

Second Language (Arabic) Acquisition of LLMs via Progressive Vocabulary Expansion

Jianqing Zhu^{†1}, Huang Huang^{†2}, Zhihang Lin^{†4}, Juhao Liang^{†3,4}, Zhengyang Tang^{†3,4},
Khalid Almubarak⁶, Abdulmohsen Alharthik¹, Bang An¹, Juncai He¹, Xiangbo Wu³,
Fei Yu^{3,4}, Junying Chen^{3,4}, Zhuoheng Ma⁴, Yuhao Du⁴, He Zhang⁴, Saied Alshahrani⁷,
Emad A. Alghamdi⁵, Lian Zhang², Ruoyu Sun^{3,4}, Haizhou Li^{3,4},
Benyou Wang^{*3,4}, Jinchao Xu¹

¹ King Abdullah University of Science and Technology ² Shenzhen International Center for Industrial and Applied Mathematics, Shenzhen Research Institute of Big Data

³ Shenzhen Research Institute of Big Data ⁴ The Chinese University of Hong Kong

⁵ King Abdulaziz University ⁶ Prince Sattam bin Abdulaziz University ⁷ University of Bisha

Abstract

This paper addresses the critical need for democratizing large language models (LLM) in the Arab world, a region that has seen slower progress in developing models comparable to state-of-the-art offerings like GPT-4 or GPT-3.5, due to a predominant focus on mainstream languages (e.g., English and Chinese). One practical objective for Arabic LLMs is to utilize Arabic-specific vocabulary in the tokenizer to accelerate decoding. However, using a different vocabulary often leads to degradation of the model’s learned knowledge, since many words become out-of-vocabulary (OOV) at the beginning of training. Inspired by the vocabulary learning during Second Language (Arabic) Acquisition for humans, the released AraLLaMA employs progressive vocabulary expansion, which is implemented by a modified BPE algorithm that progressively extends the Arabic subwords in its dynamic vocabulary during training, thereby balancing the OOV ratio at every stage. The ablation study demonstrated the effectiveness of Progressive Vocabulary Expansion. Moreover, AraLLaMA achieves decent performance comparable to the best Arabic LLMs across a variety of Arabic benchmarks. Our model weights are available at: <https://github.com/FreedomIntelligence/AraLLaMa>.

1 Introduction

In the evolving landscape of large language models (LLMs), the predominant focus has been on English and Chinese. This focus has left other linguistic communities, notably the Arab world, with slower progress in developing comparable models. Within the Arab world¹, the development

of models such as Jais (Sengupta et al., 2023) and AceGPT (Huang et al., 2024) marks a significant step forward, yet these models do not rival the capabilities of state-of-the-art models like GPT-4 (Achiam et al., 2023) or even GPT-3.5. In line with democratization (Touvron et al., 2023a,b), our development of Arabic LLMs focuses on language adaptation settings that utilize existing standard LLM architectures (like LLaMA (Touvron et al., 2023b)) and well-trained weights, thus saving computing resources and ensuring compatibility.

A primary challenge in adapting English-centric LLMs to other languages lies in vocabulary expansion (Touvron et al., 2023b; Cui et al., 2023; Huang et al., 2024; Zhao et al., 2024). For instance, AceGPT exhibits slower decoding speeds when processing Arabic, which may be attributed to limitations in its vocabulary adaptation mechanisms. It decodes Arabic words into sequences of alphabetical letters rather than at a more efficient granularity, such as Arabic subwords. This inefficiency significantly limits its broader applicability, despite its performance being nearly on par with GPT-3.5 in some benchmarks. A key concern related to vocabulary expansion is the risk that abrupt increases may result in a high incidence of out-of-vocabulary (OOV) tokens—units absent from the model’s established vocabulary. Such a surge in OOV words can compromise the linguistic knowledge embedded within the core models. Addressing this issue requires a considerable volume of pre-training data to restore and maintain the model’s linguistic capabilities effectively.

The core philosophy behind AraLLaMA is inspired by the process of vocabulary learning in human Second Language Acquisition, emphasizing that individuals typically expand their vocabulary gradually through incremental learning, rather than through instantaneous acquisition. AraLLaMA progressively extends the Arabic subwords in its vo-

*corresponding author: wangbenyou@cuhk.edu.cn.

† These authors contributed equally to this work.

¹The Arab World comprises a large group of countries, mainly located in Western Asia and Northern Africa.

cabulary during pre-training, effectively reducing the ratio of OOV words at every stage. AraLLaMA, based on the initialization of LLaMA2 (Touvron et al., 2023b), not only retains the foundational knowledge of LLaMA2, but also enables effective cross-lingual transfer from English to Arabic. Ablation on TinyLLaMA (Zhang et al., 2024) demonstrated the effectiveness of the proposed progressive vocabulary expansion, see Section 6.1.

Followed by extensive instruction tuning, AraLLaMA achieves decent performance comparable to the best Arabic LLMs across various Arabic benchmarks. The contributions of this work are three-fold: 1) We introduce progressive vocabulary expansion, utilizing a modified byte pair encoding (BPE) algorithm inspired by human Second Language Acquisition, and demonstrate its effectiveness. 2) We present AraLLaMA, a pioneering open-source Arabic Large Language Model that decodes Arabic texts three times faster than its predecessor (Huang et al., 2024) while delivering superior performance. 3) We provide the community with access to the complete data processing pipeline, pre-training/fine-tuning data, and model weights. AraLLaMA is compatible with the most popular LLM architecture (i.e., LLaMA) and can be seamlessly integrated into most LLM applications.

2 Motivation: Second Language Acquisition for Humans and LLMs

2.1 Cognitively-inspired Motivation: Second Language Acquisition for Humans

Definition 1. Second Language Acquisition (SLA) refers to the process by which people learn a language other than their native language (Krashen, 1981). SLA can occur through formal instruction in an educational setting or informally through social interaction and exposure to the language in natural settings.

In learning a second language (L2), learners pass through several developmental stages as they gain proficiency in L2, including the acquisition of phonetics, vocabulary, grammar, and pragmatics. Of these language skills, vocabulary acquisition is crucial for language learning. Several studies have posited that L2 learners mostly learn new words incidentally (Ramos and Dario, 2015; Nation, 2001). This suggests that an individual might gradually master a word or a set of words in an unconscious manner. This leads to a phenomenon:

Phenomenon 1. *In Second Language Acquisition, human individuals typically expand their vocabulary gradually, in a fashion of incremental learning rather than an instantaneous acquisition.*

A formal description of levels of language development is laid out in the Common European Framework of Reference for Languages (CEFR)². Table 8 (shown in Appendix B) showcases the required number of vocabulary size for different CEFR levels. The CEFR provides detailed descriptions of the skills language learners must achieve to effectively communicate. This can be taken as evidence of the progressive nature of vocabulary acquisition.

2.2 Problem Definition: Second Language Acquisition for LLMs

Language adaptation The focus on developing large-scale open-source language models for high-resource languages like English and Chinese has unintentionally marginalized low-resource languages, despite there being about 7,000 languages in use globally. The lack of data and computational resources makes it challenging to develop effective models for these languages. A common practice is to enhance existing models by adding specialized data for these underrepresented languages (Cui et al., 2023; Huang et al., 2024; Zhao et al., 2024), *a.k.a.*, language adaptation.

