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

Jianqing Zhu[†1], Huang Hunag[†2], Zhihang Lin[†3], Juhao Liang[†2,3], Zhengyang Tang[†2,3], Khalid Almubarak[1], Mosen Alharthi[1], Bang An[1], Juncai He[1], Xiangbo Wu[2], Fei Yu[3], Junying Chen[2,3], Zhuoheng MA[3], Yuhao Du[3], Yan Hu[3], He Zhang[3], Emad A. Alghamdi[1], Lian Zhang[2], Ruoyu Sun[2,3], Haizhou Li[2,3], Benyou Wang[2,3], and Jinchao Xu[1]

[1]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
[2]Shenzhen Research Institue of Big Data, Shenzhen, China
[3]The Chinese University of Hong Kong, Shenzhen, China

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 ChatGPT 3.5, due to a predominant focus on mainstream languages (e.g., English and Chinese). One practical objective for an Arabic LLM is to utilize an Arabic-specific vocabulary for the tokenizer that could speed up decoding. However, using a different vocabulary often leads to a degradation of learned knowledge since many words are initially out-of-vocabulary (OOV) when training starts. Inspired by the vocabulary learning during Second Language (Arabic) Acquisition for humans, the released AceGPT-v1.5 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 lowering the ratio of out-of-vocabulary (OOV) words at every stage. The ablation study demonstrated the effectiveness of Progressive Vocabulary Expansion. Moreover, AceGPT-v1.5 achieves decent performance comparable to the best Arabic LLMs across a variety of Arabic benchmarks. Models, training data, benchmarks, and codes will be all open-sourced.

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 world [1], the development of models such as Jais and AceGPT marks a significant step forward, yet these models do not rival the capabilities of state-of-the-art models like GPT-4 or even ChatGPT 3.5. In line with the democratization Touvron et al. (2023), our development of Arabic Large Language Models (LLMs) focuses on language adaptation settings that utilize existing standard LLM architectures (like LLaMA) and well-trained weights, thereby saving computing resources and ensuring compatibility.

The core challenge in language adaption for English-centric Large Language Models (LLMs) for a second language is about vocabulary expansion Touvron et al. (2023); Cui et al. (2023); Huang et al. (2023); Zhao et al. (2024). A case in point is AceGPT Huang et al. (2023), which struggles with slow decoding speeds due to its inability to adapt to the Arabic vocabulary. It decodes Arabic words into sequences of alphabetical letters rather than at a

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

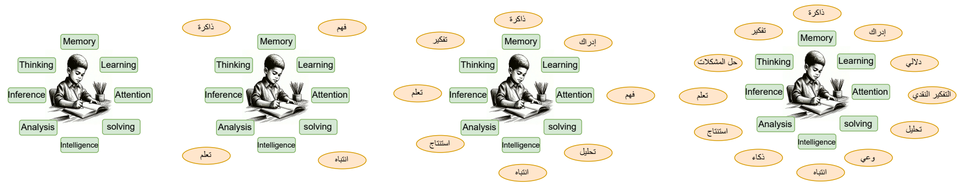

Figure 1: Second language acquisition for human, an English-speaking Child's Journey to Arabic Fluency, From Basic Vocabulary to Cultural Proficiency

more efficient granularity, such as Arabic subwords. This inefficiency significantly limits its broader applicability, despite its performance being nearly on par with ChatGPT 3.5 in some benchmarks. The primary challenge associated with vocabulary expansion is the risk that abrupt increases can lead to a high incidence of OOV words. 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 effectively restore and maintain the model's linguistic capabilities.

The core philosophy behind AceGPT-v1.5 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. AceGPT-v1.5 progressively extends the Arabic subwords in its vocabulary during pre-training, effectively reducing the ratio of out-of-vocabulary (OOV) words at every stage. By adopting this approach, AceGPT-v1.5, which is initialized with LLaMA2 13B, not only seamlessly preserves the inherent knowledge embedded in LLaMA2 13B but also facilitates a smooth transfer of knowledge 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, AceGPT-v1.5 achieves decent performance comparable to the best Arabic LLMs across a variety of 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 AceGPT-v1.5, a pioneering open-source Arabic Large Language Model that decodes Arabic texts three times faster than its predecessor Huang et al. (2023) 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. AceGPT-v1.5 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 pragmatic use. Of these language skills, vocabulary acquisition is crucial for language learning. Several studies have posited that L2 learners mostly learn new words incidentally Ramos & 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 5 (show in Appendix A) 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 adaption 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. (2023); Zhao et al. (2024), *a.k.a*, language adaption.

