ACLM: Developing a Compact Arabic Language Model

Mohamed Alkaoud
Department of Computer Science, King Saud University, Saudi Arabia
malkaoud@ksu.edu.sa

Muteb Alsaqoub
Department of Computer Science, King Saud University, Saudi Arabia
441103819@student.ksu.edu.sa

Ibrahim Aljodhi
Department of Computer Science, King Saud University, Saudi Arabia
442105900@student.ksu.edu.sa

Abdulrhman Alqadibi
Department of Computer Science, King Saud University
Saudi Arabia
442103115@student.ksu.edu.sa

Omar Altammami
Department of Computer Science, King Saud University
Saudi Arabia
442102460@student.ksu.edu.sa

Abstract: *Recent advancements in Large Language Models (LLMs) have transformed Natural Language Processing (NLP). These models have demonstrated unprecedented capabilities in understanding and generating human language. However, their large-scale nature often poses challenges related to computational resource requirements, latency, and deployment, especially in resource-constrained environments. This research focuses on the design, development, and evaluation of an Arabic Small Language Model (SLM), named the Arabic Compact Language Model (ACLM), built to be compact and efficient. ACLM aims to bridge the gap between the high resource demands of existing large-scale models and the practical needs of real-world applications by leveraging high-quality Arabic data. We began with an existing language model, Pre-Trained Transformer for Arabic Language Generation (AraGPT2)-base, and further pre-trained it on high-quality Arabic data to enhance its performance while maintaining a compact size. This approach emphasizes the importance of data quality over model size, drawing on insights from recent studies that highlight the effectiveness of high-quality data in improving model performance. To evaluate ACLM, we conducted two key assessments: 1) A survey-based evaluation involving three LLMs: ChatGPT (GPT-4o), Gemini Pro, and Command R+, and 2) A perplexity analysis on generated and real-world text. ACLM outperformed AraGPT2-base in 4 out of 5 scenarios. Additionally, ACLM demonstrated superior fluency, achieving a perplexity of 31.74 on generated text compared to 165.28 for AraGPT2-base, and a perplexity of 124.67 on real-world Arabic books, significantly lower than 2011.88 for AraGPT2-base.*

Keywords: *Arabic NLP, deep learning, efficient AI, generative AI, GPT, large language models, natural language generation, NLP, small language models.*

Received July 27, 2024; accepted March 16, 2025
https://doi.org/10.34028/iajit/22/3/9

1. Introduction

The rapid advancements in Natural Language Processing (NLP) and Artificial Intelligence (AI) have significantly transformed the landscape of computational linguistics. Among the most notable developments is the emergence of Large Language Models (LLMs), which have demonstrated unprecedented capabilities in understanding and generating human language. These models, epitomized by architectures such as OpenAI's GPT [6, 7, 24, 28, 29], have been instrumental in pushing the boundaries of AI-driven language comprehension and generation. However, their large-scale nature often poses challenges related to computational resource requirements, latency, and deployment, particularly in resource-constrained environments.

While significant progress has been made in creating largescale language models, there remains a pressing need for smaller models that can operate effectively in environments with limited computational power. This is especially critical for applications in regions with constrained access to high-end computing infrastructure or for deployment on edge devices such as smartphones and Internet of Things (IoT) devices [38].

This research paper explores the design, development, and evaluation of an Arabic Small Language Model (SLM) that is built to be compact and efficient. The model, henceforth referred to as the Arabic Compact Language Model (ACLM), aims to bridge the gap between the high resource demands of existing large-scale models and the practical needs of real-world applications.

The drive for creating ACLM stems from several key motivations. First, there is a growing demand for NLP solutions that are not only accurate but also fast and responsive [38]. In many applications, such as real-time translation, online sentiment analysis, and conversational agents, latency is a critical factor [19]. Large models, despite their accuracy, often suffer from significant delays due to their size and complexity [22, 10]. ACLM addresses this by being lightweight and

optimized for speed, making it ideal for time-sensitive applications.

Second, the resource-intensive nature of large models limits their accessibility; high-performance computing resources are not universally available, especially in developing regions [27, 31, 32]. By developing a smaller model, we aim to democratize access to advanced NLP technologies, ensuring that more users can benefit from these advancements regardless of their computational capabilities. This aligns with the broader goals of inclusivity and equity in the distribution of technological benefits.

Third, the environmental impact of large-scale models is an increasingly important consideration. Training and deploying these models consume vast amounts of energy [9, 20, 21, 26], contributing to their carbon footprint. ACLM, by virtue of its compact size, requires significantly less computational power, thus promoting more sustainable AI practices. This aligns with global efforts to mitigate the environmental impact of technology and foster sustainable development.

The approach we will use to develop our language model is to fine-tune an existing Arabic SLM on high-quality data. By high-quality data, we mean text from textbooks and books. Unlike social media posts or internet comments, these sources are well-structured, and more reliable.

Gunasekar *et al*. [13] have shown in their work that it possible to train a competitive LLM that is significantly smaller than other LLMs. They introduce phi-1 which is a 1.3 billion LLM that was trained using 'textbook quality' data. They showed that emphasizing high-quality data enhances the model's performance.

To develop ACLM, we began with an existing language model, Pre-Trained Transformer for Arabic Language Generation (AraGPT2)-base [3], which contains 135 million parameters. Building on this foundation, we further pre-train the model on high-quality Arabic data. We focused on leveraging high-quality data to improve the model's performance because increasing the model size, which is often the simplest way to enhance a language model, was not an option since we want to keep the model compact.

This paper is structured as follows: we begin with a review of the existing literature on Arabic and compact language models. This is followed by a detailed explanation of the methodologies used in the development of ACLM. The subsequent section presents the results of our evaluations, comparing ACLM's performance against AraGPT2-base, followed by a discussion on the implications of our findings. We conclude by summarizing our work and outlining directions for future research. Our contributions can be summarized as:

1. We develop ACLM, a compact Arabic language model with impressive performance compared to existing small Arabic language models.
2. We showcase the importance of training Arabic language models on high-quality text.
3. We propose a comprehensive framework to evaluate a language model's capabilities across five distinct scenarios.
4. We showcase our model's capability to run efficiently on a laptop.

2. Background

The evolution of NLP has been significantly shaped by the development of transformer-based models. Introduced by Vaswani *et al*. [37], the Transformer architecture revolutionized NLP by enabling models to process and generate text with unprecedented accuracy and efficiency. Unlike traditional Recurrent Neural Networks (RNNs) [8, 14, 17, 30], Transformers leverage self-attention to capture dependencies between words, thus overcoming the limitations of RNNs in handling long-range dependencies and parallelizing training. Prior to the widespread adoption of Transformer-based models, classical word embeddings played a crucial role in NLP [5].

Building upon the Transformer architecture, the GPT series by OpenAI marked a pivotal advancement in NLP. GPT-1 [28] demonstrated the potential of unsupervised pre-training followed by supervised fine-tuning, significantly improving performance across various NLP tasks. The subsequent iteration, GPT-2 [29], scaled up the model size to 1.5 billion parameters and showcased impressive capabilities in text generation, summarization, and translation.

GPT-3 [6], the third generation of the GPT series, further amplified these capabilities with 175 billion parameters. It exhibited remarkable proficiency in few-shot and zero-shot learning, allowing it to perform tasks with minimal task specific data. The sheer scale and versatility of GPT-3 highlighted the transformative potential of LLMs in a wide array of applications.