Vocabulary expansion in language adaptation As a preliminary study, we identified Arabic tokens from LLaMA2 vocabulary using regular expressions. It was observed that LLaMA2 vocabulary only includes the basic characters of the Arabic language, resulting in relatively slow encoding and decoding speeds compared to English. During domain adaptation, it is crucial for vocabulary expansion for the second language, since it could significantly speed up decoding speeds as the number of decoded tokens is reduced due to the adapted vocabulary. Furthermore, although augmenting the existing vocabulary with tokens from additional languages, followed by training on corresponding language corpora, appears to be a logical strategy, empirical evidence suggests that the gains from this

²The Common European Framework of Reference for Languages (CEFR) is a standard developed by the European Commission and officially published in 2001, with a revised edition in 2003. The framework serves as a guideline for language teaching and assessment across European Union countries, aiming to provide a common foundation and reference for curriculum design, syllabus development, language testing, and textbook compilation in Europe.

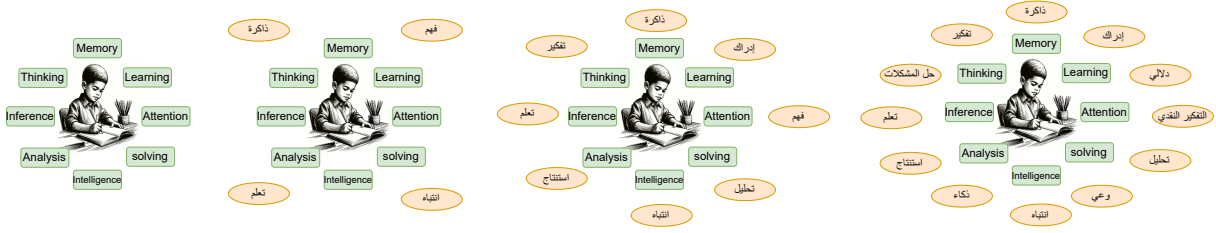


Figure 1: Second Language Acquisition for human, an English-speaking child’s journey to Arabic fluency, from basic vocabulary to cultural proficiency

method are modest. This insight underscores the complexity of enhancing support for low-resource languages within the framework of current large-scale language models.

Research question Therefore, inspired by the humans’ Second Language Acquisition, we argue for

Is it beneficial to adopt progressive vocabulary learning in language adaptation of LLMs?

3 Methodology: Progressive Vocabulary Expansion for Language Adaptation

Conventional Byte Pair Encoding (BPE) algorithms first create a complete vocabulary by iteratively merging the most frequent character pairs from a corpus, and then commence model training with this static vocabulary. This approach, while effective for monolingual models, presents challenges when adapting to new languages as it offers no mechanism for vocabulary evolution during training. To address this limitation, we propose Progressive Vocabulary Expansion.

3.1 Incremental Byte Pair Encoding Algorithm

In contrast to standard BPE algorithms (Sennrich et al., 2015) that use a static vocabulary, we introduce **Incremental Byte Pair Encoding (I-BPE)** that dynamically augments the vocabulary during training. This approach mirrors human language acquisition, where vocabulary growth occurs simultaneously with deepening language comprehension. Algorithm 1 outlines our method.

The key innovation of I-BPE is its staged approach to vocabulary expansion. At each stage, we expand the vocabulary to a predetermined size, then train the model while gradually increasing the proportion of data corresponding to newly added

Algorithm 1 Incremental Byte Pair Encoding (I-BPE) Algorithm

- 1: **Input:** (1) Initial vocabulary V ; (2) Vocabulary size at each stage: s_0, s_1, \dots, s_n ; (3) Proportion of training corpus for newly added tokens at each stage: r_0, r_1, \dots, r_n ;
- 2: **Output:** Final vocabulary V for model training and application
- 3: **for** $i = 0$ to n **do**
- 4: **while** $|V| < s_i$ **do**
- 5: Compute frequency of all adjacent token pairs in V
- 6: Identify the most frequent token pair P_{freq}
- 7: Merge P_{freq} into a new token T_{new}
- 8: Add T_{new} to vocabulary V
- 9: **end while**
- 10: Adjust corpus proportion for newly added tokens to r_i
- 11: Train model with the updated vocabulary V until convergence
- 12: **end for**
- 13: **Return** Finalized vocabulary V

tokens. This approach significantly reduces out-of-vocabulary (OOV) tokens at each training phase, enabling the model to incorporate new linguistic elements while preserving previously acquired knowledge.

3.2 Expansion Strategies Comparison

For implementing vocabulary expansion, we investigated two principled strategies (illustrated in Figure 2):

Uniform Expansion: Adds a fixed number K of tokens at each stage, resulting in $(T - 1) \times K$ total additions over T stages.

Exponential Expansion: Doubles the number of new tokens at each stage following the sequence $\{0, 1, 2, \dots, 2^{T-2}\}$, mimicking human language

acquisition patterns.

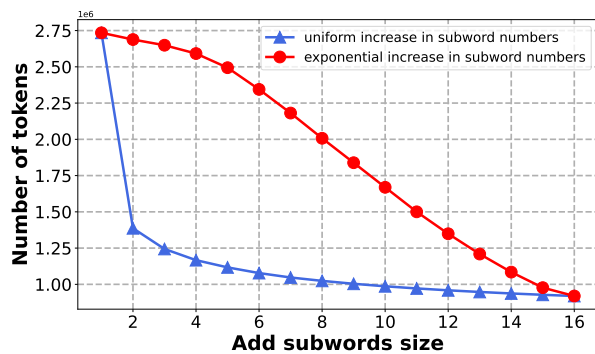


Figure 2: Compression ratio comparison between uniform and exponential vocabulary expansion strategies.

Our comparative analysis using an identical Arabic corpus through 16 progressive stages revealed crucial differences between the two approaches. As shown in Figure 2 and detailed in Table 11, uniform expansion causes abrupt improvements in compression ratio during early stages followed by diminishing returns. This sudden introduction of many tokens creates training instabilities and risks catastrophic forgetting as the model’s representation space must rapidly accommodate numerous new tokens simultaneously.

Exponential expansion, however, offers critical advantages through its graduated approach: it provides superior training stability as the gradual introduction of tokens allows smooth adaptation of the model’s representation space; it maintains consistent OOV ratios throughout training, preventing sudden vocabulary distribution shifts; and it achieves significant computational efficiency with a 3 times reduction in sequence length compared to the original LLaMA tokenizer. Based on these findings, we implemented exponential expansion with 12,800 Arabic subwords across 16 stages (\log_2^{12800}), representing the optimal saturation point for compression ratio improvement.

3.3 Compression Ratios and Tokenizer Evaluation

To rigorously assess the effectiveness of our vocabulary expansion approach, we conducted a comprehensive comparative evaluation of tokenization performance across multiple leading models. Using an identical Arabic corpus of 39 million words, we analyzed how different tokenizers processed Arabic text, with particular attention to efficiency metrics that impact both performance and computational requirements.

