Adapting Falcon3-7B Language Model for Arabic: Methods, Challenges, and Outcomes

Basma El Amel Boussaha, Mohammed Alyafeai, Ahmed Alzubaidi
Leen AlQadi, Shaikha Alsuwaidi, Hakim Hacid
Technology Innovation Institute, Abu Dhabi, UAE
basma.boussaha@tii.ae

Abstract

Under-represented languages suffer from a lack of data, and as a result, there are few LLMs that support them. Extending an existing LLM to a new language is a practical option for startups, university labs, and organizations with limited budgets. This process involves several steps. In this paper, we describe how we adapted the Falcon3-7B model to Arabic, covering everything from data collection and training to evaluation. Falcon-Arabic was trained exclusively on native data to better capture the cultural and linguistic aspects of the language. Our evaluations show that Falcon-Arabic achieves state-of-the-art results on a range of Arabic benchmarks.

1 Introduction

Despite the rapid advancement of Large Language Models (LLMs) across many languages, Arabic remains comparatively underserved. Most state-of-the-art multilingual models such as mBERT (Devlin et al., 2018), XLM-R (Conneau et al., 2020), and BLOOM (Workshop et al., 2023) have enabled substantial cross-lingual transfer, but Arabic presents unique linguistic features that introduce specific challenges requiring tailored modeling strategies (Abdul-Mageed et al., 2021). The morphological richness of Arabic, with its complex root-and-pattern system, extensive inflection, and agglutinative properties, poses significant challenges for tokenization and representation (Habash et al., 2013).

The scarcity of high-quality Arabic data exacerbates these challenges, compelling researchers to rely heavily on machine translation to augment dataset sizes (Koehn, 2005). However, this approach often results in models misaligned with Arabic cultural contexts and linguistic nuances (Bender et al., 2021). As demonstrated in Figure 1, when asked to explain an Arabic proverb, several prominent LLMs failed to grasp the underlying cultural meaning, instead providing literal word-by-word translations. This limitation underscores the need for culturally-aware Arabic language models that can capture the depth and subtlety of Arabic expression.

Prompt and Responses from Multiple LLMs

Prompt:
ما معنى "يدّك في الموية الباردة" في السياق الاجتماعي؟

Qwen2.5-7B-Instruct:
العبارة "يدّك في الموية الباردة" هي عبارة شائعة في بعض الثقافات العربية، وتستخدم عادة في السياق الاجتماعي للإشارة إلى حالة من الارتباك أو الفوضى أو عدم الراحة.

Llama-3.1-8B-Instruct:
في السياق الاجتماعي، يعني "يدّك في الموية الباردة" أنك محتاجون للمساعدة، خاصة إذا كنت في وضع صعب أو تعاني من صعوبات مالية.

Falcon3-7B-Arabic:
يدّك في الموية الباردة" هي عبارة شعبية تستخدم في السياق الاجتماعي للإشارة إلى شخص يعيش في حالة من الراحة أو الرفاهية، حيث لا يواجه أي تحديات أو صعوبات.

Figure 1: Example of LLM responses to an Arabic prompt with cultural and dialectal elements; Falcon-Arabic correctly captured the cultural nuance. The English translation of the example is provided in the Appendix (Figure 4).

Arabic LLMs can be categorized into three main model families: native models trained from scratch, multilingual models with Arabic support, and models adapted from existing multilingual LLMs (Mashaabi et al., 2024). Training Arabic models from scratch requires trillions of Arabic tokens, which are difficult to collect, along with substantial computational infrastructure (Kaplan et al., 2020).

Analysis of the Open Arabic LLM Leaderboard (El Filali et al., 2025) reveals that multilingual models such as Qwen (Qwen et al., 2025) and LLaMA (Grattafiori et al., 2024), as well as adapted models like AceGPT (Huang et al., 2024) consis-

Proceedings of The Third Arabic Natural Language Processing Conference, pages 1–15

tently rank among the top performers. Adapting existing LLMs to new languages requires significantly less data and computational resources compared to training from scratch (Wang et al., 2025). The foundation model already possesses general knowledge, reasoning capabilities, and common sense, making it a matter of aligning new language tokens with existing representations rather than learning from scratch. This approach has proven successful in recent continual pretraining studies (Gupta et al., 2023).

Motivated by these findings, we adapt Falcon3-7B (Team, 2024a) to Arabic. The adaptation process presents unique challenges since Falcon3-7B's tokenizer lacks Arabic support, requiring careful vocabulary extension and embedding initialization (Minixhofer et al., 2022). In this work, we detail the complete adaptation pipeline, from data collection and tokenizer extension to model layer adaptation, multi-stage training, and post-training procedures. We document the challenges encountered and key insights gained, contributing valuable knowledge to the community for future language adaptation efforts.

What distinguishes Falcon-Arabic is our exclusive use of native Arabic datasets without machine translation, encompassing diverse content including dialects, poetry, literature, and contemporary texts, all authentically Arabic. Through training on only 600B tokens, we achieve a model that outperforms LLMs two times its size while maintaining strong cultural relevance and linguistic authenticity for the Arabic-speaking community. Our approach demonstrates that targeted adaptation with high-quality, culturally-authentic data can achieve superior performance compared to larger, more resource-intensive alternatives (Touvron et al., 2023).

2 Related Work

The interest in building Arabic Language Models has emerged with multiple initiatives spanning various sizes from a few million parameters to billions (Mashaabi et al., 2024). Models like AraBERT (Abdul-Mageed et al., 2021) and AraGPT2 (Antoun et al., 2021) were among the first transformer-based Arabic LLMs with millions of parameters (Vaswani et al., 2017). AraBERT introduced comprehensive pre-training on Arabic text with careful preprocessing to handle the language's morphological complexity and diacritization variations. AraGPT2 demonstrated the effectiveness of generative pre-training for Arabic text generation, establishing foundational benchmarks for subsequent Arabic language models. Subsequently, increasing the number of parameters in these models showed promising performance improvements, leading to more ambitious initiatives toward building Arabic Large Language Models. Arabic LLMs can be categorized into three main categories based on how Arabic was incorporated into the training data.