InstructGPT [25], another significant development, aimed to make language models more aligned with user intentions. By fine-tuning GPT-3 using Reinforcement Learning from Human Feedback (RLHF), InstructGPT improved the model's ability to follow explicit instructions and generate responses that are more helpful and accurate. This approach addressed some of the limitations of previous models in understanding and adhering to user prompts. ChatGPT, another derivative of the GPT-3 model fine-tuned specifically for conversational contexts, epitomizes the practical application of these advancements. Designed to engage in human-like dialogue, ChatGPT has been widely deployed in customer service, virtual assistants, and other interactive systems. Its ability to generate coherent and contextually relevant responses has made it a cornerstone in the development of interactive AI systems. We notice a trend of each model having a larger size than its predecessors; a trend that we want to go

against in this paper and explore SLMs more.

3. Related Works

In this section, we discuss works related to our work. First, we have Jais [33]: An Arabic LLM developed through a collaboration between Cerebras Systems, Mohamed bin Zayed University of Artificial Intelligence, and Inception. It boasts up to 30 billion parameters and was trained on a massive dataset comprising 395 billion tokens, including 116 billion Arabic tokens and 232 billion English tokens.

Jais utilizes cutting-edge features like ALiBi position embeddings for improved context handling and accuracy, and SwiGLU for training efficiency. It was trained on the Condor Galaxy 1 AI supercomputer, showcasing significant advancements in model training and deployment. It comes in two model size: 13 billion and 30 billion. Both sizes come with instruction-tuned variants. While there's a lot of potential in Jais, even its smallest variant is still a very large model compared to what we are building.

Secondly, we have AraGPT2, an Arabic LLM developed by Antoun *et al*. [3] It is one of the first Arabic GPT-based language models, trained from scratch using a large corpus of Arabic internet text, Wikipedia, and news articles. The model comes in four variants: base, medium, large, and mega. The largest variant, AraGPT2-mega, containing 1.46 billion parameters. Additionally, the researchers developed and released an automatic discriminator model that achieves 98% accuracy in detecting model-generated text. This tool helps mitigate the potential misuse of the model, such as generating fake news. AraGPT2's smallest model has only 135 million parameters, but its generation capabilities are very limited as we will see in the rest of the paper.

SLMs often struggle with producing coherent and fluent text. This issue raises the question of whether coherence in text generation only emerges at larger scales with more complex architectures. Eldan *et al*. [11] explored this issue in depth and to address it, the TinyStories dataset was created, containing short stories understandable by 3 to 4-year-olds, generated by GPT-3.5 and GPT-4. This dataset enables training and evaluating smaller language models that still produce fluent, and grammatically correct stories with reasoning capabilities. TinyStories aims to advance the research of LMs, particularly in low-resource or specialized domains, and to explore the emergence of language capabilities in LMs.

Gunasekar *et al*. [13] explored in their work a new approach to training language models by focusing on high-quality data rather than sheer volume. They started with a 3 TB dataset of code and text from StackOverflow, selected 6 billion high-quality tokens, and used GPT-3.5 to generate 1 billion textbook-like tokens. They trained a 1.3 billion parameter model, phi-1, on this data and fine-tuned it with GPT-3.5- generated exercises. Phi-1 was impressive considering its size, demonstrating that data quality can, sometimes, outweigh data quantity and parameter size in model training.

The body of research on LLMs predominantly addresses either Arabic LLMs or compact LLMs, but there is a noticeable gap in the development of compact Arabic LLMs specifically. Existing works on Arabic LLMs, such as AraGPT2 and Jais LLM, have focused on leveraging large-scale datasets and significant computational resources to enhance Arabic NLP capabilities. Conversely, studies on compact LLMs emphasize creating smaller, efficient models that maintain high performance with fewer parameters, primarily in languages with simpler morphological structures like English. Arabic, with its complex morphology, presents unique challenges that require specialized models. Despite the advancements in both areas, no current research has effectively combined these focuses to create a compact, high-performing LLM tailored for Arabic as far as we know. This gap highlights the need for innovative approaches that integrate the principles of compact model design with the specific requirements of the Arabic language, addressing both efficiency and linguistic complexity.

4. Methodology

4.1. The Problem with Arabic Small Language Models

Before discussing our model, we aim to analyze AraGPT2 in more detail. Table 1 presents the different versions of AraGPT2 in terms of the number of parameters. In this analysis we will focus on the smallest two versions: base and medium. To test these models, we used three example prompts from the default examples on the models' pages on Hugging Face. We generated completions for each example using the Hugging Face Inference API. This approach allowed us to directly compare the outputs of the base and medium variants of AraGPT2, providing insights into their performance in generating coherent and contextually appropriate text. The results were then analyzed to evaluate the linguistic capabilities and differences between the two models. Tables 2 and 3 showcases the generated completions from both the base and medium versions, respectively.

Table 1. AraGPT2-base generated completions on three examples. Black text indicates the prompt. Blue text indicates the completion.

Prompt	Completion
يحكى أن مزارعا مخادعا قام ببيع بئر الماء الموجود في أرضه لجاره مقابل مبلغ كبير من المال	يحكى أن مزارعا مخادعا قام ببيع بئر الماء الموجود في أرضه لجاره مقابل مبلغ كبير من المال، وعندما علم أن هذا المبلغ هو ثمن البنر، قام ببيعها لجاره مقابل مبلغ كبير من المال
القدس مدينة تاريخية، بناها الكنعانيون في	القدس مدينة تاريخية، بناها الكنعانيون في القرن الاول قبل الميلاد، وهي مدينة كنعانية قديمة تقع في جنوب فلسطين، وهي مدينة كنعانية قديمة
كان يا ما كان في قديم الزمان	كان يا ما كان في قديم الزمان كان في قديم الزمان كان في قديم الزمان كان في قديم الزمان كان في قديم الزمان

Table 2. AraGPT2-medium generated completions on three examples. Black text indicates the prompt. Blue text indicates the completion.

Prompt	Completion
يحكى أن مزارعا مخادعا قام ببيع بئر الماء الموجود في أرضه لجاره مقابل مبلغ كبير من المال	يحكى أن مزارعا مخادعا قام ببيع بئر الماء الموجود في أرضه لجاره مقابل مبلغ كبير من المال، وعندما سأله عن سبب قيامه بذلك، قال له: " أنا لا اريد أن أبيع الماء،
القدس مدينة تاريخية، بناها الكنعانيون	القدس مدينة تاريخية، بناها الكنعانيون في الألف الثالثة قبل الميلاد، وهي مدينة كنعانية قديمة، تقع في شمال فلسطين، على بعد حوالي 20
كان يا ما كان في قديم الزمان	كان يا ما كان في قديم الزمان، كان في قديم الزمان ، كان في قديم الزمان، كان في قديم الزمان، كان في قديم الزمان

Table 3. The size in parameters of the different AraGPT2 models.

Model	Size (parameters)
Base	135 million
Medium	370 million
Large	792 million
Mega	1.46 billion

For the prompt about a farmer selling a well, the base variant produces repetitive and less coherent text, such as repeating the act of selling for a large sum without providing additional context. In contrast, the medium variant generates a more logically consistent continuation, explaining the farmer's reason for not wanting to sell the water.

