**59.43**	51.23	72.86	**57.56**	66.67	37.34	**19.90**	48.32
Swan-Large	44.46	64.52	66.23	48.63	72.34	50.43	69.39	38.28	44.20	55.39
+ Arabic	54.53	66.43	70.34	52.93	75.24	52.54	70.49	40.21	48.35	59.01
+ Synthetic-MSA	56.34	67.90	72.89	**57.89**	**76.90**	50.21	**70.92**	**41.76**	62.34	61.91
+ Synthetic-DOM	**58.42**	**76.54**	71.65	55.92	75.19	50.19	70.21	39.33	51.23	60.96
+ Synthetic-DIA	57.09	65.06	**77.03**	56.90	76.42	**54.89**	69.32	39.41	**65.56**	**62.41**

Table 9: The impact of Synthetic Data on Swan performance. **ArRTR:** Arabic retrieval, **DOM-RTR:** Domain-specific retrieval, and **DIA-RTR**: Dialectal Retrieval.

alyze the impact of synthetic data on the performance of Swan-Small and Swan-Large using different combinations of training datasets. Table 9 presents the results for the base models, models trained with additional human-generated Arabic data, and models enhanced using synthetic subsets such as MSA, domain-specific, and dialectal data. When comparing the initial Swan-Small (average score of 32.46) to its version trained with synthetic MSA data, we observe a significant increase in average performance to **48.42**, representing an improvement of more than **16 points**. Similarly, Swan-Large benefits from a **6.52-point** boost in average performance (from 55.39 to 61.91) with the inclusion of synthetic MSA data.

6 Conclusion

In this paper, we introduced Swan-Small and Swan-Large, along with the comprehensive ArabicMTEB benchmark for evaluating Arabic text embeddings. Our models demonstrate outstanding performance, benefiting from the strategic use of hard negatives and synthetic data in training. The evaluation across multiple benchmarks demonstrates that both Swan-Small and Swan-Large set new standards in Arabic-centric NLP tasks. They outperform existing SoTA models in both cross-lingual and Arabic-specific tasks while being cost-effective and capable of understanding cultural context—making them ideal for real-world applications in diverse Arabic language settings.

7 Limitations

While the development of the Swan models and the introduction of ArabicMTEB mark significant advancements in Arabic text embeddings, there are a number of limitations to consider. For example, although synthetic data significantly enhances model performance, it can introduce biases due to the reliance on specific patterns in the generated content. We ensured our synthetic data generation diversity by varying the data sources and generating dialectal data for multiple regions, including Egypt, Morocco, and the Gulf states, to mitigate this. We also analyzed our models by examining whether MSA data received higher accuracies in retrieval tasks in Table 9. Further, our synthetic data generation pipeline was subjected to human verification for correctness and balance across cultural contexts.

8 Ethical Statement

The societal implications of deploying dialect-aware models, such as Swan, require careful consideration. While these models can bridge gaps in NLP for Arabic-speaking regions, there is a risk of inadvertently reinforcing biases or language hierarchies, particularly in areas where particular dialects are stigmatized or underrepresented. For instance, users in communities with dialects associated with lower socioeconomic status may feel marginalized if their dialect is not adequately supported. To mitigate these concerns, we have prioritized the inclusion of low-resource dialects and ensured that our synthetic data generation pipeline accounts for dialectal diversity. Additionally, future versions of models should include further dialectal balancing, specifically focusing on underrepresented communities.

Importantly, all research and development activities for the Swan models and ArabicMTEB benchmark were conducted with a commitment to ethical standards. Data collection and usage adhered to privacy and confidentiality norms, ensuring no sensitive information was utilized without proper anonymization and consent.

Acknowledgments

We acknowledge support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[4] and UBC Advanced Research Computing-Sockeye.[5]

References

2023. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*.

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of arabic dialects.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020a. Arbert & marbert: Deep bidirectional transformers for arabic. ACL-2021 camera ready version.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021a. Arbert & marbert: Deep bidirectional transformers for arabic. pages 7088–7105.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021b. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Abdelrahim Elmadany, Chiyu Zhang, Houda Bouamor, Nizar Habash, et al. 2023a. Nadi 2023: The fourth nuanced arabic dialect identification shared task. In *Proceedings of ArabicNLP 2023*, pages 600–613.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023b. Nadi 2023: The fourth nuanced arabic dialect identification shared task. pages 600–613.