Native Arabic Models are trained on Arabic from scratch or with Arabic as a primary language. JAIS (Sengupta et al., 2023) represents a prominent example of this category, being trained on a balanced mix of Arabic, English, and code to achieve strong performance across Arabic dialects while maintaining multilingual capabilities. The model was specifically designed to handle the nuances of Arabic script and cultural context. Other small Arabic LLMs trained from scratch include ArabianGPT (Koubaa et al., 2024) and AraGPT (Antoun et al., 2021).

Multilingual Foundation Models constitute the second category, typically featuring strong English support as a primary language while demonstrating competitive results across other languages, including Arabic. The LLaMA family of models (Grattafiori et al., 2024) supports a wide range of languages through extensive multilingual pre-training, showing robust cross-lingual transfer capabilities. Qwen2.5 (Qwen et al., 2025) and Qwen3 (Yang et al., 2025) have demonstrated strong multilingual performance with particular attention to maintaining quality across diverse writing systems. The Gemma (Team et al., 2024a) and Gemma 2 (Team et al., 2024b) models have shown promising results in multilingual settings while maintaining computational efficiency through architectural innovations.

Adapted Arabic Models represent the third category, comprising models that were fine-tuned or adapted from multilingual LLMs to enhance Arabic-specific performance. Some models were adapted from LLama such as AceGPT (Huang et al., 2024), JAIS adapted family (Sengupta et al., 2023), Yehia (Navid-AI, 2025). While others were adapted from Gemma such as SILMA (Team, 2024b) and Fanar (Team et al., 2025). Each model targets specific improvements: AceGPT focuses on cultural adaptation, ALLAM emphasizes Arabic linguistic features, while Yehia and Fanar enhance regional dialect support. The JAIS adapted family and SILMA demonstrate continued progress in instruction fol-

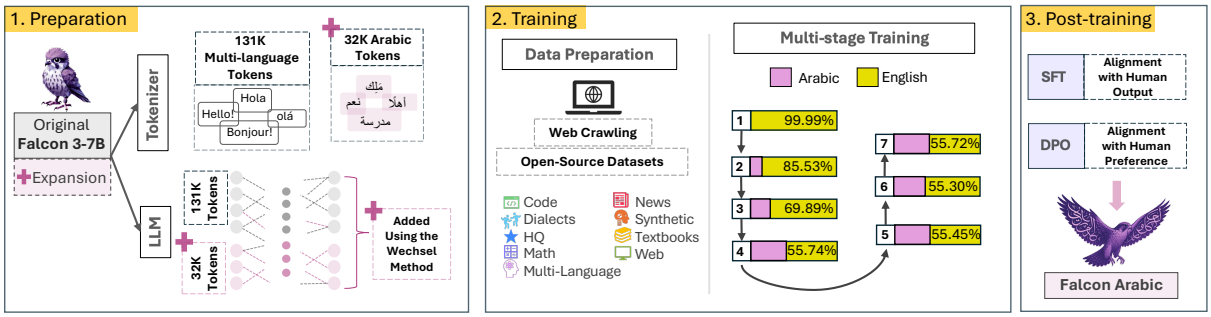

Figure 2: Schematic View of the adaptation of Falcon-3 7B Language Model for Arabic.

lowing and conversational capabilities for Arabic.

While these LLMs demonstrate competitive performance across multiple benchmarks, multilingual models such as Gemma, LLaMA, and Qwen often lack culturally-centric data related to Arabic and the Arab region, heavily relying on machine translation which may introduce cultural and linguistic biases. To address these limitations, we built Falcon-Arabic by training exclusively on native Arabic data and carefully designing training stages to smoothly integrate culturally and linguistically relevant content into Falcon3-7B, ensuring authentic representation of Arabic language nuances and cultural contexts.

3 Datasets

Addressing the significant challenge of limited Arabic data availability, we prepared a comprehensive multilingual corpus totaling approximately 600 BT tokens, with Arabic comprising 40% and English 60% of the dataset.

3.1 Arabic Datasets

Recognizing the crucial gap in Arabic datasets for LLMs, particularly in dialectal diversity and STEM-related content, we developed a comprehensive Arabic corpus addressing these limitations. The dataset covers diverse Arabic dialects including Levantine (الشام), Maghrebi (Darija), Egyptian and Gulf Arabic, ensuring broad linguistic representation across multiple textual domains: web documents (Penedo et al., 2025), educational materials, news sources, and mathematical content.

For low-resource dialects, we leveraged recent Moroccan Darija adaptations (Shang et al., 2024) and specialized OCR datasets from Arabic-Nougat (Rashad, 2024). Additionally, we actively crawled and curated new data from educational books, web documents, and news articles. A distinctive feature is our focus on grammatical details, including annotations for grammatical structures (إعراب) and various linguistic forms. Critically, we avoided machine-translated content, instead selecting authentic Arabic language data from different historical periods to maintain performance quality.

3.2 English Datasets

Acknowledging the importance of maintaining robust English performance alongside Arabic proficiency, a comprehensive English corpus comprising approximately 60% of the total dataset was curated. This dataset covers diverse textual domains including extensive collections from textbooks, web sources (Penedo et al., 2025; Lozhkov et al., 2024a; Ben Allal et al., 2024), synthetic data, code repositories (Lozhkov et al., 2024b), high-quality documents, mathematical texts (Han et al., 2024), and multilingual content. While the dataset was not fully expanded from prior training data, it strategically combines previously effective resources with newly introduced data with the aim of enhancing performance across key benchmarks.

To ensure balanced representation and address domain gaps, we further supplemented the dataset with synthetically generated data and additional crawled resources, including recent news articles and educational materials.

4 Approach

In this section we detail the steps that we followed to adapt Falcon3-7B to Arabic.

4.1 Tokenizer Extension

The original Falcon3-7B tokenizer primarily covers English, French, Spanish, and Portuguese, making it inefficient for Arabic text due to over-segmentation. To address this, we extend Falcon's vocabulary by adding 32,768 Arabic tokens to the original 131,072 tokens, resulting in a total vocabulary of 163,840 tokens which remains a reasonable tokenizer size for a 7B LLM.