In the example discussing the city of Jerusalem, the base model falls into redundancy and fails to provide a complete, coherent sentence, repeating details about the city's ancient origins. However, the medium variant produces a more complete and accurate continuation, providing additional details about the city's location and historical timeline.

Lastly, for the prompt beginning with ''Once upon a time,'' the base variant demonstrates excessive repetition, repeating the phrase multiple times without furthering the narrative. The medium variant also fails in this example. It is interesting to note that the medium version added commas between each repetition.

These examples highlight that the base model struggles with generating coherent text, often falling into redundancy and incomplete sentences. In contrast, the medium model, with more parameters, shows marked improvement in fluency, coherence, and the ability to generate more contextually appropriate continuations. Although the medium model generations still have many issues, they are clearly better than the base. This analysis emphasizes the importance of model size (number of parameters) in achieving higher-quality language generation, and the difficulty of modeling Arabic using only 135 million parameters.

Another noteworthy behavior we observed in AraGPT2- base is that its generated text occasionally includes phrases like ''you are not registered'' or ''you do not have permission to access this page,'' as illustrated in Table 4. We believe this issue arises from the nature of the OSCAR corpus [35], which was used to train AraGPT2-base. The OSCAR corpus comprises a vast array of text data extracted from the web, including content from online forums. These forums often contain access restrictions, requiring users to register or log in before viewing certain pages. Consequently, the model has inadvertently learned to replicate these access control messages from its training data, leading to their occasional appearance in generated outputs.

Table 4. Examples of AraGPT2-base ''you are not registered'' / ''you do not have permission to access this page'' generations.

#	Example
1	أنت لم تسجل الدخول بعد أو أنك لا تملك صلاحية لدخول لهذه الصفحة. هذا قد يكون عائدا لأحد هذه الأسباب : ليست لديك صلاحية أو إمتيازات كافية لدخول هذه الصفحة. هل تحاول تعديل مشاركة شخص آخر, دخول ميزات إدارية أو نظام متميز آخر ؟ لوحة تحكم العضو الرسائل الخاصة الاشتراكات المتواجدون الآن البحث في المنتدى رئيسية المنتدى [][^][الأقسام التعليمية][^][] دورات المنتدى التعليمية المجانية انتاجات المنتدى والأعمال الحصرية حقيبة البرامج والكتب والملفات اسطوانات اللغة الانجليزي الكاملة قواعد اللغة الإنجليزية الأفعال والأزمنة كلمة و
2	أنت لم تسجل الدخول بعد أو أنك لا تملك صلاحية لدخول لهذه الصفحة. هذا قد يكون عائدا لأحد هذه الأسباب : ليست لديك صلاحية أو إمتيازات كافية لدخول هذه الصفحة. هل تحاول تعديل مشاركة شخص آخر, دخول ميزات إدارية أو نظام متميز آخر ؟
3	عزيزي الزائر سجلاتنا تفيد انك لست عضو لدينا في المنتدى, في حال رغبتم بالانضمام الى أسرتنا في المنتدى ينبغي عليكم ملء النموذج التالي!التسجيل الأعضاء الأكثر تميز خلال 7 أيام التعليمات التقويم البحث مشاركات اليوم اجعل كافة الأقسام مقروءةتلاوة القرآن الكريم مركز تحميل كرم نت الرئيسية حياة الرسول الكريم الإعجاز العلمي مملكة حواء مملكة الطفل الأقسام التقنية والبرامج فيس بوك قسم الرواتر و الفضائيات والست

4.2. Data Collection

To build ACLM, the initial step involves careful data collection. We want to curate a diverse and extensive ''high-quality'' dataset comprising written text in Modern Standard Arabic (MSA), including a variety of sources such as published books and newspapers.

Constraining ourselves to published books ensures high quality, as these sources are well-established, widely recognized, and thoroughly vetted. This focus reflects a commitment to reliability, accuracy, and the depth of insight that often accompanies extensively reviewed and edited material. Books authored by subject-matter experts bring specialized knowledge, ensuring that the information presented is grounded in a deep understanding of the subject matter, contributing to the overall accuracy of the model's knowledge.

In addition to books, newspapers will be incorporated into our dataset to broaden the scope. This inclusion provides insights across diverse genres, such as politics and economics, enhancing the model's understanding and exposure to various language styles and topics. Newspapers offer real-time glimpses into societal trends, global developments, and diverse perspectives, enriching the model's ability to comprehend and respond to contemporary issues.

The structured and organized nature of published books provides a foundational framework that significantly benefits the training of LLMs. We hypothesize that the logical flow of information in books helps the language model internalize how information should be logically connected and presented, enhancing its communicative effectiveness. Considering these advantages, our decision to focus on collecting data from published books and newspapers is significant for shaping the outcome of our LLM. And here is some more information on the data we collected:

1. Hindawi book library dataset: One of the reasons we chose the Hindawi Library is its extensive collection of genres, in addition to its quality. The library includes 3, 517 books spanning different genres. It's important to note that the library is frequently updated, and it had 3, 517 when we collected the books. Each book in the library can be either viewed as multiple HTML files or downloaded in multiple formats: PDF, EPUB (e-book format), and/or KFX (Amazon Kindle File Format). We scrapped the books from the HTML files containing the books. We chose HTML due to the difficulties associated with parsing text, particularly Arabic text, from PDF/EPUB/KFX files. Figure 1 illustrates the diverse genres covered by the Hindawi book library. In contrast to other Arabic book libraries such as OpenITI [23], the Hindawi library does not focus primarily on religious texts but includes a diverse range of genres, as shown in Figure 1.
2. Newspapers data set: We augment the Arabic News Article Dataset (ANAD) [2] with news that we've scraped from Alriyadh newspaper.

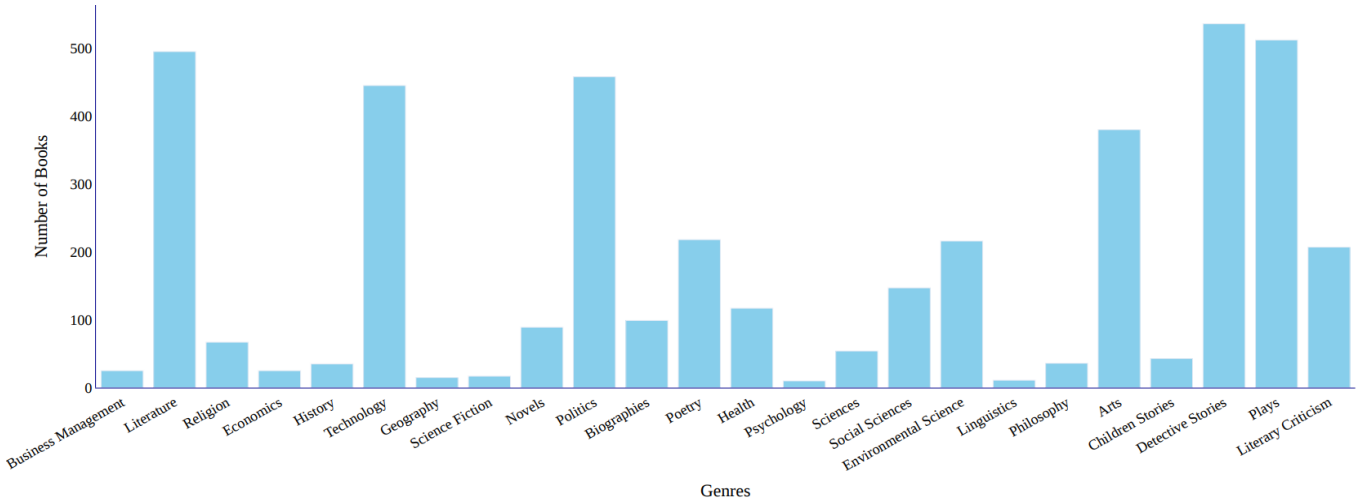


Figure 1. Hindawi book genre distribution. The x-axis represents the different genres. The y-axis represents the number of books.