The results in Table 1 reveal notable differences in how these models handle Arabic text. Our model achieved a token compression ratio of 0.3174 (ratio of tokens to original text size), representing a 68% improvement over LLaMA2’s baseline, which directly enhances inference speed and reduces memory requirements. We evaluated several key metrics established in recent tokenizer evaluation literature:

- 1. Subword Fertility (Rust et al., 2021; Moosa et al., 2023):** This metric measures the average number of tokens per word. Our model achieves the most optimal fertility (1.7063), approximately 3 times more efficient than LLaMA2 (5.3844) and Mistral (5.2833), while also outperforming multilingual models like Bloomz (Muennighoff et al., 2022) (2.0668) and Jais (1.9260) that were specifically designed with Arabic support.
- 2. Word Integrity (Moosa et al., 2023):** For Arabic’s rich morphology, preserving words as single tokens is vital. Our model achieves 63.23% word integrity, far exceeding LLaMA2 (1.8%) and outperforming Arabic-optimized models like Jais (38.95%) and Bloomz (31.76%).
- 3. Total Tokens:** For the identical test corpus, our model requires only 66.55 million tokens, compared to LLaMA2’s 210.03 million, a reduction of approximately 68% that translates directly to memory savings and computational efficiency in both training and inference phases.
- 4. Rényi Efficiency (Zouhar et al., 2023):** This information-theoretic measure (higher values indicate better vocabulary utilization) shows our model (0.7491) achieves comparable efficiency to LLaMA2 (0.7731) despite its much lower token count, indicating efficient use of vocabulary space while maintaining high word integrity.

The comparative analysis indicates that our model achieves an optimal equilibrium between morphological preservation and computational efficiency. Although models such as LLaMA2 and Mistral exhibit marginally superior Rényi Efficiency coefficients, this advantage is offset by substantial deficiencies in word integrity preservation and significantly elevated token densities. When

Tokenizer	Total Words	Total Tokens	Subword Fertility	Ratio of Words Unbroken	Rényi Efficiency
LLaMA2(AceGPT)	39,006,442	210,027,671	5.3844	0.0183	0.7731
Bloomz	39,006,442	80,617,499	2.0668	0.3176	0.7709
Mistral	39,006,442	206,082,344	5.2833	0.0185	0.7928
Jais	39,006,442	75,126,494	1.9260	0.3895	0.7343
Our model	39,006,442	66,554,771	1.7063	0.6323	0.7491

Table 1: Comprehensive tokenizer evaluation using standard metrics across different models.

compared with models specifically optimized for Arabic processing, such as Jais and Bloomz, our model consistently demonstrates superior performance across the majority of evaluation metrics, validating the efficacy of the progressive vocabulary expansion methodology for non-Latin script languages.

4 Training Methodology

Based on the Progressive Vocabulary Expansion methodology described above, we develop AraLLaMA, an Arabic Large Language Model that implements our proposed I-BPE algorithm. In this section, we detail the AraLLaMA training process, including data engineering and training specifics.

4.1 Data Engineering

Pre-training Corpora Our pre-training dataset comprises both Arabic and English corpora. We employ an array of Arabic corpora encompassing multiple categories as delineated in Table 9 (shown in Appendix D). These include filtered versions of Common Crawl, WebText, and Wikipedia1 sourced from Joud and BAAI, all of which were subjected to additional cleaning processes. Moreover, we gather and purify additional corpora, namely Wikipedia2, Books, and Newspapers. The English corpus is sourced from SlimPajama (Sobolova et al., 2023) and Proof-Pile-2 (Azerbaiyev et al., 2023).

Data for Instruction Tuning After pre-training, we aim to elicit the knowledge out of AraLLaMA via instruction tuning. Inspired by GLAN (Li et al., 2024), we introduce ALAN (Arabic Instruction Tuning for Language Models). This method utilizes specific topics targeting Arabic knowledge to generate a vast amount of synthetic instruction data.

Specifically, we identified 127 critical topics within Arabic culture, science, and engineering as our focus. ALAN decomposes these topics into a structured hierarchy of fields, sub-fields, and in-

dividual disciplines. For each discipline, ALAN compiles a comprehensive list of subjects and designs a syllabus with specific knowledge points for each one. Using GPT-4-0613, ALAN has generated 11,430 subjects and 244,812 detailed knowledge points. We provide more concrete examples in Appendix H.

Armed with this extensive collection of subjects and knowledge points, we direct the LLM to create questions and answers related to these knowledge concepts. The syllabus consists of several lectures, each with 2 to 5 knowledge points. To diversify the knowledge base, we combine knowledge points from both the same and different lectures to produce diverse instructions and answers. Additionally, to vary the instruction types, the LLM generates three kinds of questions at random: multiple-choice, open-ended, and coding questions. In total, we’ve generated 733,419 instruction tuning data pieces.

We also incorporated instruction tuning data from previous AceGPT projects. These include Quora-Arabic, Alpaca-Arabic (Taori et al., 2023), Code-Alpaca-Arabic (Chaudhary, 2023), Evol-Instruct-Arabic (Xu et al., 2023), and ShareGPT data.

4.2 Training details

The refined methodology for LLaMA2 model’s vocabulary expansion incorporated 12,800 Arabic subwords derived through the I-BPE algorithm. The initialization procedure for each new token employed decomposition into constituent subwords from the original LLaMA2 vocabulary, with embedding initialization achieved via averaging the embeddings of these component tokens. This initialization strategy preserves semantic relationships between new and existing tokens, thereby enhancing training stability and facilitating vocabulary integration.

The training procedure was structured into 16 distinct stages³, each processing 30B tokens, cul-

³Although incremental token addition is theoretically feasi-

minating in a total corpus of 480B tokens. A cosine annealing schedule governed the proportion of Arabic to English content, with Arabic representation increasing systematically from 30% to 90% across stages. This progressive exposure enables gradual adaptation to Arabic linguistic structures while preserving cross-lingual transfer capabilities via continued English exposure. Mathematical and programming content was maintained at a consistent 5% throughout all stages to ensure robust inference capabilities in these domains (see Appendix E). The final training distribution comprised approximately 251.4B Arabic tokens and 204.6B English tokens.

The pre-training framework consisted of two principal epochs: an initial epoch employing vocabulary annealing for data distribution optimization, followed by a secondary epoch utilizing the fully refined vocabulary. An additional 20B tokens of training data were processed subsequent to vocabulary expansion to further enhance model performance. Each training phase implemented a discrete cosine learning rate schedule with warm-up period, producing a vocabulary-specific model at its conclusion, thereby rendering each phase functionally independent.

This stage-wise approach facilitates systematic integration of new tokens, enabling the model to adapt to evolving data representations while developing comprehensive understanding of linguistic patterns. The graduated modulation of language distribution—progressively increasing Arabic content while decreasing English representation—optimizes the model’s capacity to process Arabic while maintaining cross-lingual capabilities.

The implementation utilized LLaMA2 architecture in 7B and 13B parameter configurations, trained on 2,368 Ascend 910A processors. Training durations were 7 and 11 days for the 7B and 13B models, respectively. The computational configuration employed model parallelism of degree 2 and pipeline parallelism of degree 4. Optimization was conducted using AdamW with 4,096-token context length. Each training stage utilized a cosine learning rate scheduler initialized at $1e-5$ and decaying to $2e-6$, with a 15% warm-up interval. Gradient accumulation factor 8 yielded an effective batch size of 4,736, enabling processing of approximately 0.019B tokens per batch.

ble, a staged implementation ($N = 16$) was chosen to simplify data preparation.