Muhammad Abdul-Mageed, Amr Keleg, Abdelrahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024a. Nadi 2024: The fifth nuanced arabic dialect identification shared task. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 709–728.

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024b. Nadi 2024: The fifth nuanced arabic dialect identification shared task. pages 709–728.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020b. Nadi 2020: The first nuanced arabic dialect identification shared task. pages 97–110.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021c. Nadi 2021: The second nuanced arabic dialect identification shared task. pages 244–259.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021d. NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. Nadi 2022: The third nuanced arabic dialect identification shared task. pages 85–97.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020c. Toward micro-dialect identification in diaglossic and code-switched environments. pages 5855–5876.

Murtadha Ahmed, Saghir Alfasly, Bo Wen, Jamal Addeen, Mohammed Ahmed, and Yunfeng Liu. 2024. Alclam: Arabic dialect language model. pages 153–159.

Asaad Alghamdi, Xinyu Duan, Wei Jiang, Zhenhai Wang, Yimeng Wu, Qingrong Xia, Zhefeng Wang, Yi Zheng, Mehdi Rezagholizadeh, Baoxing Huai, Peilun Cheng, and Abbas Ghaddar. 2023. AraMUS: Pushing the limits of data and model scale for Arabic natural language processing. pages 2883–2894.

Manel Aloui, Hasna Chouikhi, Ghaith Chaabane, Haithem Kchaou, and Chehir Dhaouadi. 2024a. 101 billion arabic words dataset. *Preprint*, arXiv:2405.01590.

Manel Aloui, Hasna Chouikhi, Ghaith Chaabane, Haithem Kchaou, and Chehir Dhaouadi. 2024b. 101 billion arabic words dataset. *Preprint*, arXiv:2405.01590.

Fakhraddin Alwajih, Gagan Bhatia, and Muhammad Abdul-Mageed. 2024. Dallah: A dialect-aware multimodal large language model for arabic. *Preprint*, arXiv:2407.18129.

[4] https://alliancecan.ca

[5] https://arc.ubc.ca/ubc-arc-sockeye

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Akari Asai, Jungo Kasai, Jonathan H. Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. Xor qa: Cross-lingual open-retrieval question answering. *Preprint*, arXiv:2010.11856.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27.

Gagan Bhatia. 2023. PolyDeDupe.

Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. 2024. Fintral: A family of gpt-4 level multimodal financial large language models. *Preprint*, arXiv:2402.10986.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of the ms marco passage ranking dataset.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of arabic. pages 1240–1245.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. pages 1–14.

Zhihong Chen, Shuo Yan, Juhao Liang, Feng Jiang, Xiangbo Wu, Fei Yu, Guiming Hardy Chen, Junying Chen, Hongbo Zhang, Li Jianquan, Wan Xiang, and Benyou Wang. 2023. MultilingualSIFT: Multilingual Supervised Instruction Fine-tuning.

Mathieu Ciancone, Imene Kerboua, Marion Schaeffer, and Wissam Siblini. 2024. Extending the massive text embedding benchmark to french. *arXiv preprint arXiv:2405.20468*.

Cohere For AI. 2024. c4ai-command-r-plus-08-2024.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ibrahim Abu El-Khair. 2016. 1.5 Billion Words Arabic Corpus. *arXiv preprint arXiv:1611.04033*.

AbdelRahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2022. Orca: A challenging benchmark for arabic language understanding.

AbdelRahim Elmadany, ElMoatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. ORCA: A challenging benchmark for Arabic language understanding. pages 9559–9586.

Kenneth Enevoldsen, Márton Kardos, Niklas Muennighoff, and Kristoffer Laigaard Nielbo. 2024. The scandinavian embedding benchmarks: Comprehensive assessment of multilingual and monolingual text embedding. *arXiv preprint arXiv:2406.02396*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. pages 878–891.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, et al. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *arXiv preprint arXiv:2204.08582*.

Gemma-Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, et al. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. Acegpt localizing large language models in arabic. pages 8139–8163.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. Aldi: Quantifying the arabic level of dialectness of text. pages 10597–10611.