Model	Fertility Score	Vocabulary Size
Falcon-Arabic	**2.17**	163,840
Gemma-3-4B	2.18	262,208
Llama-3.1-8B	2.43	128,256
Qwen2.5-7B	2.55	152,064
Falcon3-7B-Base	4.54	131,072

Table 1: Fertility scores of different LLMs. Lower is better.

We trained a BPE tokenizer on the Arabic subset of FineWeb2 (Penedo et al., 2025) using the same configuration as Falcon3-7B, then merged the vocabularies while preserving original token mappings. We evaluated the effectiveness by computing fertility scores[1] (average tokens per word) for both tokenizers on Arabic text, with results shown in Table 1.

This extension provides reduced training and inference costs, lower latency, and support for longer context windows (Gosal et al., 2024). Models with low fertility tokenizers demonstrate improved performance on downstream tasks (Ahuja et al., 2023).

4.2 Layers Extension

After training a new Arabic tokenizer and extending the Falcon3-7B tokenizer, we needed to incorporate the newly added tokens into both the input embedding layer and the output layer (lm head). The critical challenge lies in properly initializing the embeddings associated with these new tokens to maintain model performance and training stability. Multiple initialization approaches exist for newly added token embeddings, including zero, random, and averaging existing embeddings (de Vries and Nissim, 2021; Marchisio et al., 2023; Zhao et al., 2024). However, according to Gosal et al. (2024), these conventional approaches may lead to degraded performance as they deviate from the initial distribution of pre-trained word embeddings.

To address these limitations, we apply the Wechsel approach (Minixhofer et al., 2022) to initialize the newly added token embeddings. This method leverages cross-lingual alignment and subword-level correspondences to create more informed initializations that preserve the semantic structure of the original embedding space.

The Wechsel method proceeds through the following key steps: (1) tokenize bilingual dictionary words into subwords using both tokenizers, (2) compute subword embeddings e_{sw} using fastText (Bojanowski et al., 2016) as the sum of n-gram embeddings $N(sw)$ as in Equation 1 (3) align subword embeddings across languages using Orthogonal Procrustes alignment (Schönemann, 1966; Artetxe et al., 2016), (4) initialize new token embeddings e_{sw_t} as weighted averages of source embeddings using cosine similarity as weights Equation 2, and (5) copy non-embedding parameters from the source model.

$$e_{sw} = \sum_{ng \in N(sw)} e_{ng} \tag{1}$$

$$e_{sw_t} = \frac{\sum_{sw_s \in N(sw_t)} \text{sim}(sw_s, sw_t) \cdot e_{sw_s}}{\sum_{sw_s \in N(sw_t)} \text{sim}(sw_s, sw_t)} \tag{2}$$

where e_{sw} is the embedding of subword sw, $N(sw)$ is the set of n-grams occurring in the subword, e_{ng} is the embedding of n-gram ng, e_{sw_t} is the target subword embedding, $N(sw_t)$ represents the set of neighboring subwords in the source language, and $\text{sim}(sw_s, sw_t)$ denotes the cosine similarity between source and target subwords.

This approach ensures that newly added Arabic tokens receive semantically meaningful initializations that are consistent with the pre-trained embedding space, thereby facilitating more efficient adaptation and improved performance on Arabic language tasks.

4.3 Continuous Pretraining

With the tokenizer extended and the input and output embedding layers properly initialized, the model is ready for continuous pretraining. We designed a multi-stage training approach consisting of four stages to carefully control the data mixture, sequence length, and ratio between Arabic and English content. Table 2 summarizes the percentage of each data source per stage and the corresponding sequence lengths used.

The first stage represents the longest training phase with the shortest sequence length, as most datasets contain relatively short sequences. This approach is more computationally efficient and requires fewer resources while maintaining training stability. Stages 2 and 3 are designed to extend the context length capabilities of Falcon-Arabic to 16K and 32K tokens, respectively. We conclude the pretraining with a decay stage to stabilize convergence

[1]Dataset used from https://huggingface.co/spaces/wissamantoun/arabic-tokenizers-leaderboard

Stage	Seq length	Textbooks	Code	HQ	Math	Synthetic	Dialects	News	Multilang	Web
1.1	8K	11.74	13.85	14.69	2.94	15.67	0.00	0.00	0.58	40.53
1.2	8K	0.69	3.69	29.13	15.55	0.00	0.00	7.54	0.66	42.74
1.3	8K	1.72	5.23	9.97	8.20	15.31	0.11	0.46	0.83	58.17
1.4	8K	11.65	11.93	3.36	12.08	5.54	0.06	1.77	0.46	53.15
2	16K	31.59	9.71	13.51	4.74	5.89	0.13	3.38	1.27	29.78
3	32K	38.58	2.71	3.23	16.08	15.70	0.17	0.29	0.38	22.86
Decay	32K	18.89	1.61	4.85	30.25	12.56	0.16	22.08	0.20	9.40

Table 2: Training stages of Falcon-Arabic.

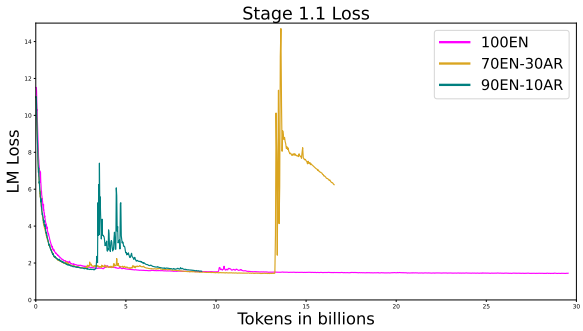

Figure 3: Training loss.

Stage	Arabic	Other Languages	Total
1.1	0.00	29.55	29.55
1.2	5.54	32.74	38.28
1.3	13.83	32.09	45.92
1.4	78.30	98.61	176.91
2	38.61	48.06	86.67
3	28.62	34.00	62.62
Decay	57.39	69.34	126.73

Table 3: Distribution of Arabic and other Languages in Billion Tokens (BT) at each training stage.

and prevent overfitting as the model approaches optimal performance. This final stage employs learning rate decay to enable smaller, more precise parameter updates, allowing the model to fine-tune its internal representations without overshooting minima or introducing instability.