Before processing the dataset, we had to ensure everything was in order so the training could proceed as smoothly as possible. The steps included removing any irrelevant symbols such as '•••', and '†' to help the model focus on linguistic patterns, concatenating all files to ensure uniformity and ease of processing, and making each line represent a single paragraph to eliminate any context confusion for the model. We fine-tuned AraGPT2-base which has 135,000,000 parameters on our dataset. In this paper, we will use the terms 'fine-tuning' and 'continuing pre-training' interchangeably.

There are different approaches to fine-tuning a pre-trained language model, each varying in scope and the extent to which model parameters are modified. Task fine-tuning involves adapting a model for a particular application, such as text classification, named entity recognition, or summarization. This approach typically relies on labeled data to optimize performance for a specific task. Another common approach is fine-tuning be freezing most of the pre-trained model's layers, and only the final layers are trained on task-specific data. More recent parameter-efficient fine-tuning techniques, such as LoRA [15], modify only a small number of parameters while keeping the majority of the model unchanged. These approaches allow for more efficient fine-tuning without significantly altering the underlying language model. However, because they only introduce minor modifications, they may not be suitable when substantial improvements to the model's overall language generation capabilities are required [4]. In contrast, continuing pre-training, also known as full-model fine-tuning, updates all model parameters by further training the model on a new dataset. Unlike other methods, continuing pre-training changes the model's behavior the most because it modifies all parameters, it results in the most significant improvements in language generation and adaptability [34].

Given AraGPT2-base's performance, as discussed in this section, we observed that the model often struggled to generate syntactically correct sentences in many cases. To address these shortcomings, we selected continuing pre-training as our fine-tuning approach. Since this method applies full-model updates rather than limited parameter modifications, it allows for deeper adaptation and more substantial improvements compared to other fine-tuning techniques.

4.3. Training

We start by continuing pre-training AraGPT2 on the book dataset we collected on a personal machine which has the following specs:

- CPU: Intel Core i5-10400F
- GPU: NVIDIA RTX 3060 12GB VRAM
- RAM: 16 GB

Unfortunately, the 12GB VRAM was a limiting factor that made training very slow due to the very small batch

size we could fit on the GPU. We then decided to rent a machine from vast.ai with the following specs:

- CPU: AMD EPYC 7763
- GPU: NVIDIA A100 SXM4 80GB VRAM
- RAM: 1032 GB

The hyperparameters used in training are shown in Table 5. However, despite these gains in language generation, the model's knowledge about the world was not as comprehensive as desired as shown in Table 7.

While it excelled in creating well-formed sentences, it occasionally lacked depth in factual accuracy. This indicates that while the model's architecture and training process succeeded in learning the structure of Arabic, further enhancements in the quality and breadth of the training data are necessary to improve its overall knowledge.

Table 5. The training hyperparameters from the first batch of books.

Hyperparameter	Value
Learning rate	5e-05
Batch size	22
Optimizer	Adam
Number of epochs	10

To address the model's lack of knowledge about the world, we continued pre-training our model by using a small sample of our news dataset, consisting of 9,116 samples. The aim was to enhance the model's understanding of real-world facts and events but, at the same time, don't forget the knowledge gained by training on the book's dataset. By integrating this curated dataset, we intended to supplement the linguistic proficiency of the model with relevant and accurate information. This approach aimed to create a more balanced language model that excels not only in generating coherent text but is also better in providing factually accurate content across various topics. The training of the second batch was done on the personal machine mentioned before. We did not use the cloud to train for two reasons:

1. The number of samples was small which means we could train on the entire data in reasonable time.
2. This research is self-funded, and we wanted to avoid paying more in cloud GPU fees.

The hyperparameters used in training are shown in Table 6. We decided to stop the training of the second batch (news) after we reach the same training loss we reached in the first batch (books); we achieved that after eight epochs. Figure 2 shows the loss over time during training for both the first batch (books dataset) and the second batch (a sample of the news dataset). The resulting model contains clear improvements compared to fine-tuning only on books as shown in Table 7.

Table 6. The training hyperparameters from the second batch of news.

Hyperparameter	Value
Learning Rate	5e-05
Batch Size	4
Optimizer	Adam
Number of Epochs	8

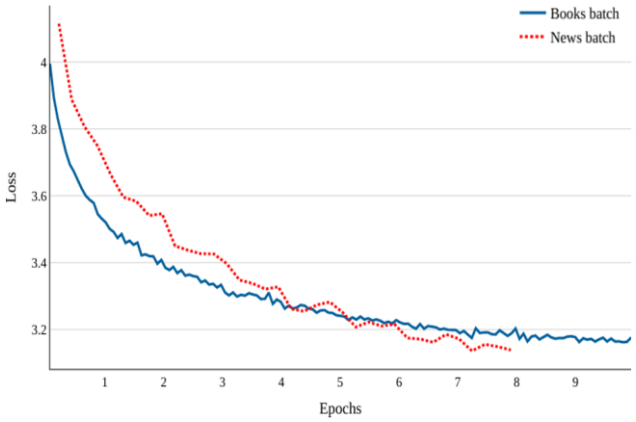


Figure 2. Loss curves plot when training ACLM. The x-axis represents the number of epochs. The y-axis represents the current loss. The blue curve corresponds to the training off the first batch. The red curve corresponds to the training of the second bat.

Table 7. ACLM generated completions on three examples (with and without news fine-tuning). Black text indicates the prompt. Blue text indicates the completion. Only 50 tokens of the completions were generated.