Amr Keleg and Walid Magdy. 2023. Arabic dialect identification under scrutiny: Limitations of single-label classification. pages 385–398.

Md Tawkat Islam Khondaker, Abdelrahim Elmadany, Muhammad Abdul-Mageed, VS Laks Lakshmanan, et al. 2022. A benchmark study of contrastive learning for arabic social meaning. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 63–75.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. 2024. Gecko: Versatile text embeddings distilled from large language models.

Weizhe Lin, Rexhina Blloshmi, Bill Byrne, Adria de Gispert, and Gonzalo Iglesias. 2023. Li-rage: Late interaction retrieval augmented generation with explicit signals for open-domain table question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1557–1566, Toronto, Canada. Association for Computational Linguistics.

Antoine Louis, Vageesh Saxena, Gijs van Dijck, and Gerasimos Spanakis. 2024. Colbert-xm: A modular multi-vector representation model for zero-shot multilingual information retrieval.

Isabelle Mohr, Markus Krimmel, Saba Sturua, Mohammad Kalim Akram, Andreas Koukounas, Michael Günther, Georgios Mastrapas, Vinit Ravishankar, Joan Fontanals Martínez, Feng Wang, Qi Liu, Ziniu Yu, Jie Fu, Saahil Ognawala, Susana Guzman, Bo Wang, Maximilian Werk, Nan Wang, and Han Xiao. 2024. Multi-task contrastive learning for 8192-token bilingual text embeddings.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *Preprint*, arXiv:2402.09906.

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

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. pages 2014–2037.

Omer Nacar and Anis Koubaa. 2024. Enhancing semantic similarity understanding in arabic nlp with nested embedding learning.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. Turjuman: A public toolkit for neural arabic machine translation. pages 1–11.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Tariq Alhindi. 2020. Machine generation and detection of Arabic manipulated and fake news. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 69–84, Barcelona, Spain (Online). Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, Ahmed El-Shangiti, and Muhammad Abdul-Mageed. 2023a. Dolphin: A challenging and diverse benchmark for Arabic NLG. pages 1404–1422.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, Ahmed El-Shangiti, and Muhammad Abdul-Mageed. 2023b. Dolphin: A challenging and diverse benchmark for arabic nlg. *Preprint*, arXiv:2305.14989.

OpenAI. 2023. Chatgpt: Optimizing language models for dialogue. *OpenAI*. `https://openai.com/research/chatgpt`.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Rafał Poświata, Sławomir Dadas, and Michał Perełkiewicz. 2024. Pl-mteb: Polish massive text embedding benchmark. *arXiv preprint arXiv:2405.10138*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Haitham Seelawi, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. 2021. ALUE: Arabic language understanding evaluation. pages 173–184.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the wmt 2023 shared task on machine translation with terminologies. pages 663–671.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore. Association for Computational Linguistics.

Xiaoyu Shen, Akari Asai, Bill Byrne, and Adrià de Gispert. 2023. xpqa: Cross-lingual product question answering across 12 languages. *Preprint*, arXiv:2305.09249.

Feng Shi, Ruifeng Ren, Xiaoying Feng, and Wenjie Li. 2023. Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *Preprint*, arXiv:2409.10173.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructure. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. *Preprint*, arXiv:2401.00368.

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

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

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2023. C-pack: Packaged resources to advance general chinese embedding.

Omar F. Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. 2019. OSIAN: Open Source International Arabic News Corpus - Preparation and Integration into the CLARIN-infrastructure. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 175–182, Florence, Italy. Association for Computational Linguistics.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. tydi: A multi-lingual benchmark for dense retrieval. pages 127–137.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2022. Making a miracl: Multilingual information retrieval across a continuum of languages. *Preprint*, arXiv:2210.09984.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

A ArMistral Training

ArMistral, is an autoregressive pretrained language model based on Mistral-7B.