5 Experiments

5.1 Experimental settings

Benchmarking Datasets To assess world knowledge, we employ four widely used benchmarks. *MMLU* (Measuring Massive Multitask Language Understanding) (Hendrycks et al., 2021a) evaluates knowledge acquired during pretraining across a broad range of subjects; we utilize both the original English version and the Arabic version introduced in (Huang et al., 2024) to ensure multilingual coverage. *RACE* (Reading Comprehension from Examinations) serves as a large-scale English reading comprehension benchmark that focuses on educational knowledge. *EXAMS* (Multi-subject High School Examinations Dataset for Cross-lingual and Multilingual Question Answering) further expands coverage by including subject-diverse exam questions drawn from multiple languages and curricula. *ArabicMMLU* complements these by providing an Arabic-specific variant of MMLU, tailored to reflect regional knowledge across various Arab countries and subjects. Beyond general knowledge evaluation, we also examine cultural and value alignment using *ACVA-all* and *ACVA-clean*, which focus on Arabic cultural relevance and localization. To comprehensively evaluate inference and reasoning abilities, we translate two commonsense reasoning benchmarks of varying difficulty—*BoolQ* and *ARC-Challenge* (*ARC-C*)—into Arabic, allowing for consistent cross-lingual assessment.

To ensure a fair comparison of candidate models, we adhere to the settings established for each benchmark separately. Furthermore, for translated benchmarks, we utilize the generation approach evaluation method as outlined in (Huang et al., 2024). Specifically, we employed GPT-3.5-Turbo-1106 to translate datasets from English to Arabic for benchmarks that were not originally in Arabic.

Baselines To compare LLMs trained or available in Arabic, we have selected several prominent Arabic LLMs or multilingual LLMs as baselines for comparison: (1) **AceGPT-[7B,13B]**: This set includes fully fine-tuned generative text models based on LLaMA2, specifically customized for the Arabic domain. (2) **Mistral-7B-Instruct-v0.2** (Jiang et al., 2023): The fine-tuned model achieves a balance between performance and efficiency. (3) **Jais-[13B,30B]** (Sengupta et al., 2023): A pre-trained bilingual large language model de-

signed for both Arabic and English. **(4) Bloom-[7B]**: A multilingual language model extensively trained on diverse textual data, allowing it to produce fluent text in 46 languages and 13 programming languages. **(5) LLaMA2-[7B,13B]**: A popular and competitive baseline model in the general domain. **(6) OpenAI GPT**: This includes GPT4 and ChatGPT, closed-source LLMs also strong at multilingual tasks.

5.2 Evaluation Results

Evaluation on Base Models In our study, the performance of base models was assessed on two Arabic-specific MMLU datasets: Arabic MMLU translate (Huang et al., 2024) and ArabicMMLU (Koto et al., 2024). The left side of Table 2 details the models’ accuracies on the Arabic MMLU translate dataset within a few-shot setting. It is evident from the data that the AraLLaMA-7B-base and AraLLaMA-13B-base models exhibit superior accuracy rates compared to models of similar scale. Notably, the AraLLaMA-13B-base model outperforms the Jais-30B model, which has a significantly larger parameter count.

Additionally, the right side of Table 2 presents the accuracy results of models in a zero-shot learning scenario. Here again, the AraLLaMA models stand out for their exceptional performance, even when compared to models with similar parameter sizes. In particular, the AraLLaMA-13B-base model demonstrates a marked advantage over the Jais-30B-base model, notwithstanding the latter’s larger size in terms of parameters.

These findings affirm the effectiveness of the AraLLaMA models, developed through an annealing algorithm to expand the vocabulary, highlighting our methodology as a productive strategy for enhancing large models’ adaptability to less prevalent languages. This contribution significantly advances the field of language model adaptation, offering a novel avenue for enriching language technology’s inclusivity and depth.

Evaluation on Chat Models Table 3 presents the comprehensive evaluation results across various benchmarks for the candidate models, spanning from Arabic to English. Overall, AraLLaMA outperforms all baseline models in the Arabic tasks. Particularly noteworthy is its proficiency in knowledge-related evaluations such as Arabic-translated MMLU and EXAMS, surpassing other models by at least 1.3%. This high-

lights the model’s expertise in addressing Arabic knowledge-related questions. Additionally, AraLLaMA demonstrates strong performance in tasks related to Arabic culture and value alignment. In terms of commonsense reasoning, AraLLaMA exhibits notable skills in tasks such as the translated versions of BoolQ and ARC-Challenge, showcasing its reasoning capabilities in Arabic. Beyond Arabic benchmarks, we also investigated the English proficiency of the models to determine whether specialization in one language affects performance in the other. The results indicate that the model maintains its English proficiency and displays robustness in multilingual assessments. It is noteworthy that the lower accuracy of the Jais is attributed to its refusal to answer for unknown reasons.

In a comprehensive evaluation of the ACVA dataset aimed at gauging the understanding of Arabic cultural nuances under a zero-shot setting, our AraLLaMA models showcased unparalleled performance. The AraLLaMA-13B-chat, in particular, stood out with exceptional Average F1 scores of 76.37% and 76.90% in “all set” and “clean Set” categories, respectively, even outperforming the renowned GPT-3.5 Turbo in the “All set” category. This performance not only highlights the AraLLaMA models’ superior grasp of Arabic culture but also establishes them as leading figures among open-source models in this nuanced domain. Compared to other top-tier open-source contenders, including the Jais-30B-chat variants, the AraLLaMA-13B-chat model’s superior results. The instruction-following tests can be found in Appendix I.

6 More Analysis

6.1 Ablation Study on Progressive Vocabulary Expansion

To further demonstrate the effectiveness of progressive vocabulary expansion in downstream task adaptation, we conduct continuous pre-training on a 1B-parameter TinyLLaMA model (Zhang et al., 2024), followed by supervised fine-tuning. More details on the experimental setup can be found in Appendix J.

A comprehensive analysis is conducted by applying the same Supervised Fine-Tuning (SFT) protocol across three pre-training configurations: the baseline TinyLLaMA model, TinyLLaMA with Progressive Vocabulary Expansion (PVE), and TinyLLaMA with Vocabulary Expansion all at once

Models	Arabic-trans MMLU (Huang et al., 2024)					ArabicMMLU (Koto et al., 2024)					Total	
	STEM	Humanities	Social Sciences	Others	Avg.	STEM	Social Sciences	Humanities	Arabic Language	Other	Avg.	Avg.
Bloomz-7B-base	33.35	29.29	37.58	34.53	33.69	-	-	-	-	-	-	-
LLaMA2-7B-base	30.30	29.33	27.46	30.78	29.47	33.7	32.8	33.5	28.4	36.7	33.4	31.43
AceGPT-7B-base	29.73	30.95	33.45	34.42	32.14	35.4	35.9	36.2	31.1	41.7	36.3	34.22
AraLLaMA-7B-base	33.03	32.08	35.39	35.59	34.03	36.7	36.5	34.1	30.0	41.2	37.0	35.52
LLaMA2-13B-base	32.94	32.30	33.42	37.27	33.76	32.9	35.0	37.8	35.8	39.3	36.1	34.93
Jais-13B-base	30.51	31.25	33.74	33.43	33.76	30.3	31.4	33.6	28.1	36.3	32.2	32.98
AceGPT-13B-base	36.60	38.74	43.76	42.72	40.45	42.7	45.5	48.3	42.4	50.7	46.1	43.28
AraLLaMA-13B-base	36.13	40.07	45.43	42.17	40.95	42.4	45.7	48.4	46.3	52.5	47.6	44.28
Jais-30B-v1-base	32.67	30.67	42.13	39.60	36.27	39.5	45.6	50.5	34.6	49.1	44.8	40.54
GPT-3.5 Turbo	43.38	44.12	55.57	53.21	49.07	53.8	57.0	57.5	57.6	63.8	57.7	53.39

Table 2: Evaluation of base models. We adopt a few-shot setting on Arabic-translated MMLU (Huang et al., 2024) and a zero-shot setting with option logit probability in ArabicMMLU (Koto et al., 2024). Numbers with the best performance are in **bold** in 7B and 13B groups.