Prompt	ACLM completion (before news fine-tuning)	ACLM completion (after news fine-tuning)
ما هي آخر المستجدات في الاقتصاد العالمي؟	ما هي آخر المستجدات في الاقتصاد العالمي؟ ما هي آخر المستجدات في الاقتصاد العالمي؟ إن هذا السؤال لا يَعني إلا ما إذا كان من الممكن أن يؤدي إلى تغيير في طريقة تفكيرنا؛ أي إن علينا ألا ندرك أننا قد أصبحنا أكثر قدرة على التفكير النقدي، أو بالأحرى أقل قدرة ...	ما هي آخر المستجدات في الاقتصاد العالمي؟ - ارتفاع معدل البطالة إلى 4 %. - انخفاض معدلات الاستثمار الأجنبي المباشر إلى 3 % في الربع الأول من العام الحالي. - تراجع معدلات النمو السكاني إلى 2%. - انخفاض أسعار النفط الخام إلى 1 % في النصف الأول من العام الجاري. - استمرار...
ما هي أهم الأحداث في تاريخ فلسطين الحديث؟	ما هي أهم الأحداث في تاريخ فلسطين الحديث؟ أليس من الممكن أن يكون هناك سببٌ آخر يدعو إلى الاعتقاد بأن اليهود قد بدءوا يشعرون بأنهم جزء لا يتجزأ من العالم. ولكن لماذا إذن لم يُحذِّروا أنفسهم من هذا الاعتقاد؟ ...	ما هي أهم الأحداث في تاريخ فلسطين الحديث؟ ما هي أهم الأحداث في تاريخ فلسطين الحديث؟ - أحداث «إسرائيل» و«الإسرائيليين»، التي جرت خلال الفترة من عام 1947 - 1948 حتى عام 1969. - ما هو الدور الذي لعبته إسرائيل في الحرب العالمية الثانية؟ - دور الاحتلال الإسرائيلي في ...
ما هي أسباب الأزمة الاقتصادية في لبنان؟» – «أجل. لقد كانت هناك أزمة أخرى.» – «ولكنك كنتِ تُريدين مني أن أقول لك: إن الأزمة لم تكن أزمة اقتصادية بالمعنى المفهوم للكلمة ... بل كانت أزمة سياسية	ما هي أسباب الأزمة الاقتصادية في لبنان؟» – «أجل. لقد كانت هناك أزمة أخرى.» – «ولكنك كنتِ تُريدين مني أن أقول لك: إن الأزمة لم تكن أزمة اقتصادية بالمعنى المفهوم للكلمة ... بل كانت أزمة سياسية ...	ما هي أسباب الأزمة الاقتصادية في لبنان؟ - أزمة اقتصادية حادة نتيجة السياسات الخاطئة التي اتبعها الرئيس ميشال عون. - هل هناك حل سياسي للأزمة؟ - يجب إيجاد حكومة إنقاذية من الشعب لا أن تكون حكومة تصريف الأعمال فقط، بل حكومة تكنوقراط، وتكون قادرة على إدارة شؤون ...

4.4. Evaluation

To further validate these observations, we conducted a systematic evaluation to compare our model's performance against AraGPT2-base since our model,

was fine-tuned on AraGPT2-base, so they have the same number of parameters and architecture. Knowing this, we conducted two types of evaluations. The first evaluation involved administering a survey to three LLMs:

1. OpenAI's ChatGPT (GPT-4o).
2. Google's Gemini Pro [36].
3. Cohere's Command R+ [12].

We designed the survey to compare responses generated by our model and AraGPT2-base across five distinct scenarios:

1. Long prompt scenario: to evaluate how a model would respond to extended prompts and whether prompt length affected output quality.
2. Open-ended prompt scenario: these prompts lack a single correct answer, allowing subjective responses, essential for evaluating a model subjective content generation ability.
3. Conversational prompt scenario: using the conversational prompt scenario, we assessed a model conversational proficiency and human-like discourse patterns.
4. Storytelling prompt scenario: a model continued given stories to assess narrative coherence and storytelling abilities.
5. Information prompt scenario: factual prompts that aimed to evaluate a model's accuracy in delivering factual information.

Table 8 shows examples of the five scenarios. For each scenario, we developed two prompts, giving us a total of ten prompts (two for each scenario). This approach aimed to minimize bias and encourage a general evaluation rather than focusing on specific aspects. The questions were carefully crafted, starting with the prompt, and followed by responses/completions from both models '1' and '2'. Afterward, each LLM had to pick between them. It was all kept anonymous, so no model knew which model was which. Another way to combat bias was that the order was switched around, so sometimes 'Model 1' was ACLM, and sometimes it was AraGPT2-base. This switch added an extra twist, keeping things unpredictable and fair. The whole setup aimed to give a balanced and honest evaluation of the models without any biases sneaking in. The instructions were given in English, but the actual prompts and responses were in Arabic. For each question, we generate the following prompt:

> You will receive a prompt followed by two outputs from different models. Your task is to determine which model's output is superior.
>
> Prompt:<prompt>
>
> 1: <model_1_generation>
>
> 2: <model_2_generation>

And feed it to each LLM to get their responses. The reason we used English instructions instead of Arabic in the prompt is because it has been shown that prompting ChatGPT using English increases performance [1, 18]. The three placeholders in the prompt above (<prompt>, <model_1_generation>, <model_2_generation>) are obviously kept in Arabic; we only change the instructions to English.

Table 8. An example of each of the five scenarios.

Scenario	Example
Long prompt	الاقتصاد في الملكة العربية السعودية يعتمد على النفط كمصدر أساسي، و أيضًا يعتمد على قطاعات أخرى مثل
Open-ended prompt	نادي الهلال السعودي هو أفضل نادٍ بسبب
Conversational prompt	- قال: لم اعد كالسابق - لماذا ذلك ؟ - لأنني
Storytelling prompt	في مدينة صغيرة تسودها السلام، كان هناك رجلان يعيشان في حي واحد. الرجل الاول يعاني من إعاقة بصرية، بينما الرجل الثاني يمتلك حاسة نظر حادة
Information prompt	أعراض مرض السكري هي

The second evaluation we conducted involved measuring perplexity using examples generated by ChatGPT. We specifically instructed ChatGPT to create examples encompassing various styles. We chose to use ChatGPT for generating these examples to avoid the risk of selecting existing text that the models might have already encountered; given that AraGPT-2, and ACLM since it is based on it, was trained on internet data, which could include virtually anything. Perplexity [16] measures how well a language model predicts a sample of text, with lower values indicating that the model is more confident and accurate in its predictions. The generated examples are shown in the Appendix I. We also evaluated the perplexity of both models on three books from Al-Maktaba Al-Shamela[1].

5. Results and Discussions

In this section, we will present and discuss the results of comparing ACLM, our model, against AraGPT2-base.

First, we asked three LLMs (ChatGPT, Gemini Pro, and Command R+) to take part in the survey.

Table 9. Evaluation of ACLM and AraGPT2-base performance across different prompt types by ChatGPT, Gemini Pro and Command R+. The top-performing model in each scenario is indicated in bold text.

Scenario	ACLM Avg. Win%	AraGPT2 Avg. Win%
Long prompt	**66.67%**	33.33%
Open-ended prompt	**66.67%**	33.33%
Conversational prompt	**100.00%**	0.00%
Storytelling prompt	**50.00%**	**50.00%**
Information prompt	**66.67%**	33.33%

A total of 10 prompts were used in this experiment; a prompt for each question. The results are shown in Table 9. In the long prompt scenario, ACLM outperformed AraGPT2 with an average win percentage of 66.67%.

[1] https://shamela.ws/

Similarly, in the open-ended prompt scenario, ACLM also achieved a 66.67%-win rate. The most striking difference is observed in the conversational prompt scenario, where ACLM achieved a 100%-win rate. In the storytelling prompt scenario, both models performed equally well, each securing a 50%-win rate. Lastly, in the information prompt scenario, ACLM again led with a 66.67%-win rate. Overall, these results suggest that ACLM consistently outperforms AraGPT2 in most scenarios, particularly in generating conversational text.

We also measured the perplexity on the set of examples generated by ChatGPT. As shown in Figure 3, our model achieved a perplexity of 31.74, significantly outperforming AraGPT2-base, which achieved a perplexity of 165.28. This stark difference in perplexity indicates that ACLM generates more coherent and fluent text compared to AraGPT2-base.

