Eagle: A Family of Advanced Arabic Large Language Models

Anonymous EMNLP submission

Abstract

In this paper, we present ***Eagle***, a suite of large language models (LLMs) designed for the Arabic language, built on Mistral, LLaMA2, and LLaMA3. We pre-train these models on the ***Oasis*** dataset, containing approximately 35 billion Arabic tokens, and further enhance through instruction fine-tuning and reinforcement learning with AI feedback. We also introduce ***Amwaj***, an Arabic embedding model for retrieval-augmented generation, and ***AraPO***, a novel alignment method for improved Arabic culture alignment. To evaluate our models, we present ***OpenArabicEval***, a diverse benchmark of 32 datasets covering comprehensive multiple-choice evaluation, natural language understanding, natural language generation, and long context evaluation. Extensive testing on ***OpenArabicEval*** demonstrates our models' exceptional performance and robustness across various NLP tasks, highlighting their effectiveness in processing Arabic. ***OpenArabicEval*** is the first benchmark to feature long context evaluation for Arabic LLMs.

1 Introduction

LLMs have revolutionized the field of natural language processing (NLP) by enabling the creation of systems capable of understanding and generating human language with remarkable accuracy and fluency. These models, built on sophisticated neural network architectures, are trained on vast amounts of text data, allowing them to learn the intricacies of language, including grammar, context, and semantics. LLMs lies in the Transformer architecture, introduced by Vaswani et al. (2017), which leverages self-attention mechanisms to handle long-range dependencies in text effectively. This architecture underpins many state-of-the-art models, such as OpenAI's GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020b) and Google's BERT (Devlin et al., 2019), which have demonstrated unprecedented capabilities in various NLP tasks, from machine translation to conversational agents.

One of the key strengths of LLMs is their adaptability to different languages. This adaptability is crucial for developing applications that cater to a global audience. For instance, multilingual models like GPT-4 (OpenAI, 2023a), Bloom(Workshop et al., 2023), XGLM (Lin et al., 2022), mGPT (Shliazhko et al., 2023), Nemotron (Parmar et al., 2024), and LLaMA-3 (AI@Meta, 2024) are designed to understand and generate text in multiple languages, overcoming the challenges posed by linguistic diversity. These models are pre-trained on large multilingual corpora, enabling them to capture the nuances of various languages and dialects.

Arabic encompasses a collection of languages and dialects, some of which (e.g., Moroccan Arabic and Egyptian Arabic) are not mutually intelligible. Classical Arabic (CA), the variety used in ancient Arabic poetry and the Qur'an, continues to coexist with other varieties today. Modern Standard Arabic (MSA) is a more contemporary form of Arabic, typically used in pan-Arab media, government, and formal education across the Arab world (Badawi, 1973). Dialectal Arabic (DA) refers to the various Arabic dialects. These dialects are often categorized regionally (e.g., Gulf, Levantine, Nile Basin, and North African (Habash, 2010; Abdul-Mageed, 2015)), but they can also be defined at the country or even provincial levels (e.g., (Bouamor et al., 2018; Abdul-Mageed et al., 2020b, 2021b, 2020a)).

The adaptability of generative LMs to specific languages such as Arabic involves further-pretraining existing LLMs on language-specific data. This process enhances their performance on tasks related to that language, making them more effective for practical applications such as language translation, sentiment analysis, and question-answering systems (Nagoudi et al., 2023; Elmadany et al., 2022). Additionally, techniques like transfer learning and zero-shot learning further en-

hance the versatility of these models, allowing them to generalize knowledge across languages and domains (Nagoudi et al., 2022a). Arabic, with its rich morphology and diverse dialects, presents unique challenges for NLP. Traditional NLP models often struggles with Arabic due to its script, right-to-left writing direction, and extensive inflectional system. However, advancements in transformer-based architectures have facilitated the creation of highly effective Arabic GPT models. Arabic GPT models such as, Jasmine (Nagoudi et al., 2022a), AceGPT (Huang et al., 2023) and Jais (Sengupta et al., 2023) are pre-trained on large, diverse Arabic text corpora to capture the syntactic, semantic, and contextual nuances of the language. The pre-training phase involves predicting the next word in a sentence, which enables the model to learn language patterns and context effectively.

In this paper we presents an extensive suite of LLMs specifically designed for the Arabic language. Our contributions are as follows: **(1)** We introduce ***Eagle***, three cutting-edge, powerful Arabic LLMs built on Mistral-7B (Jiang et al., 2023), LLaMA2-7B (Touvron et al., 2023a), and LLaMA3-8B (Zhang et al., 2024). **(2)** ***Oasis*****:** An efficient training dataset for Arabic Large Language models that include three subsets: Pretraining, Instruction and Reward modelling **(3)** ***Amwaj***, two robust Arabic embedding models for enhanced Arabic language understanding, **(4)** ***OpenArabicEval***, a comprehensive and diverse benchmark for evaluating Arabic LLMs, and **(5)** A novel alignment method AraPO for Arabic cultural alignment.