Vocabulary expansion in language adaption As a preliminary study, we identified Arabic tokens from the LLaMA2 vocabulary using regular expressions. It was observed that the LLaMA2 vocabulary only includes the basic characters of the Arabic language, resulting in relatively slow encoding and decoding speeds compared to English. During domain adaption, 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 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 adaption of LLMs?*

3 Methodology: Progressive Vocabulary Expansion for Language Adaption

The standard Byte Pair Encoding (BPE) process expands the initial vocabulary by iteratively merging frequent character pairs or sequences from training data into new tokens, until reaching a desired size. Training commences once this process is completed, rendering the vocabulary static. To investigate the posed question, this section introduces Progressive Vocabulary Expansion. This method incrementally incorporates new tokens in a dynamic vocabulary during training, mimicking a human-like paradigm of digesting and then learning during time.

In contrast to BPE algorithm Sennrich et al. (2015) that uses a static vocabulary during LLM training, we propose an incremental BPE(I-BPE) that uses dynamic vocabulary to implement Progressive Vocabulary Expansion, see Algorithm 3. Similar to the BPE process of repeatedly merging the most frequent pairs, gradually adding new tokens and training them equates to introducing new characters or subwords into the vocabulary, thus expanding and updating it. New tokens are continually added to the vocabulary until the vocabulary size is equal to the given number in each stage, and then the model is trained to adapt to the new

[2]The 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.

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

Initialized vocabulary V
Define the vocabulary size of each stage: $s_0, s_1, s_2, ..., s_n$
Define the proportion of training corpus corresponding to newly added tokens in each stage: $r_0, r_1, r_2, ..., r_n$
for $i = 0 : n$ **do**
    while $|V| < s_i$ **do**
        Calculate the frequency of adjacent token pairs in V
        Identify the most frequent pair or sequence, P_{freq}
        Merge P_{freq} to form a new token T_{new}
        Add T_{new} to vocabulary V
    end while
    Increase the proportion of corpus corresponding to newly added tokens to r_i
    Train model with this new V until convergence
end for
Finalize the vocabulary V for model training and application

vocabulary while increasing the proportion of corpus corresponding to newly added tokens. It repeats this expansion and annealing by gradually increasing both the vocabulary size and proportion of the corresponding corpus until the vocabulary is expanded to a preset size. This iterative approach could improve stability during language adaptation and maintain adaptability to existing data. Technically, this approach could substantially reduce the Out-of-Vocabulary (OOV) ratio at every step of the training process, thereby enhancing the model's capability to gradually recognize previously unknown words.

As seen in Figure 2, there exist two distinct strategies for vocabulary expansion: exponential addition of subwords or uniform addition.

- The **uniform expansion** involves adding K tokens at each stage. It results in a total number of $(T-1) \times K$ over T stages while the first stage does not add new tokens.
- The **exponential expansion** adds new tokens exponentially, mimicking the vocabulary learning mechanism observed in humans. Consistent with the uniform expansion, there is a stage at the beginning where no new tokens are added and then this approach starts with integrating one new token, with the number of tokens introduced in each subsequent stage doubling, following the sequence $\{0, 1, 2, \cdots, 2^{T-2}\}$, until reaching the desired expansion size.

Exponential expansion versus uniform expansion We conducted a comparative analysis of the impact of uniform and exponential vocabulary expansion strategies on token count using the same corpus. The encoding process was segmented into 16 distinct stages, with the token count computed at each stage using the correspondingly expanded vocabulary. Figure 2 illustrates the trend in token counts for both vocabulary expansion methods as the number of stages progresses. As observed in Figure 2, the uniform expansion leads to a significant increase in the compression ratio during the initial stages, but it becomes saturated later on. This could introduce training instability at the beginning, as it suddenly encounters a high ratio of new words. Such instability could harm large language models in terms of exacerbating catastrophic forgetting. Exponential growth facilitates a gradual adjustment in the compression ratio, Therefore, we opted for the exponential addition of subwords. Finally, it shortens the length of the decoded sequence by threefold, which could lead to significant speedup during both training and inference.