Since Falcon3-7B was not originally exposed to Arabic data during its pretraining, introducing Arabic datasets requires careful consideration to avoid catastrophic forgetting and important distribution shifts (Çagatay Yildiz et al., 2024). We conducted multiple experiments for the first training stage to identify the optimal proportion of Arabic data while monitoring training loss stability. As shown in Figure 3, initiating training with 30% Arabic data resulted in significant training instability, evidenced by substantial loss spikes. Reducing the Arabic percentage to 10% improved stability but still exhibited spikes, suggesting the model required additional English data for stabilization.

To address this challenge, we implemented a short stabilization stage of 29BT consisting of 100% English data, allowing the model to adjust to the newly added tokens gradually. Following this adjustment period, we employed three additional substages where we progressively increased the Arabic data percentage to achieve 45%, which we maintained across the remaining training stages as detailed in Table 3. This gradual approach ensures smooth integration of Arabic content while preserving the model's existing capabilities and maintaining training stability throughout the continuous pretraining process.

Checkpoints of each training stage were evaluated separately on Arabic and English benchmarks to monitor the evolution of the training process and detect early signs of catastrophic forgetting or bad data. More details are provided in Section 6. Falcon-Arabic was trained on 566B tokens using 32 H100 nodes (8k toks/GPU/s), corresponding to 3.4 days of wall-clock training and 2.5×10^{22} FLOPs.

5 Post-training

At this stage, we trained our base model to engage in conversations and follow user instructions. We employed Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) to obtain an instruct version of the Falcon-Arabic.

5.1 Supervised Fine-tuning (SFT)

We started by performing SFT, to make the model capable of conducting conversations, making it capable of following instructions and answering questions. In line with continuous pretraining, both arabic and english data were fed to the model at this stage, ensuring that it could chat in both languages. Next, we discuss the SFT datasets used.

Hyperparameter	SFT	DPO
Batch Size	512	128
Epochs	3	1
KL Penalty (B)	-	5
Optimizer		
Optimizer	AdamW	AdamW
B_1	0.9	0.9
B_2	0.99	0.99
ϵ	1×10^{-15}	1×10^{-8}
Learning Rate		
Decay Type	linear	linear
Max lr	1×10^{-6}	1×10^{-7}
Min lr	6×10^{-8}	-
Warmup	3%	5%

Table 4: SFT/DPO Optimal Hyperparameters.

5.1.1 SFT Data

A wide range of datasets was used in terms of category and task type, curated from public datasets and curated sources. Examples of Arabic public datasets used are Aya (Singh et al., 2024), WikiReading (Albilali et al., 2022), and Bactrian-X (Li et al., 2023). Furthermore, an in-house synthetic SFT dataset was created that expands the list of covered topics and allows the model to handle multi-turn chats. To ensure the chat model remains multilingual, the publicly available tulu-3 dataset (Lambert et al., 2025) was used. The resulting SFT dataset comprised 4.3 million samples, with a language distribution of approximately 55% Arabic and 45% English.

5.1.2 SFT Recipe

An extensive search was performed on the SFT hyperparameters to select the optimal set of hyperparameters values that maximizes the model performance. Table 4 shows the optimal SFT configuration we used during the SFT stage.

5.2 Direct Preference Optimization (DPO)

In the second stage of the post-training, we leveraged DPO (Rafailov et al., 2024) to align the model with generating more human-like responses. DPO offered an offline training approach, where the need for a reward model is alleviated. Typically, DPO is applied to binary preference data, where each sample has a pair of accepted and rejected responses for the same prompt. The ultimate objective of this stage is to steer the model to become aligned with human preference while maintaining its knowledge and capabilities from the SFT stage. Several public binary preference datasets were utilized, such as argilla[2], orca [3], and tulu-3 (Lambert et al., 2025). The optimal hyperparameters found for DPO is shown in Table 4.

6 Evaluation

To assess the performance of Falcon-Arabic, the pre-trained and intruct models were evaluated using several benchmarks[4]. The backend of our evaluation setup leveraged lighteval (Habib et al., 2023) and lm-eval (Gao et al., 2024), which are both established evaluation tools within the NLP community. We compared Falcon-Arabic against several open-source SOTA models ($< 14B$), chosen based on the OALL (El Filali et al., 2025). The benchmarks used in this work are discussed in the following subsections.

6.1 Benchmarks

General benchmarks AlGhafa (Almazrouei et al., 2023) is an Arabic benchmark that targets the evaluation of tasks that include comprehension, sentiment analysis, and question-answering. Only the native Arabic datasets were used. ArabicMMLU is a native Arabic benchmark, which includes 40 tasks and nearly 15k MCQs (Koto et al., 2024). ArbMMLU-HT is a human translated version of the original English MMLU dataset containing 57 tasks. Subjects covered in ArabicMMLU and ArabicMMLU-HT span various topics such as history and social science, which are of varying complexity (Sengupta et al., 2023). Exams (Hardalov et al., 2020) is a benchmark of questions that targets high school level of difficulty, and only the Arabic samples were used. MadinahQA (Koto et al., 2024) is a benchmark with 983 QA pairs that focuses generally on the syntax and grammar of the Arabic language.

Reasoning To access the reasoning capabilities of our model, we integrated the publicly available dataset, called Arabic-GSM8K [5] with lighteval, which is a translation of the GSM8K (Cobbe et al., 2021a).