The rest of the paper is organized as follows: We discuss the literature review and related work in Section 2, describe our training dataset (i.e., Oasis) in Section 3, our LLMs in Section 4, and our Chat Models in Section 5. The OpenArabicEval benchmark is presented in Section 6. Section 7 details the experiment and evaluation results. Finally, we conclude in Section 8.

2 Related Works

LLMs have undergone significant advancements in recent years, transitioning from primarily English-centric designs to sophisticated multilingual architectures. This evolution is particularly notable within the realm of causal language models (CLMs), which predict the next word in a sequence based on preceding words. In this section we describe the English-Centric, multilingual, and Arabic CLMs.

English-Centric CLMs. Initially, the majority of CLMs were developed with a strong focus on English. Models such as GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020b), developed by OpenAI, exemplify this trend. In addition, there are many open-source LLMs such as LLaMA (Touvron et al., 2023b,a), Mistral (Jiang et al., 2023), and Gemma (Team et al., 2024) have demonstrated remarkable proficiency in in natural language understanding and generation, but their performance in non-English languages was often limited due to the predominance of English in their training data (Zhang et al., 2023).

Mulilingual CLMs. The increasing global demand for multilingual AI applications has spurred the development of LLMs that can handle multiple languages proficiently. This shift is not just about expanding vocabulary; it involves training models on diverse linguistic structures, syntaxes, and cultural contexts. mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), mBART (Liu et al., 2020) and mT5 (Xue et al., 2021) are notable examples for encode-only and encode-decoder, trained on extensive multilingual datasets to achieve robust performance across a wide range of languages. More recently, developing multilingual CLMs involves several challenges, such as managing the variance in language structures and ensuring equitable performance across languages with differing amounts of available data. Techniques like tokenization strategies, transfer learning, and the use of large-scale multilingual corpora have been employed to address these issues. Additionally, models like GPT-4 (OpenAI, 2023a), Bloom (Workshop et al., 2023), XGLM (Lin et al., 2022), mGPT (Shliazhko et al., 2023), Nemotron (Parmar et al., 2024), and LLaMA-3 (AI@Meta, 2024) have incorporated cross-lingual training methods to enhance their multilingual capabilities.

Arabic CLMs. As the CLMs have demonstrated significant proficiency in various NLP tasks, adapting pre-existing LLMs is the recent trending approach to building language specific LLMs. This can be achieved by using two main ingredients: a robust base LLM, such as LLaMA or Mistral, and a large corpus from target language. Recent work in Arabic LLMs has shown promising results. For instance, Jasmine (Nagoudi et al., 2022a) is a robust Arabic text decoder capable of handling Ara-

bic text generation and classification tasks. Meanwhile, models like AceGPT (Huang et al., 2023) and Jais (Sengupta et al., 2023) have followed the instruction fine-tuning approach of ChatGPT (ChatGPT, 2024), introducing language decoders based on LLaMA2 (Touvron et al., 2023b) that can follow instructions and maintain coherent conversations in Arabic. These advancements, along with contributions from the open-source community, have set new standards and opened up new opportunities for research in the NLP field.

3 Oasis

In this section we introduce the most efficient training dataset for Arabic LLMs. We have three subsets of Oasis: Pretraining, Instruction and Reward modelling.

3.1 Pretraining Dataset

Our pretraining dataset boasts a rich linguistic diversity, including all categories of Arabic, namely Classical Arabic (CA), Dialectal Arabic (DA), and Modern Standard Arabic (MSA). The following details provide an overview of our data sources.

MSA and DIA Data. We use a large and diverse 100 GB of Arabic text, amounting to 35 billion tokens. This data is aggregated from various sources: AraNews$_{v2}$(Nagoudi et al., 2020), El-Khair(El-Khair, 2016), Gigaword,[1] OSCAR (Suárez et al., 2019), OSIAN (Zeroual et al., 2019), Wikipedia Arabic, and Hindawi Books.[2] We also derived ArabicWeb22 (A) and (B) from the open source Arabic text 2022[3], AraNews$_{v2}$ (Nagoudi et al., 2020), El-Khair (El-Khair, 2016), Gigaword,[4] OSCAR (Suárez et al., 2019), OSIAN (Zeroual et al., 2019), 101 Billion Arabic words (Aloui et al., 2024), Wikipedia Arabic, and Hindawi Books.[5] We also derived ArabicWeb22 (A) and (B) from the open source Arabic text 2022.[6] This pretraining dataset was cleaned, filtered and deduplicated using Bhatia (2023).