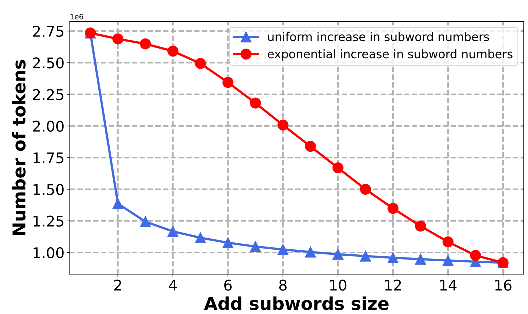

Figure 2: The impact on compression ratios for *uniform* or *exponential* vocabulary expansion.

4 Training

We will discuss the details of data engineering in Section 4.1, along with additional training details in Section 4.2.

4.1 Data Engineering

Pre-training Corpora Our pre-training dataset comprises both Arabic and English corpora. We employed an array of Arabic corpora encompassing multiple categories as delineated in Table 6 (shown in Appendix B). These include a filtered version of Common Crawl, WebText, and Wikipedia1 sourced from Joud and BAAI, all of which were subjected to an additional cleaning process. Moreover, we gathered and methodically purified additional corpora, namely Wikipedia2, Books, and Newspapers. The English corpus is sourced from SlimPajama Soboleva et al. (2023) and Proof-Pile-2 Azerbayev et al. (2023).

Incorporating the insights gained from our discussion on token annealing, this study further delves into the pre-training process, showcasing the integral role of the token annealing strategy in shaping our pre-training stages. Our pre-training framework is meticulously segmented into two epochs, with the inaugural epoch deploying the vocabulary annealing algorithm to fine-tune data distribution, as previously delineated. The subsequent epoch advances with training predicated on the refined vocabulary. The process of vocabulary expansion is methodically organized into 16 delineated stages, each uniquely composed of a calibrated mix of data from English, Arabic, mathematical, and coding domains, with the precise ratios detailed in Table 7 (show in Appendix C). A corpus of 30 billion tokens is employed for training across each stage, underscoring the extensive scale of our pre-training efforts.

The strategic design of these stages showcases a deliberate, phased approach towards the integration of new tokens. This facilitates a seamless adaptation of the model to a broad spectrum of data representations, ensuring a comprehensive understanding and engagement with various linguistic and symbolic nuances. By judiciously modulating the data composition at every stage—wherein the percentage of Arabic data steadily increases, reflecting a focused effort to bolster the model's proficiency with Arabic, simultaneously with a corresponding decrement in the English data percentage—we guarantee the model's agility and proficiency across a wide linguistic spectrum.

Data for Instruction Tuning After pre-training, we aim to elicit the knowledge out of AceGPT-v1.5 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, subfields, and individual 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 F.

Armed with this extensive collection of subjects and knowledge points, we direct the large language model (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 using `GPT-3.5-Turbo`. The distribution of topics in this data is shown in Figure 4 (shown in Appendix 4).

We also incorporated instruction tuning data from previous AceGPT projects Huang et al. (2023), including 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

In refining our methodology for the LLaMA2 model's vocabulary expansion to enhance its handling of Arabic, we not only identified and integrated 12,800 new Arabic subwords using the Incremental Byte Pair Encoding (BPE) method but also adjusted the language content ratio at each of the 16 training stages [3], see details in Appendix C. Following the expansion of the vocabulary through the aforementioned stages, to further enhance the model's performance, we continued training on an additional 20B data based on the expanded vocabulary.

In this paper, we continue pre-training on LLaMA2 models, which have 7 billion (7B) and 13 billion (13B) parameters, using a computational framework composed of 2,396 GPUs. We employ a model parallelism of 2 and a pipeline parallelism of 4. Optimization was carried out using the AdamW optimizer, with a context length of 4,096 tokens for each model. At the start of every training stage, we reintroduced a cosine learning rate scheduler with an initial rate of 1e-5 and decreased to 2e-6, ensuring a gradual adaptation through a 15% warm-up period at the beginning of each stage. Gradient accumulation was set at 8, achieving a total batch size of 4,736 and enabling the processing of approximately 0.019 billion tokens per batch.