Pretraining data We further pretrain it on a large and diverse Arabic dataset, including all categories of Arabic, namely Classical Arabic (CA), Dialectal Arabic (DA), and MSA. This data is aggregated from various sources: AraNews$_{v2}$ (Nagoudi et al., 2020), El-Khair (El-Khair, 2016), Gigaword,[6] OSCAR (Suárez et al., 2019), OSIAN (Zeroual et al., 2019), 101 Billion arabic words (Aloui et al., 2024a), Wikipedia Arabic, and Hindawi Books.[7] We also derived ArabicWeb22 (A) and (B) from the open source Arabic text 2022.[8] This pretraining dataset was cleaned, filtered and deduplicated using Bhatia (2023). We have also ensured that the model is pretrained in multiple domains, enhancing its results as seen in Table 10.

Instruction Finetuning. To enhance the capabilities of our ArMistral, we instruct-tuning it

[6] LDC Catalog Link

[7] OpenITI corpus (v1.6).

[8] ArabicText-2022 data

on three datasets: Alpaca-GPT4, Evol-instruct, and ShareGPT extracted from MultilingualSIFT datasets (Chen et al., 2023).

Alignment Dataset. We collected an alignment dataset from Quora and Mawdoo websites and then we took the gold answers as the choosen and we generated the rejected using AceGPT-7B (Huang et al., 2024).

Results. As seen from Table 10, Our ArMistral-Chat model outperforms all existing Arabic LLMs.

B Datasets overview

The Table 11 provides a comprehensive summary of the various datasets utilized in the study. It categorizes datasets based on their type, such as Reranking, Bitext Mining, Retrieval, Crosslingual Retrieval, STS, Pair Classification, Clustering, and Classification. Each entry specifies the dataset name, language, citation, and category, reflecting the diversity and scope of data sources for evaluating the model's performance across different tasks and linguistic contexts.

C Polydedupe: versatile cleaning Pipeline

PolyDeDupe is a Python package designed for efficient and effective data deduplication across over 100 languages. It supports syntactic and semantic deduplication, making it a versatile tool for high-quality data preprocessing in NLP tasks. Key features include customizable Jaccard similarity thresholds, a performance speed twice that of other tools like SlimPajama, and support for deduplicating instruction tuning data. It can be easily installed via pip to deduplicate datasets, display original and filtered dataset sizes, and identify duplicate clusters. Supported languages span Western, Central, and Eastern European languages, Slavic languages using Cyrillic script, Greek, various Arabic and Devanagari script languages, and more.

D Prompts for evaluation

Table 12 provides an overview of the prompts used for evaluating various tasks. It includes instructions for Reranking, Bitext Mining, Retrieval, Crosslingual Retrieval, Semantic Textual Similarity (STS), Pair Classification, Clustering, and Classification. Each entry outlines the specific task and the corresponding instruction used to guide the model's evaluation process.

Figure 4: Latency vs Performance.

E Full Leaderboard

Table 13 presents the performance comparison of various models on different tasks within the `ArabicMTEB` benchmark. It includes metrics for Retrieval, Semantic Textual Similarity (STS), Pair Classification (PairCLF), Classification (CLF), Reranking, Clustering, and Bitext Mining (BTM). The table lists each model, its dimensionality, and the scores for each task, along with an overall average score. The results highlight the strengths and weaknesses of each model across a range of tasks, providing a comprehensive overview of their performance.

F Inference Latency.

Inference latency is very critical in deploying machine learning models, especially in real-time applications with crucial response time. It refers to the time taken by a model to predict received input. In the context of text embedding models such as `Swan-Small` and `Swan-Large`, lower latency is particularly valuable for user-facing services that rely on fast processing of natural language input, such as chatbots and search engines. From Figure 4, we find that `Swan-Large`, despite its larger size indicated by a larger bubble, has optimized inference times due to architectural efficiencies, and `Swan-Small` strikes the perfect balance between size, performance, and latency. We compare the performance of the models from Table 5.