RAG ALRAGE (El Filali et al., 2025) is a benchmark composed of 2.1k QA pairs that were generated from 40 Arabic books. ALRAGE is intended for the evaluation of LLMs' retrieval-augmented generation (RAG) capabilities in Arabic. The tasks

[2] 2A2I/argilla-dpo-mix-7k-arabic

[3] multilingual/orca_dpo_pairs

[4] chat-template was used for Instruct models

[5] https://huggingface.co/datasets/Omartificial-Intelligence-Space/Arabic-GSM8K

Model	Size	ALGhafa	ArabicMMLU	EXAMS	MadinahQA	AraTrust	ALRAGE	ArbMMLU-HT	Avg
Qwen2.5	7B	**72.17**	61.42	49.16	**51.13**	77.56	64.83	51.67	61.13
jais-adapted	7B	32.92	27.33	26.44	24.84	33.91	41.43	27.4	30.61
	13B	40.62	36.97	34.26	29.04	61.18	62.53	33.12	42.53
AceGPT-v2	8B	46.32	50.41	43.58	40.81	69.25	57.76	35.62	49.12
AceGPT	13B	48.23	41.38	36.87	35.37	56.51	**79.96**	32.12	47.21
Llama-3.1	8B	64.34	52.28	40.04	43.08	71.98	47.08	42.67	51.64
Falcon3-7B-Base	7B	37.89	31.81	24.77	24.87	49.89	60.23	25.88	36.48
Falcon-Arabic	7B	67.17	**64.85**	**52.89**	48.79	**85.36**	63.71	**55.25**	**62.57**

Table 5: Falcon-Arabic compared to the best open source SOTA Models. **Bold** indicates the best score in each column; underline indicates the second best.

in this benchmark include questions and target answers, and candidate context, where outputs are judged by `Qwen2.5-72B-Instruct`.

Truthfulness AraTrust (Alghamdi et al., 2024) is a benchmark with 522 human written MCQs, with the aim of assessing the safety and truthfulness of a model.

Dialect and Culture ArabCulture (Sadallah et al., 2025) was used to assess arab cultural understanding and awareness with questions spanning countries in the Gulf, Levant, North Africa, and the Nile valley. AraDiCE (Mousi et al., 2024) is benchmark composed of 45k samples that includes the dialects translation of major benchmarks, which are ArabicMMLU, boolQ, truthfulqa, piqa, openbookqa, and winogrande in both the Egyptian and the Levantine dialect. Furthermore, the benchmark includes a range of cultural questions related to several Arab countries. For AraDiCE, we report three scores which are Aradice-CULT, Aradice-LEV and Aradice-EGY that corresponds to the mean scores obtained in cultural questions, Levantine questions, and Egyptian samples, respectively.

English benchmarks Considering that Falcon-Arabic was trained to be a multilingual model, its capabilities were evaluated on english tasks too. Therefore, Falcon-Arabic was benchmarked on the open source LLM leaderboard v1 and v2 tasks, which are GSM8K (Cobbe et al., 2021b), HellaSwag (Zellers et al., 2019), ARC Challenge (Clark et al., 2018), Winogrande (Sakaguchi et al., 2021), TruthfulQA (Lin et al., 2022), MMLU (Hendrycks et al., 2021a), IFEval (Zhou et al., 2023), GPQA (Rein et al., 2023), MMLU-pro (Wang et al., 2024), MATH (Hendrycks et al., 2021b), BBH (Suzgun et al., 2022), and MUSR (Sprague et al., 2024).

The evaluation metric used with most of the mentioned benchmarks is normalized accuracy, with the exception of ALRAGE and Arabic-GSM8K. For ALRAGE, an LLM judge was used specifically `Qwen2.5-72B-Instruct`, whereas *exact match* was used for Arabic-GSM8K.

6.2 Results and Discussion

In this section, we discuss the evaluation results of Falcon-Arabic and other SOTA models on general Arabic, reasoning, cultural and English benchmarks.

6.2.1 Arabic General Benchmarks

Table 5 presents the scores of the `Falcon-Arabic` model against SOTA models. From the results, it is evident that `Falcon-Arabic` significantly outperforms the SOTA models in ArabicMMLU, ArbMMLU-HT, and EXAMS. This indicates that our base model excels in general knowledge and STEM subjects. Similar observations can be made in AraTust, which suggests that `Falcon-Arabic` is performing the best in terms of safety. Looking at the Alghafa and MadinahQA benchmarks, our model came second to `Qwen2.5-7B`. Furthermore, in terms of RAG capabilities, our model ranked third, with clear superiority to `AceGPT-13B`. By viewing the average column, it can be deduced that `Falcon-Arabic` is superior to all competitors, as manifested by the highest average score of 62.57.

Next, the evaluation of the instruct models' scores are depicted in Table 6. In the general knowledge and STEM benchmarks, `Falcon-Arabic-Instruct` obtained the highest scores in ArabicMMLU and ArbMMLU-HT, and ranked second EXAMS benchmark. Looking at MadinahQA, it can be inferred that `Falcon-Arabic-Instruct` model excelled in grammar tasks, as it achieved the highest score. Despite not performing the best with AraTrust, our instruct model is still on par with the best instruct models, where `Yehia-7B-preview` scored the highest.

The same observation can be made with Alghafa, where our instruct model is comparable with the best performing models, namely `c4ai-command-r7b-arabic`. To compare the overall performances, the average score indicates that the `Falcon-Arabic-Instruct` is superior to all other SOTA models of similar scale ($< 14B$). By com-

Model	Size	ALGhafa	ArabicMMLU	EXAMS	MadinahQA	AraTrust	ALRAGE	ArbMMLU-HT	Avg
Qwen2.5-Instruct	7B	65.6	52.25	39.66	62.73	80.68	77.37	40.33	59.8
Jais-adapted-chat	7B	63.38	49.9	47.71	34.79	66.02	63.6	37.97	51.05
	13B	67.28	54.23	47.3	44.2	79.68	68.41	45.45	58.08
AceGPT-v2	8B	73.48	61.32	49.72	55.89	74.19	70.94	50.89	62.35
AceGPT	13B	59.18	49.84	40.97	33.08	65.7	79.75	39.31	52.55
Llama-3.1-Instruct	8B	70.91	53.58	50.28	39.72	75.57	49.89	47.94	55.41
c4ai-command-r7b-arabic	7B	**74.84**	59.34	**64.99**	63.84	80.47	75.9	50.14	67.07
aya-expanse-8b	8B	66.71	57.55	45.44	48.74	82.54	75.78	49.22	60.85
ALLaM-Instruct-preview	7B	69.49	64.9	51.58	54.24	86.93	76.81	52.81	65.25
Yehia-preview	7B	70.81	64.9	52.14	54.37	**87.49**	76.64	53.4	65.68
SILMA-Instruct-v1.0	9B	33.99	62.16	51.4	52.48	82.83	**80.39**	40.32	57.64
Falcon3-Instruct	7B	55.75	41.2	29.42	34.4	57.85	43.21	33.59	42.3
Falcon-Arabic-Instruct	7B	72.37	**68.27**	53.45	**73.63**	82.62	72.26	**55.47**	**68.3**

Table 6: Falcon-Arabic-Instruct compared to the best open source SOTA instruct models on OALL benchmark. **Bold** indicates the best score in each column; underline indicates the second best.

paring Tables 5 and 6, it can be concluded that `Falcon-Arabic-Instruct` showed an improvement over `Falcon-Arabic` in all benchmarks, except with the AraTrust benchmark.