5 Experiments

5.1 Experimental settings

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. (2021a)	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. (2023)	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 1: Overview of Evaluation benchmarks

Benchmarking Datasets As shown in Table 1, we employ four popular benchmarks aimed at assessing world knowledge: (1) ***MMLU*** (Multiple-Choice Multimodal Language Understanding) - This dataset is designed to measure the knowledge acquired during pretraining. For this benchmark, we employ both the original English version from Hendrycks et al. (2021b) and the Arabic version proposed by Huang et al. (2023), ensuring comprehensive coverage. (2) ***RACE*** (Reading Comprehension from Examinations) - A large-scale reading comprehension dataset designed to evaluate the educational knowledge of the models. (3) ***EXAMS*** (Multi-subject High School Examinations Dataset for Cross-lingual and Multilingual Question Answering) - Different from the previous benchmarks, EXAMS provides a diverse range of subjects for evaluation. (4) ***ArabicMMLU*** - Similar to the global MMLU, this dataset is specifically tailored for original Arabic LLMs, encompassing various countries and subjects. Additionally, evaluating Arabic cultural and value alignment is crucial. To assess this, we utilize ***ACVA-all*** and ***ACVA-clean*** for localization testing. To comprehensively evaluate model performance on inference and reasoning ability, we translate two commonsense reasoning benchmarks of varying difficulty: ***BoolQ*** and ***ARC-Challenge (ARC-C)***.

[3] In principle, a stageless solution could be employed, allowing the addition of one token after another without the need to define the boundaries between stages. However, for the sake of simplifying the implementation, particularly in terms of data preparation, we have opted for a staged approach where we make the number of stages $N = 16$.

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. (2023). 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] Huang et al. (2023)**: This set includes fully fine-tuned generative text models based on LlaMA2, specifically customized for the Arabic language 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 designed 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. (2023) and ArabicMMLU Koto et al. (2024). The left side of Table 8 details the models' accuracies on the Arabic MMLU translate dataset within a few-shot setting. It is evident from the data that the AceGPT-v1.5-7B and AceGPT-v1.5-13B models exhibit superior accuracy rates compared to models of similar scale. Notably, the AceGPT-v1.5-13B model outperforms the Jais-30B model, which has a significantly larger parameter count.

Additionally, the right side of Table 8 presents the accuracy results of models in a zero-shot learning scenario. Here again, the AceGPT-v1.5 models stand out for their exceptional performance, even when compared to models with similar parameter sizes. In particular, the AceGPT-v1.5-13B-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 AceGPT-v1.5 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.

	Arabic-trans MMLU Huang et al. (2023)					ArabicMMLU Koto et al. (2024)						**Total**
Model	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
AceGPT-v1.5-7B-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
AceGPT-v1.5-13B-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
ChatGPT 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. (2023) 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.

Evaluation on Chat Models Table 3 presents the comprehensive evaluation results across various benchmarks for the candidate models, spanning from Arabic to English. Overall,

Models	Arabic								English			Total
	MMLU (trans)	MMLU Koto et al. (2024)	EXAMS	ACVA clean	ACVA all	BoolQ (trans)	ARC-C (trans)	**Avg.**	BoolQ	RACE	**Avg.**	**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
AceGPT-v1.5-7B-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
AceGPT-v1.5-13B-chat	**47.33**	**61.70**	**48.37**	**76.90**	**76.37**	**69.33**	**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
ChatGPT 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.

AceGPT-v1.5 outperforms all baseline models in the Arabic language 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 highlights the model's expertise in addressing Arabic knowledge-related questions. Additionally, AceGPT-v1.5 demonstrates strong performance in tasks related to Arabic culture and value alignment. In terms of commonsense reasoning, AceGPT-v1.5 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 AceGPT-v1.5 models showcased unparalleled performance. The AceGPT-v1.5-13B-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 ChatGPT 3.5 Turbo in the "All set" category. This performance not only highlights the AceGPT-v1.5 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 AceGPT-v1.5-13B-chat model's superior results.