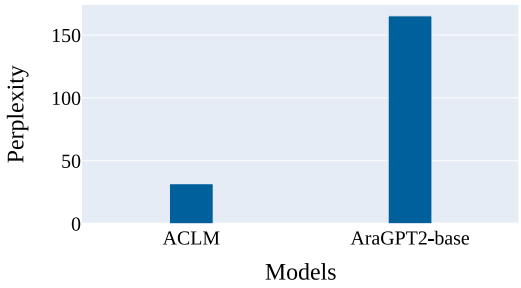


Figure 3. Perplexity of ACLM vs AraGPT2-base on the generated examples.

To further assess the models' generalization, we computed perplexity on three books sourced from Al-Maktaba Al-Shamela. ACLM achieved a perplexity of 124.67, while AraGPT2-base recorded 2011.88 as shown in Figure 4. These results further reinforce the superiority of ACLM over AraGPT2-base.

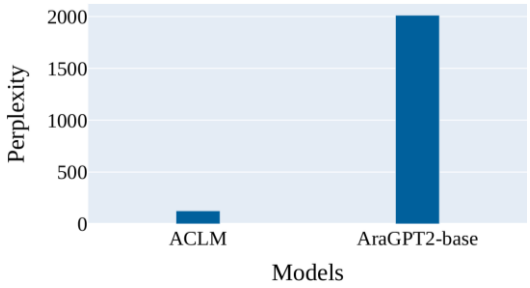


Figure 4. Perplexity of ACLM vs AraGPT2-base on the three books from Al-Maktaba Al-Shamela.

This substantial improvement in perplexity aligns with our qualitative observations, where ACLM consistently generated more fluent and syntactically accurate sentences compared to AraGPT2-base. To demonstrate that, we took the same prompts in Table 2 and used ACLM to complete them. The results are shown in Table 10. As we can notice, the completions are much better than AraGPT2-base. In the "كان يا ما كان في قديم الزمان" prompt, AraGPT2-base has a tendency to repeat phrases verbatim, especially in its response to the prompt, where it loops the phrase multiple times. ACLM, in contrast, provides a more fluid and natural-sounding completion, avoiding excessive repetition and instead offering a thoughtful narrative expansion. AraGPT2-base struggles with logical consistency in some cases, particularly in the first prompt. It redundantly states that the farmer sold the well twice for a high amount without adding new meaningful content. ACLM introduces more context and logical flow, explaining why the well was difficult to access, making the response richer and more coherent.

Table 10. ACLM vs AraGPT2-base generated completions on three examples. Black text indicates the prompt. Blue text indicates the completion.

Prompt	AraGPT2-base completion	ACLM completion
يحكى أن مزارعا مخادعا قام ببيع بئر الماء الموجود في أرضه لجاره مقابل مبلغ كبير من المال	يحكى أن مزارعا مخادعا قام ببيع بئر الماء الموجود في أرضه لجاره مقابل مبلغ كبير من المال، وعندما علم أن هذا المبلغ هو ثمن البئر، قام ببيعها لجاره مقابل مبلغ كبير من المال	يحكى أن مزارعا مخادعا قام ببيع بئر الماء الموجود في أرضه لجاره مقابل مبلغ كبير من المال، حيث إن البئر كانت تحت الأرض ولم يستطع أحد الوصول إلى البئر إلا بشق الأنفس.
القدس مدينة تاريخية، بناها الكنعانيون في	القدس مدينة تاريخية، بناها الكنعانيون في القرن الاول قبل الميلاد، وهي مدينة كنعانية قديمة تقع في جنوب فلسطين، وهي مدينة كنعانية قديمة	القدس مدينة تاريخية، بناها الكنعانيون في القرن الـ13 قبل الميلاد.وتعد القدس من أكثر الأماكن قدسية لدى اليهود، وهي المكان الوحيد الذي احتلته قوات الاحتلال الإسرائيلي عام 1967م.ومنذ ذلك الحين، لم تعد المدينة التاريخية في نظر الفلسطينيين مجرد مدينة تراثية
كان يا ما كان في قديم الزمان	كان يا ما كان في قديم الزمان كان في قديم الزمان كان في قديم الزمان كان في قديم الزمان كان في قديم الزمان	كان يا ما كان في قديم الزمان لا بد أنه قد جاء اليوم الذي كنا نعيش فيه في زمن مضى، ولم نكن نعرف أننا نعيش الآن في عصر ذهبي، بل كنا في عهدنا الحاضر، وكنا على الدوام نتشارك حياتنا اليومية من خلال هذه الحياة التي كانت موجودة بيننا

To further demonstrate the practicality and efficiency of our model, we measured the number of Tokens it generates Per Second (TPS) by running it on a laptop with an entry-level GPU. We took three prompts, generated completions from them, and measured the TPS. We then report the average TPS. The laptop used in this experiment had the following specs:

- CPU: AMD Ryzen 7 7735HS
- GPU: NVIDIA GeForce RTX 4050 Laptop GPU
- RAM: 16 GB

We also reran the same experiment running it only the Central Processing Unit (CPU). The results shown in Table 11 demonstrate that it is actually possible to run our model with just a CPU. although the TPS is not high (6.53) it is still impressive that the model would run with an acceptable TPS. Using a laptop with the weakest NVIDIA laptop GPU of this generation (40 series), we get an impressive 38.44 tokens per second. This highlighted the model's ability to perform efficiently even on a laptop. This experiment shows the versatility and accessibility of our model, paving the way for high-quality language processing that can be achieved on commonly available devices.

Table 11. TPS measurements for running the model on 1) An entry-level laptop GPU, and 2) A laptop's CPU.

Experiment	Avg. TPS
1. With GPU	38.44
2. Without GPU	6.53

6. Conclusions

This research paper presents the design, development, and evaluation of the ACLM, a small and efficient language model tailored for MSA. The study addresses the limitations of existing large-scale language models, which often require extensive computational resources, leading to challenges in deployment, especially in resource-constrained environments.

By leveraging high-quality Arabic data, ACLM was built upon the AraGPT2-base model, which has 135 million parameters. Our approach emphasized the importance of high-quality data in enhancing model performance without increasing the model size. Through rigorous pre-training on carefully curated datasets of published books and newspapers, ACLM achieved impressive linguistic capabilities and produced coherent and contextually accurate text.

The evaluation framework designed for this study assessed ACLM's performance across five distinct scenarios: long prompt, open-ended prompt, conversational prompt, storytelling prompt, and information prompt. The results demonstrated that ACLM outperformed AraGPT2- base, showcasing its ability to generate more coherent, fluent, and contextually appropriate text.

In conclusion, ACLM represents a significant advancement in Arabic NLP, offering a compact and efficient solution that bridges the gap between high resource demands of large models and practical needs. This work highlights the potential of leveraging high-quality data to create small yet powerful language models, paving the way for more inclusive and sustainable AI technologies.

Future work can explore several avenues to enhance the capabilities and applications of ACLM. One potential direction is to optimize the model for deployment on edge devices by further reducing its size and improving efficiency through techniques like model pruning and quantization. Enhancing the model's understanding of context and improving its ability to handle complex queries and instructions through advanced fine-tuning methods, including instruction fine-tuning, could also be valuable.