6.2.2 Cultural and Reasoning Benchmarks

Table 7, where scores on cultural knowledge and reasoning benchmarks are presented. Looking at Arabic-GSM8K, our model obtained 54.89 Qwen2.5-Instruct scoring the highest in the range of 62. The columns ArabCulture and Aradice-CULT in Table 7, depict the performance of our model and SOTA in existing cultural benchmarks. In both columns, we see solid performance of `Falcon-Arabic-Instruct` compared to SOTA, evident by sharing the best score in Aradice-CULT and being only 6 points away from the highest scoring model in ArabCulture. Looking at Table 7, we see that `Falcon-Arabic-Instruct` obtained comparable scores to high performaning models in both Levantine and Egyptian dialects by being approximately 2 points away from the best model.

6.2.3 English Benchmarks

Although our primary goal was Arabic adaptation of Falcon3-7B, maintaining English performance remained crucial. We monitored Falcon-Arabic's English benchmark performance throughout training (detailed in Section 6.1). Figure 6 reveals minimal English performance gains, likely because our English data overlapped with Falcon3-7B's original training corpus, providing no additional benefit. Table 8 confirms this observation, showing performance degradation in English capabilities. Future work should focus on incorporating novel, high-quality English data during both training and post-training phases to address this limitation.

In summary, Table 5 shows that `Falcon-Arabic` outperformed all base models shown by the highest average achieved without any close competition from other models, making it one of the best base models in Arabic tasks. Table 6 shows that `Falcon-Arabic-Instruct` outscored all competing SOTA models, with solid performance on STEM subjects, Arabic grammar understanding, and truthfulness. However, scores in ALRAGE, indicated that `Falcon-Arabic-Instruct` is still lacking in RAG capabilities. Table 7 indicates that our instruct model slightly trails in cultural awareness and reasoning, although the performance gap with the leading model is relatively small.

7 Limitations

As with any Large Language Model, Falcon-Arabic is subject to inherent limitations that users must carefully consider (Ashraf et al., 2025). The model can exhibit hallucination behaviors, generating factually incorrect information or fabricating details that appear plausible but are not grounded in reality (Huang et al., 2025). Additionally, despite our efforts to train on high-quality, culturally-authentic Arabic datasets, Falcon-Arabic may still produce toxic, biased, or unsafe content that could be harmful or offensive to users (Mubarak et al., 2024).

The Arabic adaptation of Falcon3-7B reveals a common trade-off in language-specific fine-tuning: while Arabic capabilities improved, English performance declined slightly, indicating that the current adaptation methodology may not optimally balance multilingual retention with Arabic enhancement.

Furthermore, the model's performance on Arab culture and Arabic-GSM8K benchmarks highlights domain-specific limitations. The cultural knowledge gaps likely stem from insufficient exposure to diverse regional content during training, limiting representation of varied cultural contexts across Arabic-speaking regions. The mathematical reasoning deficiencies on Arabic-GSM8K reflect a

Model	Size	Arabic-GSM8K	ArabCulture	Aradice-CULT	Aradice-LEV	Aradice-EGY
Qwen2.5-Instruct	7B	**62.55**	53.27	38.89	43.50	45.00
Jais-adapted-chat	7B	10.16	56.86	35	44.87	46.04
	13B	46.25	71.45	40.56	48.41	49.10
AceGPT-v2	8B	45.87	35.44	47.78	49.9	51.04
Llama-3.1-Instruct	8B	49.58	47.53	37.78	43.38	44.79
c4ai-command-r7b-arabic	7B	60.05	67	45	48.44	48.70
aya-expanse-8b	8B	57.77	50.46	47.22	47.66	50.02
ALLaM-Instruct-preview	7B	52.01	67.49	**51.67**	**53.40**	**53.26**
Yehia-preview	7B	50.04	67.58	51.11	51.81	52.52
SILMA-Instruct-v1.0	7B	33.28	**71.6**	41.67	52.13	52.30
Falcon-Arabic-Instruct	7B	54.89	65.16	**51.67**	51.01	51.96

Table 7: Falcon-Arabic-Instruct vs. best open source SOTA instruct models on cultural, dialectal and reasoning benchmarks. **Bold** indicates the best score in each column; underline indicates the second best.

Model	IFEval	GPQA	MMLU-pro	BBH	MUSR	MATH	GSM8K	Hellaswag	ARC Challenge	Winogrande	TruthfulQA	MMLU	Avg
	0-shot	0-shot	5-shot	3-shot	0-shot	4-shot	5-shot	10-shot	25-shot	5-shot	0-shot	5-shot	
Falcon3-7B	**33.9**	**12.8**	**32.34**	**31.8**	**18.1**	**18.5**	**76.6**	**75.54**	**51.0**	**71.0**	**37.3**	**67.4**	**43.86**
Falcon-Arabic	29.1	8.7	28.9	26.6	7.4	12.8	62.0	73.4	49.7	69.9	31.5	60.1	38.34
Falcon3-7B-Instruct	**76.12**	**8.05**	**34.3**	**37.92**	**21.17**	**40.86**	**81.5**	**78.43**	**62.6**	**70.4**	**55.42**	**70.5**	**53.11**
Falcon-Arabic-Instruct	57.6	4.5	28.3	28.5	19.4	12.3	67.7	71.4	53.5	68.42	31.5	63.34	42.21

Table 8: Falcon model evaluation scores on English benchmarks.

domain mismatch: our model, trained on native Arabic mathematical discourse, struggles with the translated benchmark's English-centric reasoning patterns and problem formulations that don't align with authentic Arabic mathematical conventions.