6 More Analysis

6.1 Ablation Study on Progressive Vocabulary Expansion

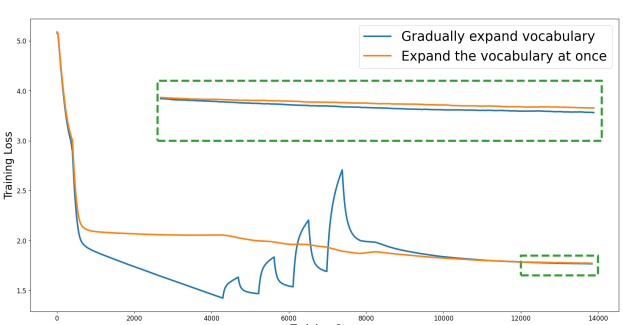

Figure 3: Loss curve of TinyLLaMa with sliding window average

Model	ArabicMMLU Koto et al. (2024) STEM	Social Sciences	Human-ities	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 4: Zero-shot evaluation for TinyLLaMA in ArabicMMLU with option logit probabiltiy

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 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.

In the analysis associated with Figure 3, which applies a sliding window average technique, it is observed that the strategy of progressively expanding the vocabulary yields a reduced final loss. Furthermore, as evidenced in Table 4, 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.

6.2 Compression Ratios

An encoding comparison was conducted on a consistent corpus to evaluate the compression efficiency of the vocabularies from LLaMA (AceGPT) and AceGPT-v1.5, using LLaMA as the benchmark. AceGPT-v1.5 notably enhanced the baseline by achieving a token compression ratio of 0.3174, following the augmentation of its vocabulary with 12,800 Arabic subwords.

6.3 Benchmarking in English dataset

We evaluated the accuracy of both base and chat models on the English MMLU dataset. As illustrated in Table 8 (shown in Appendix E), in the base model category, AceGPT-v1.5'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 Supervised Fine-Tuning (SFT), AceGPT-v1.5 achieves the highest accuracy among models of similar size and surpasses the Jais-30B model, which has a greater number of parameters.

7 Conclusion

The adaptation of large-scale models to less commonly spoken languages is fraught with challenges, notably the hurdles of knowledge transfer and the prevalence of out-of-vocabulary (OOV) terms. To address these issues specifically for Arabic, we developed a novel annealing training algorithm. This strategy methodically expands the vocabulary and employs a phased training process, leading to the development of the AceGPT-v1.5 7B and 13B models. Subsequent evaluations of both the base and chat configurations across diverse datasets have unequivocally established AceGPT-v1.5's superior accuracy compared to peers within the same parameter range. Remarkably, AceGPT-v1.5 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 access 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.

References

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics, 2023.

Sahil Chaudhary. Code alpaca: An instruction-following llama model for code generation. `https://github.com/sahil280114/codealpaca`, 2023.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

Yiming Cui, Ziqing Yang, and Xin Yao. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*, 2023.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 5427–5444. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.438. URL `https://doi.org/10.18653/v1/2020.emnlp-main.438`.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021a. URL `https://openreview.net/forum?id=d7KBjmI3GmQ`.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021b.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Ziche Liu, et al. Acegpt, localizing large language models in arabic. *arXiv preprint arXiv:2309.12053*, 2023.

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. Mistral 7b, 2023.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, et al. Arabicmmlu: Assessing massive multitask language understanding in arabic. *arXiv preprint arXiv:2402.12840*, 2024.

Stephen Krashen. Second language acquisition. *Second Language Learning*, 3(7):19–39, 1981.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1082. URL `https://aclanthology.org/D17-1082`.

Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, et al. Synthetic data (almost) from scratch: Generalized instruction tuning for language models. *arXiv preprint arXiv:2402.13064*, 2024.

I. S. P. Nation. *Learning Vocabulary in Another Language*. Cambridge Applied Linguistics. Cambridge University Press, 2001.

Restrepo Ramos and Falcon Dario. Incidental vocabulary learning in second language acquisition: A literature review. *Profile Issues in TeachersProfessional Development*, 17(1): 157–166, 2015.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, et al. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*, 2023.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

Daria Soboleva, Al-Khateeb Faisal, Myers Robert Steeves Jacob R, Hestness Joel, and Dey Nolan. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. `www.cerebras.net/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama`, June 2023. URL `https://huggingface.co/datasets/cerebras/SlimPajama-627B`.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

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

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

Jun Zhao, Zhihao Zhang, Qi Zhang, Tao Gui, and Xuanjing Huang. Llama beyond english: An empirical study on language capability transfer. *arXiv preprint arXiv:2401.01055*, 2024.