Models	MMLU			Arabic				Avg.	English			Total
	(trans)	(Koto et al., 2024)	EXAMS	ACVA clean	ACVA all	BoolQ (trans)	ARC-C (trans)		BoolQ	RACE	Avg.	
LLaMA2-7B-chat	13.78	33.40	13.05	20.99	21.80	34.92	23.72	21.09	71.31	50.49	60.90	31.49
Phoenix-7b	29.72	44.74	31.93	43.80	41.86	66.70	33.53	41.75	62.23	60.97	61.60	46.16
AceGPT-7B-chat	30.69	36.31	33.73	53.87	53.07	60.70	38.05	43.77	54.74	53.97	54.36	46.12
Mistral-7B-Instruct-v0.2	27.93	41.44	21.56	64.56	63.47	60.18	35.67	44.97	84.53	73.17	78.85	52.50
AraLLaMA-7B-chat	45.77	56.62	43.69	69.46	70.86	72.45	60.49	59.90	75.78	72.13	73.96	63.02
Jais-13B-chat	19.52	54.83	19.71	66.75	61.41	41.25	11.95	39.34	28.13	20.08	24.10	35.96
LLaMA2-13B-chat	8.92	36.12	16.11	35.12	35.71	54.13	27.47	30.51	62.87	48.28	55.58	36.08
AceGPT-13B-chat	35.59	52.61	38.72	70.82	70.21	66.85	44.20	54.14	60.55	45.22	52.88	53.86
AraLLaMA-13B-chat	47.33	61.70	48.37	76.90	76.37	63.33	63.99	63.42	83.67	80.82	82.24	67.61
Jais-30B-chat-v1	38.12	59.33	40.45	74.46	72.41	73.76	50.94	58.49	65.05	75.26	70.16	61.09
Jais-30B-chat-v3	35.68	62.36	32.24	73.63	73.66	76.30	51.02	57.84	79.54	85.23	82.43	63.29
GPT-3.5 Turbo	46.07	57.72	45.63	74.45	76.88	76.12	60.24	62.44	85.32	84.65	84.99	67.45

Table 3: Chat Models Evaluation in zero-shot setting. Numbers with best performance are in **bold** in 7B and 13B groups.

(VE). The performance of these models is evaluated on the Arabic MMLU (see Table 4) and Arabic Vicuna-80 (see Table 5) benchmarks. Experiment results demonstrate that vocabulary expansion significantly enhances model performance, with the PVE approach yielding superior results across various categories in the Arabic MMLU benchmark, achieving an average score of 40.7 compared to 38.5 for VE and 36.5 for the baseline model. Similarly, in the Arabic Vicuna-80 comparison, the PVE method led to the highest accuracy of 29.18%, outperforming VE (22.61%) and the baseline model (21.3%). These results underscore the effectiveness of progressive vocabulary expansion in enhancing language model performance, particularly in complex language tasks.

6.2 Benchmarking in English dataset

We evaluated the accuracy of both base and chat models on the English MMLU dataset. As illustrated in Table 2 (shown in Appendix G), in the base model category, AraLLaMA’s accuracy is slightly lower than that of the original LLaMA model but

notably higher than the AceGPT model, which is also trained on the LLaMA architecture. This indicates that expanding Arabic capabilities via an annealing algorithm does not compromise the model’s inherent English proficiency. This offers a viable solution for language transfer in large models. After undergoing SFT, AraLLaMA achieves the highest accuracy among models of similar size and surpasses the Jais-30B model, which has a greater number of parameters.

6.3 Decoding Efficiency Analysis

We conducted a systematic evaluation of generation efficiency between LLaMA2 and AraLLaMA 7B chat models on Arabic text generation tasks. Each model was tested on standardized Arabic prompts with a maximum output length of 100 tokens. To ensure statistical reliability, we performed five independent trials and analyzed only Arabic language outputs, excluding any non-Arabic tokens from the performance calculations.

Table 6 shows that while both models achieve similar token processing speeds (30 tokens/second,

Model	STEM	Social Sciences	Humanities	Arabic Language	Other	Avg.
TinyLLaMA chat	35.1	36.9	38.5	28.6	39.8	36.5
TinyLLaMA (VE) chat	35.3	39.7	40.1	33.8	41.6	38.5
TinyLLaMA (PVE) chat	36.3	40.7	44.2	33.5	45.7	40.7

Table 4: Performance comparison on ArabicMMLU (Koto et al., 2024) across different domains.

Model	Accuracy (%)
TinyLLaMA chat	21.30 (<i>baseline</i>)
TinyLLaMA (VE) chat	22.61 (+1.31)
TinyLLaMA (PVE) chat	29.18 (+7.88)

Table 5: Performance Comparison on Arabic Vicuna-80 Benchmark

Model	Tokens/Second	Words/Second
LLaMA2	29.68 ± 0.04	4.55 ± 0.50
AraLLaMA	30.12 ± 0.06	20.37 ± 0.04

Table 6: Comparative analysis of generation speed between LLaMA2 and AraLLaMA on Arabic text.

$p > 0.05$), AraLLaMA generates words 4.5× faster. This efficiency gain (from 4.55 ± 0.50 to 20.37 ± 0.04 words/second) demonstrates the effectiveness of our vocabulary expansion approach. The improved word-level performance while maintaining similar token-level speeds indicates that our language-specific tokenization strategy successfully optimizes text generation for Arabic’s morphological complexity.

7 Conclusion

Adapting large-scale models to less commonly spoken languages is fraught with challenges, notably the hurdles of knowledge transfer and the prevalence of OOV terms. We developed a novel annealing training algorithm to address these issues specifically for Arabic. This strategy methodically expands the vocabulary and employs a phased training process, leading to the development of the AraLLaMA 7B and 13B models. Subsequent evaluations of both the base and chat configurations across diverse datasets have unequivocally established AraLLaMA’s superior accuracy compared to peers within the same parameter range. Remarkably, the AraLLaMA also exhibits robust performance advantages over models with significantly more parameters. The proven efficacy of our algorithm is supported by robust empirical evidence. Moving forward, we aim to further democratize ac-

cess to advanced model technology by making our models, along with their code and datasets, openly available, thus making a meaningful contribution to the progress of the field.

Limitation

This paper exhibits several limitations. Due to constraints in resources and budget, the models has not undergone evaluation by native Arabic speakers, which could affect its practicality and adoption. Consequently, its use is currently confined to academic research rather than online deployment. Additionally, the writing of this paper was supported by AI tools for grammar correction and refinement.