8 Conclusion

In this work, we present Falcon-Arabic, a successful adaptation of Falcon3-7B to Arabic through vocabulary extension, multi-stage training, and exclusive use of native Arabic datasets. Our methodology involved extending Falcon3-7B tokenzier, implementing a gradual training recipe that preserves existing capabilities while incorporating diverse Arabic linguistic varieties. Post-training phases including SFT and DPO further enhanced instruction-following and cultural alignment.

The resulting Falcon-Arabic demonstrates that targeted adaptation with high-quality, native data can achieve exceptional performance, outperforming models two times its size while maintaining strong cultural relevance and linguistic authenticity. Our work provides valuable insights for effective language model adaptation strategies, showing that careful attention to tokenization, training design, and data authenticity can yield powerful models for underrepresented languages with limited computational resources. Future work will focus on improving the model on multiple areas including math, culture and RAG style of questions.

Acknowledgments

We would like to express our sincere gratitude to Younes Belkada for his invaluable assistance with training the Arabic tokenizer and integrating Falcon-Arabic in the English evaluation pipeline. We also extend our special thanks to Mikhail Lubinets for his support with the training infrastructure, and to Mohammed Chami for his efforts in crawling Arabic news and STEM websites. Finally, we are deeply grateful to Puneesh Khanna and Iheb Chaabane for our insightful discussions regarding the training codebase and hyperparameter optimization.

References

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

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Eman Albilali, Nora Al-Twairesh, and Manar Hosny. 2022. Constructing arabic reading comprehension datasets: Arabic wikireading and kaiflematha. *Language Resources and Evaluation*, 56(3):729–764.

Emad A. Alghamdi, Reem I. Masoud, Deema Alnuhait, Afnan Y. Alomairi, Ahmed Ashraf, and Mohamed Zaytoon. 2024. Aratrust: An evaluation of trustworthiness for llms in arabic. *Preprint*, arXiv:2403.09017.

Ebtesam Almazrouei, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, Guilherme Penedo, Giulia Campesan, Mugariya Farooq, Maitha Alhammadi, Julien Launay, and Badreddine Noune. 2023. AlGhafa evaluation benchmark for Arabic language models. In *Proceedings of ArabicNLP 2023*, pages 244–275, Singapore (Hybrid). Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraGPT2: Pre-trained transformer for Arabic language generation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 196–207, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Yasser Ashraf, Yuxia Wang, Bin Gu, Preslav Nakov, and Timothy Baldwin. 2025. Arabic dataset for LLM safeguard evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5529–5546, Albuquerque, New Mexico. Association for Computational Linguistics.

Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. 2024. Cosmopedia.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021a. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021b. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Wietse de Vries and Malvina Nissim. 2021. As good as new. how to successfully recycle English GPT-2 to make models for other languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 836–846, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Ali El Filali, Manel ALOUI, Tarique Husaain, Ahmed Alzubaidi, Basma El Amel Boussaha, Ruxandra Cojocaru, Clémentine Fourrier, Nathan Habib, and Hakim Hacid. 2025. The open arabic llm leaderboard 2. https://huggingface.co/spaces/OALL/Open-Arabic-LLM-Leaderboard.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Gurpreet Gosal, Yishi Xu, Gokulakrishnan Ramakrishnan, Rituraj Joshi, Avraham Sheinin, Zhiming Chen, Biswajit Mishra, Sunil Kumar Sahu, Neha Sengupta, Natalia Vassilieva, and Joel Hestness. 2024. Bilingual adaptation of monolingual foundation models. In *ICML 2024 Workshop on Foundation Models in the Wild*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pre-training of large language models: How to (re)warm your model? *Preprint*, arXiv:2308.04014.

Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. Morphological analysis and disambiguation for dialectal Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 426–432, Atlanta, Georgia. Association for Computational Linguistics.

Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. 2023. Lighteval: A lightweight framework for llm evaluation.

Xiaotian Han, Yiren Jian, Xuefeng Hu, Haogeng Liu, Yiqi Wang, Qihang Fan, Yuang Ai, Huaibo Huang, Ran He, Zhenheng Yang, and Quanzeng You. 2024. Infimm-webmath-40b: Advancing multimodal pre-training for enhanced mathematical reasoning. *Preprint*, arXiv:2409.12568.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. Acegpt, localizing large language models in arabic. *Preprint*, arXiv:2309.12053.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

Anis Koubaa, Adel Ammar, Lahouari Ghouti, Omar Najar, and Serry Sibaee. 2024. Arabiangpt: Native arabic gpt-based large language model. *Preprint*, arXiv:2402.15313.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation. *Preprint*, arXiv:2305.15011.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024a. Fineweb-edu: the finest collection of educational content.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, Tianyang Liu, Max Tian, Denis Kocetkov, Arthur Zucker, Younes Belkada, Zijian Wang, Qian Liu, Dmitry Abulkhanov, Indraneil Paul, and 47 others. 2024b. Starcoder 2 and the stack v2: The next generation. *Preprint*, arXiv:2402.19173.

Kelly Marchisio, Patrick Lewis, Yihong Chen, and Mikel Artetxe. 2023. Mini-model adaptation: Efficiently extending pretrained models to new languages via aligned shallow training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5474–5490, Toronto, Canada. Association for Computational Linguistics.

Malak Mashaabi, Shahad Al-Khalifa, and Hend Al-Khalifa. 2024. A survey of large language models for arabic language and its dialects.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2024. Aradice: Benchmarks for dialectal and cultural capabilities in llms. *Preprint*, arXiv:2409.11404.

Hamdy Mubarak, Hend Al-Khalifa, and Khaloud Suliman Alkhalefah. 2024. Halwasa: Quantify and analyze hallucinations in large language models: Arabic as a case study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8008–8015, Torino, Italia. ELRA and ICCL.

Navid-AI. 2025. Yehia 7b preview. `https://huggingface.co/Navid-AI/Yehia-7B-preview`.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language. *Preprint*, arXiv:2506.20920.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Mohamed Rashad. 2024. Arabic-nougat: Fine-tuning vision transformers for arabic ocr and markdown extraction. *Preprint*, arXiv:2411.17835.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *Preprint*, arXiv:2311.12022.