References

[1] Alkaoud M., "A Bilingual Benchmark for Evaluating Large Language Models," *PeerJ Computer Science*, vol. 10, pp. 1-22, 2024. https://pmc.ncbi.nlm.nih.gov/articles/PMC10909174/

[2] Altamimi M. and Alayba A., "ANAD: Arabic News Article Dataset," *Data in Brief*, vol. 50, pp. 109460, 2023. DOI:10.1016/j.dib.2023.109460

[3] Antoun W., Baly F., and Hajj H., "AraGPT2: Pre-Trained Transformer for Arabic Language Generation," *in Proceedings of the 6th Arabic Natural Language Processing Workshop*, Kyiv, pp. 196-207, 2021. https://aclanthology.org/2021.wanlp-1.21/

[4] Biderman D., Dan Biderman., Portes J., Ortiz J., Paul M., Greengard P., Jennings C., King D., Havens S., Chiley V., Frankle J., Blakeney C., and Cunningham J., "LoRA Learns Less and Forgets Less," *arXiv Preprint*, vol. arXiv:2405.09673v2, pp. 1-39, 2024. https://doi.org/10.48550/arXiv.2405.09673

[5] Bourahouat G., Abourezq M., and Daoudi N., "Word Embedding as a Semantic Feature Extraction Technique in Arabic Natural Language Processing: An Overview," *The International Arab Journal of Information Technology*, vol. 21, no. 2, pp. 313-325, 2024. https://doi.org/10.34028/iajit/21/2/13

[6] Brown T., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., and Neelakantan A., et al., "Language Models Are Few-Shot Learners," *in Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, pp. 1877-1901, 2020. https://dl.acm.org/doi/abs/10.5555/3495724.3495883

[7] Bubeck S., Chandrasekaran V., Eldan R., Gehrke J., Horvitz E., Kamar E., Lee P., Lee Y., Li Y., Lundberg S., Nori H., Palangi H., Ribeiro M., and Zhang Y., "Sparks of Artificial General Intelligence: Early Experiments with GPT-4," *arXiv Preprint*, vol. arXiv:2303.12712v5, pp. 1-155, 2023. https://arxiv.org/abs/2303.12712

[8] Cho K., Van Merrienboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., and Bengio Y., "Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation," *in Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Doha, pp. 1724-1734, 2014. https://aclanthology.org/D14-1179.pdf

[9] De Vries A., "The Growing Energy Footprint of Artificial Intelligence," *Joule*, vol. 7, no. 10, pp. 2191-2194, 2023. https://doi.org/10.1016/j.joule.2023.09.004

[10] Donisch L., Schacht S., and Lanquillon C., "Inference Optimizations for Large Language Models: Effects, Challenges, and Practical Considerations," *arXiv Preprint*, vol. arXiv:2408.03130v1, pp. 1-12, 2024. https://arxiv.org/abs/2408.03130

[11] Eldan R. and Li Y., "TinyStories: How Small can Language Models be and Still Speak Coherent English?," *arXiv Preprint*, vol.

arXiv:2305.07759v2, 2023. https://doi.org/10.48550/arXiv.2305.07759

[12] Gomez A., "Introducing Command R+: A Scalable LLM Built for Business, 2024 https://cohere.com/blog/command-r-plus-microsoft-azure, Last Visited, 2024.

[13] Gunasekar S., Zhang Y., Aneja J., Mendes T., and Giorno A., et al., "Textbooks are all you Need," *arXiv Preprint*, vol. arXiv:2306.11644v2, pp. 1-26, 2023. https://arxiv.org/abs/2306.11644

[14] Hochreiter S. and Schmidhuber J., "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997. https://doi.org/10.1162/neco.1997.9.8.1735

[15] Hu E., Shen Y., Wallis P., Allen-Zhu Z., Li Y., Wang S., Wang L., and Chen W., "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv Preprint*, vol. arXiv:2106.09685v2, pp. 1-26, 2021. https://arxiv.org/abs/2106.09685

[16] Jelinek F., Mercer R., Bahl L., and Baker J., "Perplexity-a Measure of the Difficulty of Speech Recognition Tasks," *The Journal of the Acoustical Society of America*, vol. 62, no. S1, pp. S63-S63, 1977. https://doi.org/10.1121/1.2016299

[17] Jordan M., *Advances in Psychology*, Elsevier, 1997. https://doi.org/10.1016/S0166-4115(97)80111-2

[18] Lai V., Ngo N., Pouran Ben Veyseh A., Man H., Dernoncourt F., Bui T., and Nguyen T., "ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning," *in Proceedings of the Findings of the Association for Computational Linguistics: EMNLP*, Singapore, pp. 13171-13189, 2023. https://aclanthology.org/2023.findings-emnlp.878/

[19] Li Y., Li Z., Yang W., and Liu C., "RT-LM: Uncertainty-Aware Resource Management for Real-Time Inference of Language Models," *in Proceedings of the IEEE Real-Time Systems Symposium*, Taipei, pp. 158-171, 2023. https://ieeexplore.ieee.org/document/10405961

[20] Luccioni A., Jernite Y., and Strubell E., "Power Hungry Processing: Watts Driving the Cost of AI Deployment?," *in Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, Rio de Janeiro, pp. 85-99, 2024. https://doi.org/10.1145/3630106.3658542

[21] Luccioni A., Viguier S., and Ligozat A., "Estimating the Carbon Footprint of Bloom, a 176b Parameter Language Model," *The Journal of Machine Learning Research*, vol. 24, no. 1, pp. 11990-12004, 2024. https://dl.acm.org/doi/10.5555/3648699.3648952

[22] Mei T., Zi Y., Cheng X., Gao Z., Wang Q., and Yang H., "Efficiency Optimization of Large-Scale Language Models Based on Deep Learning in Natural Language Processing Tasks," *in Proceedings of the IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering*, Jinzhou, pp. 1231-1237, 2024. https://ieeexplore.ieee.org/document/10729518

[23] Nigst L., Romanov M., Savant S., Seydi M., Verkinderen P., and Hakimi H., OpenITI: A Machine-Readable Corpus of Islamicate Texts, 2023, Last Visited, 2024. https://zenodo.org/records/10007820

[24] OpenAI, GPT-4, Technical Report, 2023. https://cdn.openai.com/papers/gpt-4.pdf

[25] Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C., Mishkin P., and Zhang C., et al., "Training Language Models to Follow Instructions with Human Feedback," *in Proceedings of the 36th Conference on Neural Information Processing Systems*, New Orleans, pp. 27730-2774, 2022. https://dl.acm.org/doi/10.5555/3600270.3602281

[26] Patterson D., Gonzalez J., Le Q., Liang C., Munguia L., Rothchild D., So D., Texier M., and Dean J., "Carbon Emissions and Large Neural Network Training," *arXiv Preprint*, vol. arXiv:2104.10350, pp. 1-22, 2021. https://arxiv.org/abs/2104.10350

[27] Perrine P., "Inaccessible Neural Language Models Could Reinvigorate Linguistic Nativism," *arXiv Preprint*, vol. arXiv:2301.05272v1, pp. 1-6, 2023. https://arxiv.org/abs/2301.05272

[28] Radford A., Narasimhan K., Salimans T., and Sutskever I., Improving Language Understanding by Generative Pre-Training, Technical Report, 2018. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

[29] Radford A., Wu J., Child R., Luan D., Amodei D., and Sutskever I., "Language Models are Unsupervised Multitask Learners," *OpenAI blog*, vol. 1, no. 8, pp. 1-24, 2019. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[30] Rumelhart D., Hinton G., and Williams R., *Learning Internal Representations by Error Propagation*, MIT Press, 1985. https://stanford.edu/~jlmcc/papers/PDP/Volume%201/Chap8_PDP86.pdf