G STS Dataset Creation:

The Arabic Semantic Textual Similarity (Arabic-STS) datasets was developed to facilitate research

Model	ARC	Hellaswag	Exams	MMLU	Truthfulqa	ACVA	AlGhafa	Average
ArMistral-7B-Chat	43.20	55.53	45.54	**43.50**	**52.44**	77.06	**35.57**	**50.41**
Jais-13b-chat	41.10	**57.70**	**46.74**	42.80	47.48	72.56	34.42	48.97
AceGPT-13B-chat	**43.80**	52.70	42.09	41.10	49.96	**78.42**	31.95	48.57
AceGPT-13B-base	39.90	51.30	39.48	40.50	46.73	75.29	30.37	46.22
AraLLama-7B-Chat	39.45	50.23	38.24	41.03	50.44	70.45	32.54	46.05
ArMistral-7B-Base	41.50	52.50	38.92	37.50	51.27	69.64	30.24	45.94
Jais-13b-base	39.60	50.30	39.29	36.90	50.59	68.09	30.07	44.98
AceGPT-7B-chat	38.50	49.80	37.62	34.30	49.85	71.81	31.83	44.81
AraLLama-7B-Base	38.40	50.12	38.43	40.23	45.32	69.42	31.52	44.78
AceGPT-7B-base	37.50	48.90	35.75	29.70	43.04	68.96	33.11	42.42

Table 10: Comparison of ArMistral with other Arabic LLMs.

in semantic similarity for the Arabic language. The dataset is derived from the Arabic Billion Words (Aloui et al., 2024b) corpus, which serves as a foundation for extracting a diverse collection of sentence pairs. Each pair is annotated with a similarity score that captures the degree of semantic equivalence between the sentences. The dataset generation process was guided by the capabilities of the GPT-4 model developed by OpenAI, ensuring that the resulting sentence pairs are of high quality and reflect nuanced linguistic characteristics. The creation involved several steps, including selecting representative sentences from the corpus, generating semantically varied sentence pairs, and annotating similarity scores using both automated methods and human reviewers to maintain consistency and reliability.