Abdelrahman Sadallah, Junior Cedric Tonga, Khalid Almubarak, Saeed Almheiri, Farah Atif, Chatrine Qwaider, Karima Kadaoui, Sara Shatnawi, Yaser Alesh, and Fajri Koto. 2025. Commonsense reasoning in arab culture. *Preprint*, arXiv:2502.12788.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale.

Peter H. Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, and 13 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Guokan Shang, Hadi Abdine, Yousef Khoubrane, Amr Mohamed, Yassine Abbahaddou, Sofiane Ennadir, Imane Momayiz, Xuguang Ren, Eric Moulines, Preslav Nakov, Michalis Vazirgiannis, and Eric Xing. 2024. Atlas-chat: Adapting large language models for low-resource moroccan arabic dialect. *Preprint*, arXiv:2409.17912.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *Preprint*, arXiv:2402.06619.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *Preprint*, arXiv:2310.16049.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Falcon-LLM Team. 2024a. The falcon 3 family of open models.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. 2025. Fanar: An arabic-centric multimodal generative ai platform. *Preprint*, arXiv:2501.13944.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024a. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024b. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Silma Team. 2024b. Silma.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xingjin Wang, Howe Tissue, Lu Wang, Linjing Li, and Daniel Dajun Zeng. 2025. Learning dynamics in continual pre-training for large language models. *Preprint*, arXiv:2505.07796.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, and 375 others. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

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

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

Çagatay Yildiz, Nishaanth Kanna Ravichandran, Prishruit Punia, Matthias Bethge, and Beyza Ermis. 2024. Investigating continual pretraining in large language models: Insights and implications. *CoRR*, abs/2402.17400.

A English Example

Prompt and Responses from Multiple LLMs

Prompt: What does ”your hand is in cold water” mean in a social context?

Qwen2.5-7B-Instruct:
The phrase ”your hand is in cold water” is a common expression in some Arab cultures and is usually used in a social context to indicate a state of confusion, chaos, or discomfort.

Llama-3.1-8B-Instruct:
In a social context, ”your hand is in cold water” means that you need help, especially if you are in a difficult situation or facing financial hardships.

Falcon3-7B-Arabic:
”Your hand is in cold water” is a popular expression used in a social context to refer to someone who is living in comfort or luxury, and is not facing any challenges or difficulties.

Figure 4: Enlgish translation of the example in Figure 1.

B Arabic Benchmarks

Throughout training, we assessed each checkpoint across training stages using Open Arabic LLM leaderboard benchmarks (El Filali et al., 2025). As shown in Figure 5, the majority of performance gains occurred during the initial stage when the model first encountered Arabic data, demonstrating the effectiveness of our tokenizer extension and layer initialization approach. Subsequent stages provided marginal improvements while primarily serving to expand the context window and prevent overfitting.

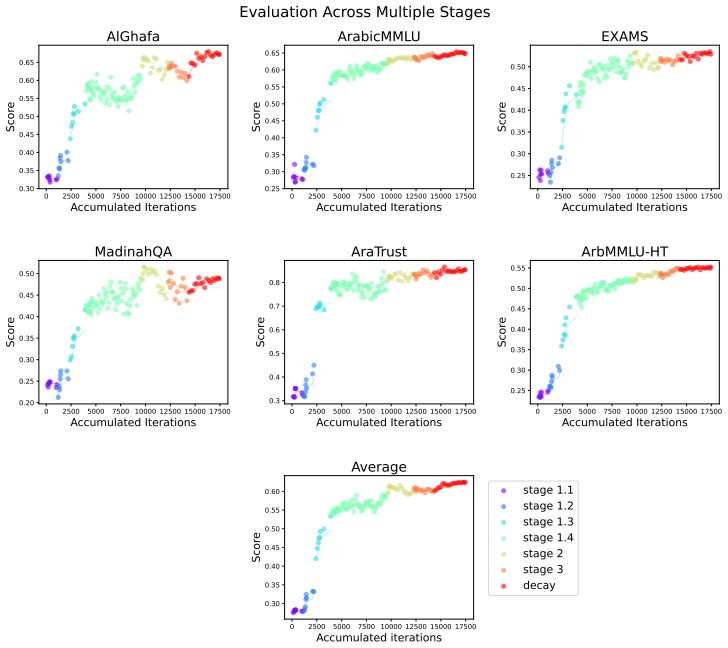

Figure 5: Scores evolution across multiple training stages of Falcon-Arabic on Arabic benchmarks.

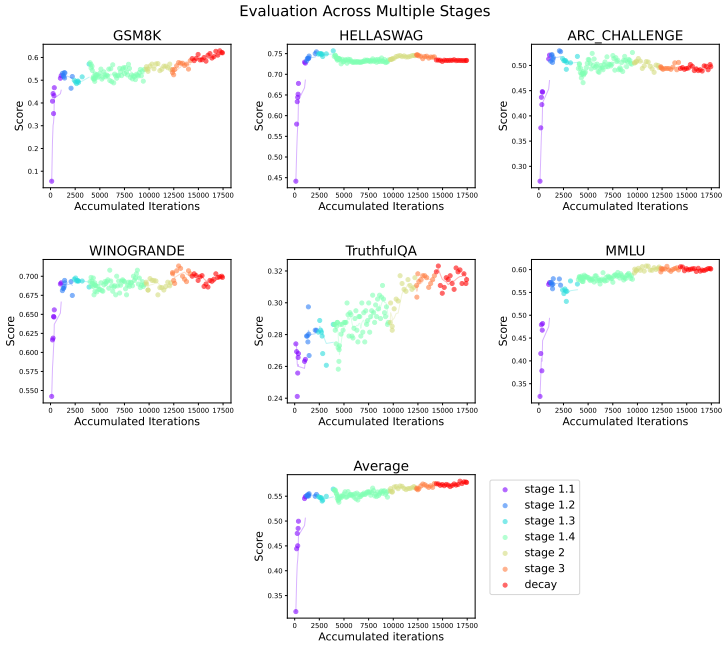

Figure 6: Scores evolution across multiple training stages of Falcon-Arabic on English benchmarks.