Acknowledgements

This work was conducted under the platform of the KAUST-SRIBD Joint Lab on Scientific Computing and Machine Learning. We would like to acknowledge the support of Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone Project (No. HZQSW-S-KCCYB-2024016), the Shenzhen Science and Technology Program (JCYJ20220818103001002), Shenzhen Doctoral Startup Funding (RCBS20221008093330065), Tianyuan Fund for Mathematics of National Natural Science Foundation of China (NSFC) (12326608), Shenzhen Key Laboratory of Cross-Modal Cognitive Computing (grant number ZDSYS20230626091302006), Shenzhen Stability Science Program 2023, Shenzhen Key Lab of Multi-Modal Cognitive Computing, and KAUST Baseline Research Fund.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2023. **Llemma: An open language model for mathematics**. *Preprint*, arXiv:2310.10631.
- Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. *arXiv preprint arXiv:2004.03720*.
- Sahil Chaudhary. 2023. Code alpaca: An instruction-following llama model for code generation. <https://github.com/sahil280114/codealpaca>.
- Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, et al. 2023. Phoenix: Democratizing chatgpt across languages. *arXiv preprint arXiv:2304.10453*.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. **Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality**.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.
- James Coady. 1996. L2 vocabulary acquisition through extensive reading. In *Second language vocabulary acquisition: A rationale for pedagogy*, pages 225–237. Cambridge University Press.
- Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. **Free dolly: Introducing the world’s first truly open instruction-tuned llm**.
- Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.
- Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. **EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5427–5444. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. **Measuring massive multitask language understanding**. *Preprint*, arXiv:2009.03300.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. **Measuring massive multitask language understanding**. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, et al. 2024. Acegpt, localizing large language models in arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2023. **Mistral 7b**. *Preprint*, arXiv:2310.06825.
- Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, et al. 2024. Arabicmmlu: Assessing massive multitask language understanding in arabic. *arXiv preprint arXiv:2402.12840*.
- Stephen Krashen. 1981. Second language acquisition. *Second Language Learning*, 3(7):19–39.
- Stephen Krashen. 1982. *Principles and practice in second language acquisition*. Pergamon Press.
- Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. **RACE: Large-scale ReAding comprehension dataset from examinations**. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.
- H ctor Daniel Le n Romero et al. 2016. Spaced retrieval practice applied to vocabulary learning in secondary education.

- Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, et al. 2024. Synthetic data (almost) from scratch: Generalized instruction tuning for language models. *arXiv preprint arXiv:2402.13064*.
- Ibraheem Muhammad Moosa, Mahmud Elahi Akhter, and Ashfia Binte Habib. 2023. Does transliteration help multilingual language modeling? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 670–685, Dubrovnik, Croatia. Association for Computational Linguistics.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.
- Tatsuya Nakata. 2015. Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition*, 37(4):677–711.
- I. S. P. Nation. 2001. *Learning Vocabulary in Another Language*. Cambridge Applied Linguistics. Cambridge University Press.
- Ian SP Nation and I. S. P. Nation. 2001. *Learning vocabulary in another language*, volume 10. Cambridge University Press.
- Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, et al. 2023. Seallms—large language models for southeast asia. *arXiv preprint arXiv:2312.00738*.
- Restrepo Ramos and Falcon Dario. 2015. Incidental vocabulary learning in second language acquisition: A literature review. *Profile Issues in Teachers Professional Development*, 17(1):157–166.
- Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135. Association for Computational Linguistics.
- Elizabeth Salesky, Andrew Runge, Alex Coda, Jan Niehues, and Graham Neubig. 2020. Optimizing segmentation granularity for neural machine translation. *Machine Translation*, 34(1):41–59.
- Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.
- Daria Soboleva, Al-Khateeb Faisal, Myers Robert Steeves Jacob R, Hestness Joel, and Dey Nolan. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.
- Jingjing Xu, Hao Zhou, Chun Gan, Zaixiang Zheng, and Lei Li. 2020. Vocabulary learning via optimal transport for neural machine translation. *arXiv preprint arXiv:2012.15671*.
- Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.
- Songshan Zhang, Hai Xu, and Xian Zhang. 2021. The effects of dictionary use on second language vocabulary acquisition: A meta-analysis. *International Journal of Lexicography*, 34(1):1–38.
- Jun Zhao, Zhihao Zhang, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *arXiv preprint arXiv:2401.01055*.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. 2023. Tokenization and the noiseless channel. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5184–5207, Toronto, Canada. Association for Computational Linguistics.

A Related Work

Our work primarily focuses on two key areas: low-resource language models and vocabulary expansion.

Low-resource language models Recent efforts have centered on developing open-source LLMs as alternatives to proprietary models like GPT-3.5 Turbo and GPT-4 (Taori et al., 2023; Chiang et al., 2023; Conover et al., 2023; Chen et al., 2023; Sen Gupta et al., 2023). These initiatives have expanded beyond English, addressing languages with fewer available resources and creating models specifically tailored to diverse linguistic landscapes (Chen et al., 2023; Üstün et al., 2024). SeaLLMs (Nguyen et al., 2023) are adapted from English-centric models by extending vocabulary and fine-tuning to better capture regional language complexities. Jais (Sen Gupta et al., 2023) introduces a model trained from scratch based on GPT architecture, while AceGPT Huang et al., 2024 offers a model designed to adapt to local Arabic culture, specifically tailored to regional nuances. This trend highlights the growing need for multilingual LLMs that perform well in low-resource environments while maintaining competitive performance against more established models.

Vocabulary expansion Vocabulary expansion for large language models (LLMs) has become a crucial area of research, particularly for improving performance in low-resource languages. Traditional methods like Byte Pair Encoding (BPE), while effective at handling out-of-vocabulary (OOV) words, are suboptimal for pretraining larger models, as discussed by Tay et al. (Bostrom and Durrett, 2020), who propose alternative tokenization methods to better capture linguistic nuances. Pham et al. (Xu et al., 2020) advance this by introducing optimal transport-based vocabulary learning, which optimizes the distribution of subword units, enhancing translation tasks, particularly in multilingual and low-resource settings.

Kudo et al. (Kudo, 2018) propose subword regularization and offer another avenue for improvement by allowing models to learn from multiple subword segmentation rather than a fixed one, increasing robustness and flexibility. In contexts with limited data, Liu et al. (Salesky et al., 2020) have demonstrated that combining subword-based methods with additional pretraining steps significantly improves model performance. These works show

that moving beyond traditional vocabulary methods allows for more dynamic and context-aware modeling, enhancing LLMs’ scalability and adaptability across diverse linguistic landscapes.

B CEFR Language Proficiency Levels

Table 8 illustrates the vocabulary size that learners are expected to acquire at various stages of second language acquisition. The vocabulary size is gradually expanding when humans acquire a second language, as one cannot achieve proficiency in all second-language words at once, as it takes time to digest these words.

C Cognitive Mechanisms of Vocabulary Acquisition

We reviewed relevant literature to confirm the phenomenon of exponential vocabulary expansion in second language acquisition and the cognitive theories that support it. Studies indicate that learners typically begin by mastering a small set of high-frequency vocabulary in the early stages of language learning. As they progress, their vocabulary size grows rapidly. This process can be explained through the following two aspects:

Cognitive Mechanisms of Incremental Learning

In the initial stages, learners build their understanding by repeatedly encountering and using simple foundational words. Research by (Krashen, 1982) and (Nation and Nation, 2001) shows that mastering high-frequency vocabulary is crucial for understanding more complex linguistic structures. These foundational words provide a stable cognitive base, allowing learners to gradually expand their vocabulary (Zhang et al., 2021; Nakata, 2015).