H Country level Cultural Evaluation

Task	Dataset	Type	Language	Citation	Size
BitextMining	Darija	S2S	Moroccan Arabic Dialect to English	(Nagoudi et al., 2023b)	2000
BitextMining	Narabizi	S2S	Arabizi to French	(Nagoudi et al., 2023b)	144
BitextMining	Mt_en2ar	S2S	English to MSA	(Nagoudi et al., 2023b)	4000
BitextMining	Mt_fr2ar	S2S	French to MSA	(Nagoudi et al., 2023b)	4000
BitextMining	Mt_es2ar	S2S	Spanish to MSA	(Nagoudi et al., 2023b)	4000
BitextMining	Mt_ru2ar	S2S	Russian to MSA	(Nagoudi et al., 2023b)	4000
BitextMining	Cs_dz_fr	S2S	Algerian Arabic Dialect to French	(Nagoudi et al., 2023b)	200
BitextMining	Cs_eg_en	S2S	Egyptian Arabic Dialect to English	(Nagoudi et al., 2023b)	200
BitextMining	Cs_jo_en	S2S	Jordanian Arabic to English	(Nagoudi et al., 2023b)	200
BitextMining	Cs_ma_fr	S2S	Moroccan Arabic to French	(Nagoudi et al., 2023b)	200
BitextMining	Cs_ps_en	S2S	Palestinian Arabic to English	(Nagoudi et al., 2023b)	200
BitextMining	Cs_ye_en	S2S	Yemeni Arabic to English	(Nagoudi et al., 2023b)	200
Classification	MassiveIntent	S2S	Multilingual (Arabic subset)	(FitzGerald et al., 2022)	100
Classification	MassiveScenario	S2S	Multilingual (Arabic subset)	(FitzGerald et al., 2022)	100
Classification	OrcaSentiment	S2S	Arabic	(Elmadany et al., 2022)	5000
Classification	OrcaDialect_region	S2S	Arabic	(Elmadany et al., 2022)	5000
Classification	OrcaDialect_binary	S2S	Arabic	(Elmadany et al., 2022)	5000
Classification	OrcaDialect_country	S2S	Arabic	(Elmadany et al., 2022)	5000
Classification	OrcaAns_claim	S2S	Arabic	(Elmadany et al., 2022)	5000
Classification	OrcaMachine_generation	S2S	Arabic	(Elmadany et al., 2022)	5000
Classification	OrcaAge	S2S	Arabic	(Elmadany et al., 2022)	5000
Classification	OrcaGender	S2S	Arabic	(Elmadany et al., 2022)	5000
Classification	OrcaAdult	S2S	Arabic	(Elmadany et al., 2022)	5000
Classification	OrcaDangerous	S2S	Arabic	(Elmadany et al., 2022)	5000
Classification	OrcaEmotion	S2S	Arabic	(Elmadany et al., 2022)	5000
Classification	OrcaHate_speech	S2S	Arabic	(Elmadany et al., 2022)	5000
Classification	OrcaOffensive	S2S	Arabic	(Elmadany et al., 2022)	5000
Classification	OrcaIrony	S2S	Arabic	(Elmadany et al., 2022)	5000
Classification	OrcaSarcasm	S2S	Arabic	(Elmadany et al., 2022)	5000
Classification	OrcaAbusive	S2S	Arabic	(Elmadany et al., 2022)	5000
Clustering	Arabic_news	P2P	Arabic	Our Paper	2500
Clustering	Arabic_topic	S2S	Arabic	Our Paper	30
Clustering	Arabic_baly_stance	P2P	Arabic	(Elmadany et al., 2022)	1000
Clustering	Arabic_baly_stance	S2S	Arabic	(Elmadany et al., 2022)	100
PairClassification	Arabic_xnli	S2S	Arabic	Our Paper	538
PairClassification	Arabic_sts	S2S	Arabic	Our Paper	1256
PairClassification	Arabic_mq2q	S2S	Arabic	Our Paper	244
Reranking	Miracl_ar	S2P	Multilingual (Arabic subset)	(Zhang et al., 2023)	750
Reranking	Mmarco_arabic	S2P	Arabic	Our Paper	3000
Reranking	MedicalQA_arabic	S2P	Arabic	Our Paper	4350
Reranking	Mmarco_en2ar	S2P	English to MSA	Our Paper	500
Reranking	Mmarco_ar2en	S2P	MSA to English	Our Paper	500