A CEFR Language Proficiency Levels

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 5: CEFR Language Proficiency Levels. 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.

B Arabic data distribution

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 6: Arabic data distribution and elapsed epochs

C Data mixture

The details of incremental vocabulary and data mixture and the are shown in Tab. 7.

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

D SFT data distribution of topics

Figure 4: A word cloud showing the distribution of topics, with the font size indicating the relative amount of data for each topic.

E Evaluation of models in English MMLU dataset

Model	few shot English MMLU					zero shot English MMLU					**Total**
	STEM	Humanities	Social Sciences	Others	Avg.	STEM	Social Sciences	Humanities	Other	Avg.	**Avg.**
LLaMA2-7B	**40.00**	**51.95**	52.42	**50.89**	**48.81**	31.49	31.26	38.35	38.80	34.97	41.89
AceGPT-7B	36.09	46.33	49.19	46.23	44.46	33.91	43.85	49.47	45.38	43.15	43.81
AceGPT-v1.5-7B	38.44	49.62	**53.32**	50.61	48.00	**45.49**	**63.55**	**66.05**	**59.25**	**58.59**	**53.30**
LLaMA2-13B	47.28	**63.55**	64.33	57.97	**58.28**	-	-	-	-	-	-
Jais-13B	27.14	14.38	45.64	41.13	32.11	44.33	55.14	61.39	56.06	54.23	43.17
AceGPT-13B	46.66	61.39	63.37	56.12	56.88	39.88	52.18	58.51	49.61	50.04	53.46
AceGPT-v1.5-13B	**47.31**	62.47	**64.77**	**58.14**	58.17	**51.48**	**66.71**	**71.65**	**61.72**	**62.89**	**60.53**
Jais-30B	27.42	14.60	45.84	41.43	32.32	39.23	44.51	52.96	50.91	46.90	39.61
ChatGPT 3.5 Turbo	58.39	72.12	78.02	69.95	69.62	-	-	-	-	-	-

Table 8: Evaluation of models in English MMLU dataset: few-shot on base model and zero-shot on chat model. Under the zero-shot setting, the LLaMA2-13B model does not follow instructions for unknow reason.

F 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.

F.0.1 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"
```

F.0.2 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"
```

F.0.3 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
- Mach number
- Key characteristics of hypersonic flows
Lecture title: Fundamentals of Shock Waves
Knowledge points:
- Definition of shock waves
- Formation of shock waves
- Types of shock waves
Lecture title: High-Temperature Gas Dynamics
Knowledge points:
- Definition of high-temperature gas dynamics
- Behavior of high-temperature gases
- Effects of high-temperature gases on materials
Lecture title: Principles of Rarefied Gas Dynamics
Knowledge points:
- Definition of rarefied gas dynamics
- The continuum hypothesis
- Governing equations
Lecture title: High-Speed Flow Over Bodies
Knowledge points:
- High-speed flow characteristics
- Impact on the body
- Aerodynamic heating
Lecture title: Hypersonic Vehicle Configurations
Knowledge points:
- Types of hypersonic vehicles
- Vehicle configurations
- Advantages and limitations of each configuration
Lecture title: Aerothermodynamics of Hypersonic Flows
Knowledge points:
- Definition of aerothermodynamics
- Aerothermodynamics in hypersonic flows
- Heat transfer in hypersonic flows
Lecture title: Hypersonic Flow Control
Knowledge points:
- Importance of flow control
- Methods of hypersonic flow control
- Challenges in hypersonic flow control
Lecture title: Hypersonic Propulsion Systems
Knowledge points:
```

```
- Types of hypersonic propulsion systems
- Working principles
- Advantages and disadvantages
Lecture title:  Future Trends in Hypersonic and High-Speed Flows
Knowledge points:
- Current research in the field
- Potential future trends
- Challenges and opportunities
```

F.0.4 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.
```