Exponential Vocabulary Growth Once learners acquire foundational vocabulary, the rate of vocabulary expansion accelerates. Through extensive reading and structured learning strategies such as spaced retrieval practice (León Romero et al., 2016), learners are able to acquire complex vocabulary in a relatively short period. (Coady, 1996) emphasize that extensive reading provides a large amount of language input, enabling learners to incrementally encounter and absorb more advanced vocabulary.

D Arabic data distribution

Table 9 show the Arabic dataset primarily draws from several key sources, with the largest contri-

bution coming from the Common Crawl (filtered) dataset, which accounts for 55.5% of the total data. Other significant sources include WebText, which contributes 26.7%, and Books+Newspapers, providing 8.9% with 2.5 billion tokens. Additionally, Wikipedia is divided into two parts, contributing 3.76% and 5.14%. These diverse sources collectively form the foundation for training the Arabic model.

E Data mixture

Table 10 shows the data distribution across the pre-training stages is carefully adjusted, with the proportions of Arabic and English data determined using a cosine annealing schedule. Initially, the Arabic data constitutes 30% of the total, while English data makes up 65% and math & coding data consistently accounts for 5%. As the training progresses and new subwords are added, the proportion of Arabic data increases steadily, reaching 90% by the final stage. Concurrently, the English data proportion decreases to 5%, while the math & coding data remains constant at 5% throughout all stages. This dynamic adjustment ensures that the model effectively balances the learning of Arabic and English content, with a strong emphasis on Arabic in the later stages.

F Comparison of compression ratio and OOV changes at different stages between exponential and uniform expansion

Table 11 illustrates the trends in compression ratio and OOV (Out-Of-Vocabulary) ratio as vocabulary size is incrementally expanded using both Exponential and Uniform methods. In the case of ****Exponential Vocabulary Expansion****, both the compression ratio and OOV ratio change gradually, ensuring a more balanced progression as new subwords are added. This gradual change is beneficial for maintaining stability during model training, as it allows the system to adjust incrementally to the growing vocabulary.

G Evaluation of models in English MMLU dataset

In the evaluation of English MMLU performance, AraLLaMA models, both 7B and 13B, consistently outperform their counterparts across most categories in both few-shot and zero-shot settings (shown in Table 2). Particularly, AraLLaMA-13B

Aspect	Benchmark	Language (+ translation)	Size	Evaluation Types	Metrics
Knowledge Ability	RACE (Lai et al., 2017)	EN	4.9K	Multiple-choice Questions	Accuracy
	MMLU (Hendrycks et al., 2021b)	EN (+AR)	14K	Multiple-choice Questions	Accuracy
	ArabicMMLU (Koto et al., 2024)	AR	14.5K	Multiple-choice Questions	Accuracy
	EXAMS (Hardalov et al., 2020)	AR	0.56K	Multiple-choice Questions	Accuracy
Arabic Cultural and Value Alignment	ACVA-all (Huang et al., 2024)	AR	9K	Yes/No binary Questions	F1-score
	ACVA-clean	AR	2.48K	Yes/No binary Questions	F1-score
Commonsense Reasoning	BoolQ (Clark et al., 2019)	EN (+AR)	3.27K	Yes/No binary Questions	Accuracy
	ARC-Challenge (Clark et al., 2018)	(+AR)	1.17K	Multiple-choice Questions	Accuracy

Table 7: Overview of Evaluation benchmarks

CEFR Level	Description	Learning Hours	Vocabulary Size
Basic User	A1 Beginner Level	110-130	2000 words
	A2 Elementary Level	150-180	3000 words
Independent User	B1 Intermediate Level	200-230	5000 words
	B2 Upper Intermediate Level	200-230	8000 words
Proficient User	C1 Advanced Level	150-200	10000 words
	C2 Mastery Level	250-300	30000 words

Table 8: CEFR Language Proficiency Levels.

achieves the highest average score of 62.89 in zero-shot tasks, demonstrating its superior capability in generalization and task adaptability.

H ALAN examples

We provide concrete examples of ALAN below. Note that we translate examples into English using GPT-3.5-Turbo. In practice, our data is in Arabic.

H.1 Topics

A set of 30 topics, randomly chosen, is listed below:

"Arabic Language and Literature" "Mathematics"
 "Islamic Studies" "Middle Eastern History and Politics"
 "Computer science" "Economics" "Healthcare industry"
 "Social work" "Business" "Geography"
 "Mining" "Chemical Engineering" "Languages and Literature"
 "Materials Science and Engineering"
 "Transport industry" "Chemistry" "Food industry"
 "Systems science" "Astronomy" "Cultural industry"
 "Energy industry" "Radiology" "Pediatrics"
 "Dentistry" "Civil Engineering" "Aerospace industry"
 "Public administration" "Infectious disease" "Public policy"
 "Environmental studies and forestry"

H.2 Subjects

A set of 30 subjects, randomly chosen, is listed below:

"Hypersonic and High-Speed Flows" "Mental Health Nursing"
 "Mechanical Systems and Energy Efficiency" "Obstetrics and Gynecological Nursing"
 "Immunology" "Interdisciplinary Geriatric Care"
 "Signal Processing" "Geography research methods and techniques"
 "Public Administration and Management" "An introduction to space exploration"
 "Environmental and Safety Management" "Social and Ethical Aspects of Agriculture"
 "Folk and Cultural Dance" "Power System Protection and Control"
 "Collage and Mixed Media" "Advanced Game Theory"
 "Pediatric Critical Care" "Transport Modeling and Forecasting"
 "Foundations of Mathematics"
 "Carbon Capture, Storage, and Utilization" "Customer Service and Relationship Management"
 "Introduction to Probability" "Virtual Reality and Augmented Reality"
 "Reservoir Management and Enhanced Oil Recovery"
 "Safety and Standards in Industrial Robotics" "Social Work with LGBTQ+ populations"
 "Nutritional Science" "Advanced Gynaecology Courses"
 "Bioinformatics and Computational Chemistry"
 "Reusable Launch Vehicle Technology"

H.3 A syllabus with specific knowledge points

We provide an example syllabus with specific knowledge points as below.

Subject title: Hypersonic and High-Speed Flows
 Lecture title: Introduction to Hypersonic Flows
 Knowledge points:
 - Definition of hypersonic flows

Dataset	# tokens	Weight in training mix
Common Crawl (filtered)	101.3 billion	55.5%
WebText	10.62 billion	26.7%
Books+Newspapers	2.5 billion	8.9%
Wikipedia1	0.36 billion	3.76%
Wikipedia2	0.51 billion	5.14%

Table 9: Arabic data distribution and elapsed epochs

Stage	New subwords added	Arabic data	English data	math & coding data
1	0	30.00%	65.00%	5.00%
2	1	30.33%	64.47%	5.00%
3	2	31.31%	63.69%	5.00%
4	4	32.94%	62.06%	5.00%
5	8	35.19%	59.81%	5.00%
6	16	38.04%	56.96%	5.00%
7	32	41.46%	53.54%	5.00%
8	64	45.41%	49.59%	5.00%
9	128	49.85%	45.15%	5.00%
10	256	54.73%	40.27%	5.00%
11	512	60.00%	35.00%	5.00%
12	1024	65.60%	29.40%	5.00%
13	2048	71.46%	23.54%	5.00%
14	4196	77.53%	17.47%	5.00%
15	8192	83.73%	11.27%	5.00%
16	12800	90.00%	5.00%	5.00%

Table 10: Detailed distribution of Arabic, English and math & coding data across each pre-training stage.