Retrieval	MultiLongDoc	S2P	Multilingual (Arabic subset)	MDQA	
Retrieval	XPQA	S2S	Multilingual (Arabic subset)	XPQA	
Retrieval	Mintaka	S2S	Multilingual (Arabic subset)	Mintaka	
Retrieval	Lareqa	S2P	Arabic	(Nagoudi et al., 2023b)	220
Retrieval	Dawqs	S2S	Arabic	(Nagoudi et al., 2023b)	318
Retrieval	Exams	S2S	Arabic	(Nagoudi et al., 2023b)	2600
Retrieval	Mkqa	S2S	Arabic	(Nagoudi et al., 2023b)	340
Retrieval	Mlqa	S2S	Arabic	(Nagoudi et al., 2023b)	517
Retrieval	Arcd	S2S	Arabic	(Nagoudi et al., 2023b)	693
Retrieval	Tydiqa	S2S	Arabic	(Nagoudi et al., 2023b)	5700
Retrieval	Xsquad	S2S	Arabic	(Nagoudi et al., 2023b)	5700
Retrieval	Crosslingual_ar2de	S2P	MSA to German	Our Paper	1831
Retrieval	Crosslingual_ar2en	S2P	MSA to English	Our Paper	1831
Retrieval	Crosslingual_ar2es	S2P	MSA to Spanish	Our Paper	1831
Retrieval	Crosslingual_ar2hi	S2P	MSA to Hindi	Our Paper	1831
Retrieval	Crosslingual_ar2vi	S2P	MSA to Vietnamese	Our Paper	1831
Retrieval	Crosslingual_ar2zh	S2P	MSA to Chinese	Our Paper	1831
Retrieval	Crosslingual_de2ar	S2P	German to MSA	Our Paper	1831
Retrieval	Crosslingual_en2ar	S2P	English to MSA	Our Paper	1831
Retrieval	Crosslingual_es2ar	S2P	Spanish to MSA	Our Paper	1831
Retrieval	Crosslingual_hi2ar	S2P	Hindi to MSA	Our Paper	1831
Retrieval	Crosslingual_vi2ar	S2P	Vietnamese to MSA	Our Paper	1831
Retrieval	Crosslingual_zh2ar	S2P	Chinese to MSA	Our Paper	1912
Retrieval	MoroccoCultural	S2P	Arabic	Our Paper	100
Retrieval	SyriaCultural	S2P	Arabic	Our Paper	100
Retrieval	LibyaCultural	S2P	Arabic	Our Paper	100
Retrieval	LebanonCultural	S2P	Arabic	Our Paper	100
Retrieval	QatarCultural	S2P	Arabic	Our Paper	100
Retrieval	SudanCultural	S2P	Arabic	Our Paper	100
Retrieval	AlgeriaCultural	S2P	Arabic	Our Paper	100
Retrieval	MauritaniaCultural	S2P	Arabic	Our Paper	100
Retrieval	TunisiaCultural	S2P	Arabic	Our Paper	100
Retrieval	IraqCultural	S2P	Arabic	Our Paper	100
Retrieval	EgyptCultural	S2P	Arabic	Our Paper	100
Retrieval	SomaliaCultural	S2P	Arabic	Our Paper	100
Retrieval	UAE_Cultural	S2P	Arabic	Our Paper	100
Retrieval	OmanCultural	S2P	Arabic	Our Paper	100
Retrieval	KuwaitCultural	S2P	Arabic	Our Paper	100
Retrieval	BahrainCultural	S2P	Arabic	Our Paper	100
Retrieval	Saudi_ArabiaCultural	S2P	Arabic	Our Paper	100
Retrieval	JordanCultural	S2P	Arabic	Our Paper	100
Retrieval	PalestineCultural	S2P	Arabic	Our Paper	100
Retrieval	YemenCultural	S2P	Arabic	Our Paper	100
Retrieval	MoroccoDIA	S2P	Moroccan Arabic Dialect	Our Paper	100
Retrieval	EgyptDIA	S2P	Egyptian Arabic Dialect	Our Paper	100
Retrieval	NewsDomainSpecific	S2P	Arabic	Our Paper	1000
Retrieval	LegalDomainSpecific	S2P	Arabic	Our Paper	1000
Retrieval	MedicalDomainSpecific	S2P	Arabic	Our Paper	1000
Retrieval	FinanceDomainSpecific	S2P	Arabic	Our Paper	1000
Retrieval	WikipediaDomainSpecific	S2P	Arabic	Our Paper	1000
STS	STS17	S2S	Arabic	(Cer et al., 2017)	8060
STS	STS22	P2P	Arabic	(Semenov et al., 2023)	500
STS	Arabic_sts	S2S	Arabic	Our Paper	750
STS	Arabic_stsb_multi_dialect	S2S	Arabic Dialectal	Our Paper	1500
STS	Arabic_sts	P2P	Arabic	Our Paper	500