[31] Sathish V., Lin H., Kamath A., and Nyayachavadi A., "LLeMpower: Understanding Disparities in the Control and Access of Large Language Models," *arXiv Preprint*, vol. arXiv:2404.09356v1, pp. 1-11, 2024. https://arxiv.org/abs/2404.09356

[32] Selvan R., Pepin B., Igel C., Samuel G., and Dam E., "PePR: Performance Per Resource Unit as a Metric to Promote Small-Scale Deep Learning in Medical Image Analysis," *in Proceedings of the*

6th Northern Lights Deep Learning Conference, Tromso, pp. 1-10, 2025. https://arxiv.org/abs/2403.12562

[33] Sengupta N., Sahu S., Jia B., Katipomu S., Li H., Koto F., Marshall W., and Gosal G., et al., "Jais and Jais Chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models," *arXiv Preprint*, vol. arXiv:2308.16149v2, pp. 1-5, 2023. https://doi.org/10.48550/arXiv.2308.16149

[34] Shuttleworth R., Andreas J., Torralba A., and Sharma P., "LoRA vs Full Fine-Tuning: An Illusion of Equivalence," *arXiv Preprint*, vol. arXiv:2410.21228v1, pp. 1-21, 2024. https://doi.org/10.48550/arXiv.2410.21228

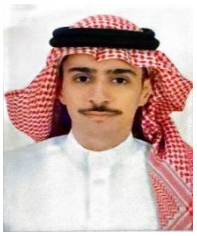
[35] Suarez P., Romary L., and Sagot B., "A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages," *in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Virtual, pp. 1703-1714, 2020. https://aclanthology.org/2020.acl-main.156/

[36] Team G., Anil R., Borgeaud S., Wu Y., and Alayrac J., et al., "Gemini: A Family of Highly Capable Multimodal Models," *arXiv Preprint*, vol. arXiv:2312.11805v4, pp. 1-90, 2024. https://doi.org/10.48550/arXiv.2312.11805

[37] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Kaiser L., and Polosukhin I., "Attention is all you Need," *in Proceedings of the 31st Conference on Neural Information Processing Systems*, Long Beach, 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[38] Zheng Y., Chen Y., Qian B., Shi X., Shu Y., and Chen J., "A Review on Edge Large Language Models: Design, Execution, and Applications," *ACM Computing Surveys*, vol. 57, no. 82024, pp. 1-35, 2025. https://arxiv.org/abs/2410.11845

Mohamed Alkaoud received the B.S. degree in Computer Science from King Saud University, Riyadh, Saudi Arabia, in 2014; the M.S. degree in Computer Science from the University of California Davis, Davis, CA, USA in 2018; and the Ph.D. degree in Computer Science from the University of California Davis, Davis, CA, USA in 2021. Since 2021, he has been an Assistant Professor with the Computer Science Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. His research interest includes Artificial Intelligence, Machine Learning, Natural Language Processing, Computational Linguistics, Computer Vision, Digital Humanities, and Education.



Muteb Alsaqoub received the B.S. degree in Computer Science from King Saud University, Riyadh, Saudi Arabia, in 2024. His research interest includes Artificial Intelligence and Machine Learning.

Ibrahim Aljodhi received the B.S. degree in Computer Science from King Saud University, Riyadh, Saudi Arabia, in 2024. His research interest includes Artificial Intelligence and Machine Learning.

Abdulrhman Alqadibi received the B.S. degree in Computer Science from King Saud University, Riyadh, Saudi Arabia, in 2024. His research interest includes Artificial Intelligence and Machine Learning.

Omar Altammami received the B.S. degree in Computer Science from King Saud University, Riyadh, Saudi Arabia, in 2024. His research interest includes Artificial Intelligence and Machine Learning.

Appendix I. A list of the generated examples that we used to calculate perplexity.

أعلنت وزارة الصحة اليوم عن تسجيل 500 حالة جديدة من الإصابة بفيروس كورونا، مع تعافي 400 حالة وخروجهم من المستشفيات. كما حثت الوزارة المواطنين على الالتزام بالإجراءات الوقائية وارتداء الكمامات في الأماكن العامة للحفاظ على سلامة الجميع".

"إفي تلك الليلة الهادئة، كانت النجوم تتلألأ في السماء، والأشجار تتمايل مع النسيم العليل، مما خلق جوًا من السكينة والهدوء. جلس الكاتب في شرفته يحتسي كوبًا من الشاي، وبدأ يفكر في الأحداث التي مرت عليه خلال النهار".

"تأسست مدينة بغداد في القرن الثامن الميلادي على يد الخليفة العباسي المنصور، وأصبحت مركزًا للعلم والثقافة والتجارة. شهدت المدينة تطورًا كبيرًا خلال العصور الوسطى، حيث ازدهرت المكتبات والمدارس وجذب العلماء والمفكرين من مختلف أنحاء العالم".

"تعتبر عملية البناء الضوئي من أهم العمليات الحيوية في النباتات، حيث تقوم الأوراق بامتصاص ضوء الشمس وتحويله إلى طاقة كيميائية من خلال تفاعل كيميائي معقد. يتم تخزين هذه الطاقة في جزيئات الجلوكوز التي تستخدمها النباتات للنمو والتطور".

"يعمل الحاسوب على معالجة البيانات الرقمية باستخدام وحدات المعالجة المركزية والذاكرة العشوائية لتشغيل البرامج المختلفة. يتم تخزين البيانات في وحدات التخزين مثل الأقراص الصلبة ومحركات الأقراص الصلبة، بينما يتم عرض النتائج على الشاشة".

"قال أحمد: 'متى ستزورنا يا علي؟'، فأجاب علي: 'سأحاول القدوم في نهاية الأسبوع إن شاء الله. لقد اشتقت لكم جميعًا وأود قضاء بعض الوقت معكم.' ابتسم أحمد وقال: 'نحن أيضًا ننتظرك بفارغ الصبر'."

"في القلب شوق وفي العين دمعة، يحن الفؤاد إلى أيامنا الباسمة، فما أطيب الذكرى وما أعذبها. تسري الليالي ونحن نتذكر، تلك اللحظات التي عشناها سويًا، فنرى في أحلامنا شمس الصباح".
من جد وجد، ومن زرع حصد. إن الاجتهاد والعمل الدؤوب هما مفتاح النجاح في الحياة. لا يمكن لأي شخص أن يصل إلى أهدافه دون أن يبذل جهدًا ويواجه التحديات بصبر وإصرار".

"لفتح الملف، اضغط على الزر الأيمن للفأرة واختر 'فتح باستخدام'، ثم حدد البرنامج المناسب من القائمة المنسدلة. إذا لم يكن البرنامج موجودًا، يمكنك النقر على 'اختيار برنامج آخر' وتصفح جهاز الكمبيوتر للعثور على البرنامج المطلوب".

"في أحد أيام الصيف الحارة، قررنا الذهاب إلى الشاطئ للتمتع بنسمات البحر الباردة. كانت الأمواج تتلاطم بلطف على الرمال الذهبية، والأطفال يلعبون ويبنون قلاعًا من الرمل. جلسنا تحت مظلة كبيرة نستمتع بالمشهد الجميل ونشرب عصير الليمون الطازج".