- Mach number	- Vehicle configurations
- Key characteristics of hypersonic flows	- Advantages and limitations of each configuration
Lecture title: Fundamentals of Shock Waves	Lecture title: Aerothermodynamics of Hypersonic Flows
Knowledge points:	Knowledge points:
- Definition of shock waves	- Definition of aerothermodynamics
- Formation of shock waves	- Aerothermodynamics in hypersonic flows
- Types of shock waves	- Heat transfer in hypersonic flows
Lecture title: High-Temperature Gas Dynamics	Lecture title: Hypersonic Flow Control
Knowledge points:	Knowledge points:
- Definition of high-temperature gas dynamics	- Importance of flow control
- Behavior of high-temperature gases	- Methods of hypersonic flow control
- Effects of high-temperature gases on materials	- Challenges in hypersonic flow control
Lecture title: Principles of Rarefied Gas Dynamics	Lecture title: Hypersonic Propulsion Systems
Knowledge points:	Knowledge points:
- Definition of rarefied gas dynamics	- Types of hypersonic propulsion systems
- The continuum hypothesis	- Working principles
- Governing equations	- Advantages and disadvantages
Lecture title: High-Speed Flow Over Bodies	Lecture title: Future Trends in Hypersonic and High-Speed Flows
Knowledge points:	Knowledge points:
- High-speed flow characteristics	- Current research in the field
- Impact on the body	- Potential future trends
- Aerodynamic heating	- Challenges and opportunities
Lecture title: Hypersonic Vehicle Configurations	
Knowledge points:	
- Types of hypersonic vehicles	

Add Subword Size	Compress Ratio (Exponential)	OOV Ratio (Exponential)	Add Subword Size	Compress Ratio (Uniform)	OOV Ratio (Uniform)
0	0.90	0.000	0	0.90	0.000
1	0.88	0.017	853	0.45	0.669
2	0.87	0.018	1736	0.40	0.116
4	0.85	0.022	2559	0.37	0.068
8	0.82	0.038	3412	0.35	0.049
16	0.77	0.061	4265	0.34	0.039
32	0.72	0.076	5118	0.33	0.031
64	0.65	0.094	5971	0.32	0.026
128	0.60	0.093	6824	0.31	0.021
256	0.54	0.105	7677	0.31	0.019
512	0.48	0.116	8530	0.30	0.017
1024	0.43	0.110	9383	0.30	0.015
2048	0.39	0.118	10236	0.30	0.013
4096	0.34	0.120	11089	0.29	0.012
8192	0.31	0.116	11942	0.29	0.011
12800	0.28	0.070	12800	0.28	0.010

Table 11: Comparison of Exponential and Uniform Vocabulary Expansion Methods

H.4 Synthetic QA data

We provide a synthetic QA example using knowledge points generated by GPT-3.5-Turbo.

Subject title:
Computer Vision for Industrial Robotics

Lecture title:
Stereo Vision and 3D Reconstruction

Knowledge points:
- Principles of stereo vision
- Stereo camera calibration
- Depth estimation and 3D reconstruction
- Point cloud processing

Synthetic question:
In stereo vision, the process of determining the depth of objects in a scene is known as:
A. Image rectification
B. Disparity mapping
C. Camera calibration
D. Point cloud processing

Synthetic solution to the question:
B

Explanation:
The correct answer is B. Disparity mapping. In stereo vision, the depth of objects in a scene is determined by calculating the disparity between corresponding points in the left and right images. Disparity mapping involves finding the pixel-level differences between the two images to estimate the depth information.

I Instruction-following test

We evaluated the models’ instruction-following capabilities using the Arabic versions of Vicuna-80 (Chiang et al., 2023), translated by GPT-4 and refined by native speakers. Following the methodology in (Chiang et al., 2023), GPT-4 was used as the evaluator, assigning scores to each model’s performance compared to GPT-3.5 Turbo, with a temperature setting of 0.2. For each question, GPT-4 independently scored the responses from both the evaluated model and GPT-3.5 Turbo. The average performance ratio of the evaluated model was calculated by dividing its overall score by that of GPT-3.5 Turbo. Results in Table 13 indicate that AraLLaMA models outperform their counterparts in Arabic Vicuna-80. Notably, AraLLaMA-7B exceeds Jais-13B by approximately 17%, despite having a smaller model size.

Model	Ratio of GPT-3.5
Jais-13B	75.40%
Llama-7B	78.99%
AraLLaMA-7B	92.71%

Table 13: Performance ratio of GPT-3.5 Turbo in Arabic Vicuna-80.

J Details of Ablation Study

J.1 Experiment Settings:

We undertook continuous pre-training on a 1B-parameter TinyLLaMA model (Zhang et al., 2024), which is derived from the LLaMA architecture and was initially trained on an English corpus comprising 3 trillion tokens. The pre-training regimen was segmented into five distinct stages, during which

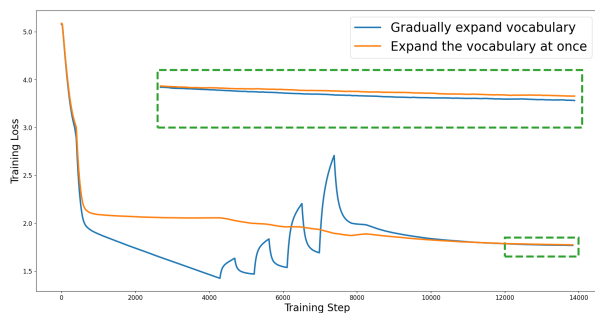


Figure 3: Loss curve of TinyLLaMa with sliding window average

0, 16, 64, 256, and 1024 Arabic subwords were progressively added to the vocabulary. Each stage allocated a different volume of data, totaling 80 billion tokens, with the proportion of Arabic to English data gradually shifting from 0:10 to 9:1. In a parallel experiment, we introduced 1024 subwords to the vocabulary in a single step, maintaining the same total token count and data distribution as in the phased approach. Both experiments adhered to an identical learning rate strategy, reinstating a cosine learning rate scheduler at the onset of each stage, starting with an initial rate of $1e-5$ and tapering to $2e-6$, with the initial 5 billion tokens of each stage designated for warm-up. Utilizing 192 GPUs, the experiments were conducted with a batch size of 3072.

J.2 Progressive Vocabulary Expansion

Pre-training

The results shown in Figure 3 demonstrate that the strategy of progressively expanding the vocabulary, which applies a sliding window average technique, yields a reduced final loss. Furthermore, as evidenced in Table 14, within the ArabicMMLU dataset, the approach of incrementally introducing new vocabulary items consistently outperforms the method of a one-time vocabulary expansion. This pattern underscores the effectiveness of gradual vocabulary enhancement in optimizing model performance.

Model	STEM	Social Sciences	Humanities	Arabic Language	Other	Avg
Expand vocab at once	28.6	26.7	28.1	24.4	30.1	27.0
Gradually expand vocab (ours)	29.8	27.1	27.2	24.6	31.4	27.3

Table 14: Zero-shot evaluation for TinyLLaMA in ArabicMMLU (Koto et al., 2024) with option logit probability