Table 11: Overview of ArabicMTEB datasets. **S2S:** Sentence to Sentence. **S2P:** Sentence to Paragraph. **P2P:** Paragraph to Paragraph.

Task	Instructions
Reranking	Given an Arabic search query, retrieve web passages that answer the question in {Lang}. Query:{query}.
BitextMining	Retrieve parallel sentences in {Lang}.
Retrieval	Given an Arabic search query, retrieve web passages that answer the question. Query:{query}.
Crosslingual Retrieval	Given an Arabic search query, retrieve web passages that answer the question in {Lang}. Query:{query}.
STS	Retrieve semantically similar text. Text: {text}.
Pair Classification	Retrieve texts that are semantically similar to the given text. Text: {text}.
Clustering	Identify the topic or theme of the given news article. Article:{article}.
Classification	Classify the text into the given categories {options}.

Table 12: Prompts used for evaluation.

Model	Dim.	Retrieval	STS	PairCLF	CLF	Re-rank	Cluster	BTM	Avg
Number of datasets		23	5	3	18	5	4	12	70
Swan-Large	4096	**65.63**	59.10	**75.62**	52.55	69.42	41.24	**71.24**	**62.11**
multilingual-e5-lg	1024	64.01	**59.45**	75.06	53.43	**70.79**	42.49	66.33	61.65
e5-mistral-7b-inst	4096	56.34	57.02	70.24	53.21	66.24	39.44	70.50	59.00
Swan-Small	768	58.42	58.44	74.93	**57.34**	68.43	40.43	42.45	57.21
multiling-e5-b	768	56.91	57.99	74.30	52.30	69.07	**42.56**	33.90	55.29
multiling-e5-s	384	55.14	56.73	73.97	50.85	67.92	42.37	38.47	55.06
LaBSE	768	34.98	54.15	70.60	49.57	62.17	41.42	33.28	49.45
text2vec-base	384	27.69	59.37	71.41	47.94	57.76	37.26	38.32	48.54
ARBERTv2	768	15.12	37.88	62.87	56.85	62.21	39.25	1.99	39.45
CamelBERT-msa	768	9.21	47.69	67.43	55.77	60.20	39.89	1.85	40.29
arabertv02-large	1024	7.34	34.26	63.63	54.32	56.71	37.26	10.97	37.78
arabertv02-base	768	8.62	39.77	66.30	55.77	60.03	41.74	0.70	38.99
CamelBERT-mix	768	7.19	46.47	67.23	56.68	57.50	38.72	0.41	39.17
MARBERTv2	768	5.88	45.21	70.89	54.89	58.64	40.81	0.45	39.54
ARBERT	768	8.07	29.89	61.86	56.92	61.09	37.10	2.28	36.74
CamelBERT-da	768	4.07	41.05	65.82	53.75	54.44	37.63	0.31	36.72
MARBERT	768	2.22	40.62	66.46	54.35	53.09	36.33	0.40	36.21
CamelBERT-ca	768	2.74	36.49	62.26	46.26	51.34	35.77	0.09	33.56

Table 13: ArabicMTEB Results.

Model	Swan-lg	Me5-lg	Coh-lt-v3.0	Swan-s	OpenAI-3-lg	Coh-v3.0	Me5-s	Me5-b	ATM-V2	ARBERTv2	MARBERT
Algeria	89.34	93.34	89.44	90.45	86.95	88.99	91.23	90.66	84.99	18.27	1.50
Bahrain	93.71	93.77	93.52	86.48	91.98	92.40	93.08	89.04	90.49	27.48	5.74
Egypt	98.34	94.58	91.37	95.66	91.45	87.81	93.02	91.65	88.45	11.54	1.63
Iraq	92.45	90.90	86.98	88.34	92.43	87.83	89.02	90.78	81.22	17.34	1.92
Jordan	92.34	92.79	90.07	89.70	94.56	91.18	93.67	92.25	87.95	27.46	4.50
Kuwait	93.45	96.34	96.10	90.44	88.53	92.51	96.17	94.94	89.97	36.67	4.92
Lebanon	95.66	93.05	92.38	90.45	90.23	91.04	91.92	92.85	87.14	22.55	1.82
Libya	89.56	88.43	87.27	85.45	89.66	85.75	87.21	85.32	79.95	28.88	2.46
Mauritania	92.44	92.92	92.61	89.45	90.31	92.05	20.99	3.32	0.63	0.50	0.00
Morocco	90.34	85.49	83.19	86.34	83.56	85.47	81.73	86.59	4.75	0.32	0.00
Oman	94.45	94.26	92.37	91.98	92.45	92.61	93.00	93.04	84.21	11.24	3.43
Palestine	90.45	90.67	87.50	91.18	87.45	83.33	85.22	86.49	77.83	27.25	3.63
Qatar	98.79	93.44	91.80	92.35	95.66	89.98	91.20	90.49	85.50	29.15	7.00
Saudi_Arabia	95.34	93.49	92.98	91.47	90.45	92.12	92.72	91.47	86.48	25.06	2.50
Somalia	90.23	94.78	93.67	88.34	89.55	92.30	21.25	2.50	20.81	2.62	0.00
Sudan	92.36	91.99	86.90	90.89	91.45	90.72	89.49	87.60	82.47	24.51	2.50
Syria	91.46	91.83	90.56	90.45	90.56	86.97	88.69	88.75	87.45	13.81	3.63
Tunisia	94.57	94.64	93.46	95.54	85.34	90.92	93.79	92.04	84.40	25.04	4.15
UAE	96.09	95.14	93.41	94.12	97.66	93.53	94.45	91.56	91.79	31.92	2.00
Yemen	92.34	91.24	89.40	92.12	89.54	89.70	88.25	89.89	83.08	5.29	1.29
Avg.	93.19	92.65	90.75	90.56	90.49	89.86	83.81	81.56	73.98	19.34	2.73

Table 14: Country-level cultural evaluation.