CA Data. Our primary source for CA data is the Open Islamicate Texts Initiative (OpenITI) corpus (v1.6) (Nigst et al., 2020). The corpus comprises 11,195 premodern Islamic books, primarily sourced from Shamela Library,[7] the Al-Jami Al-Kabir collection (JK),[8] texts digitized by the Jordanian publisher Markaz Al-Turāth, and the Shia Library.[9]

3.2 Instruction Dataset

Table 1 presents the composition of the $Oasis_{Inst}$ dataset, aggregating a wide array of data sources and sample sizes. To build the Oasis instruct dataset, we first collected all open-source high-quality Arabic NLU (Elmadany et al., 2022) and NLG (Nagoudi et al., 2023) datasets. We also use the Arabic split of the MultilingualSIFT dataset by Chen et al. (2023); Huang et al. (2024).To further enrich our dataset, we follow the methodology of (Teknium, 2023) to build the Arabic Generation Dataset. This is generated synthetically using multiple different datasets as seed data. We also collect high-quality human answers to various common questions from `https://mawdoo3.com/`. Finally, we built a long context Arabic instruction dataset which has an average length of $60k$ tokens per sample. The methodology to build this dataset was inspired by (Zhang et al., 2024). Finally, we use the Human Inst dataset from (Alwajih et al., 2024) to further improve the model. We have $9.5k$ annotations from 6 different countries; this dataset was built as a part of a larger ongoing project.

3.3 Reward Dataset

Creating a Reward Dataset (RD) is crucial for training reward models that can effectively steer LLMs to produce high-quality responses that align with human preferences. The process involves generating some rejected samples for the human-accepted samples to train models effectively. Hence, we generate the rejected samples for our reward dataset using 4 different models and randomly select one from the four generations. We use the following models: AceGPT-13B Chat (Huang et al., 2024), Command-R+, GPT-3.5-Turbo (Brown et al., 2020a) and GPT-4-Turbo (OpenAI, 2023a).

4 Arabic Language Models

Traditional approaches to training LLMs have either involved training entire models on vast datasets, which is computationally expensive (Sengupta et al., 2023), or continuously pre-training on

[1] https://catalog.ldc.upenn.edu/LDC2009T30.
[2] https://www.hindawi.org/books.
[3] https://data.baai.ac.cn/details/ArabicText-2022
[4] LDC Catalog Link
[5] OpenITI corpus (v1.6) (Nigst et al., 2020).
[6] ArabicText-2022 data
[7] https://shamela.ws.
[8] http://kitab-project.org/docs/openITI.
[9] https://shiaonlinelibrary.com.

Task	#Datasets	Source	#Samples
AG Inst.	1	**Our Paper**	$981k$
ORCA	60	Elmadany et al. (2022)	$500k$
Quora QA	1	Huang et al. (2024)	$247k$
Transliteration	3	Ameur et al. (2019); Talafha et al. (2021); Merhav and Ash (2018)	$209K$
MultiSIFT	3	Huang et al. (2024)	$114k$
Summarization	5	Chouigui et al. (2021); Bhattacharjee et al. (2021); Varab and Schluter (2021); Hasan et al. (2021); Gaanoun et al. (2022)	$75k$
Diacritization	1	Fadel et al. (2019)	$50k$
Style transfer	2	Mubarak (2018); Alhafni et al. (2022)	$40k$
QA	7	Mozannar et al. (2019); Lewis et al. (2020); Artetxe et al. (2020); Roy et al. (2020); Ismail and Nabhan Homsi (2018); Hardalov et al. (2020)	$29.6k$
MT	6	Eisele and Chen (2010); Ziemski et al. (2016); Seddah et al. (2020); Outchakoucht and Es-Samaali (2021); Nagoudi et al. (2022b); Bouamor et al. (2014)	$21.7k$
Mawdoo QA	1	**Our Paper**	$20k$
GEC	3	Mohit et al. (2014); Rozovskaya et al. (2015); Habash and Palfreyman (2022)	$19.4k$
AYA	1	Aryabumi et al. (2024)	$11.5k$
CIDAR	1	Alyafeai et al. (2024)	$10k$
LC Inst.	1	**Our Paper**	$10k$
Human Inst.	1	**Our Paper**	$9.5k$
Paraphrasing	3	Cer et al. (2017); Alian et al. (2019); Scherrer (2020)	$2.1k$
Total	**101**		**2.3M**
Enhanced	**101**		**2.2M**

Table 1: $Oasis_{Inst}$ Dataset. Here LC Inst. Refers to Long context instructions.

Arabic data (Huang et al., 2024), which faces limitations due to an inefficient tokenizer optimized primarily for high-resource languages. To address these challenges, we introduce three models based on different architectures: AraMistral, AraLlama, and AraMax, which are based on Mistral-7B (Jiang et al., 2023), LLaMA2-7B (Touvron et al., 2023a), and LLaMA3-8B (Zhang et al., 2024), respectively.

Vocabulary Extension. Our first model is ***AraLlama-7B*** is based on LLaMA-2 (7B) (Touvron et al., 2023a). The original LLaMA-2 vocabulary only contains 28 Arabic letters, limiting its effectiveness. Expanding the vocabulary significantly enhances document-level understanding and encoding efficiency (Li et al., 2023). We increase the vocabulary size from 32,000 to 60,000 tokens, initializing new embeddings using a mean method based on the original LLaMA-2 (7B) model. This preserves English proficiency while effectively transferring capabilities to the Arabic model. The model is then trained using LoRA, utilizing all linear layers in the attention module.

Continually Pretraining. Our next model, ***AraMistral-7B***, is pre-trained using the Mistral-7B (Jiang et al., 2023) model with a next token prediction objective. Mistral-7B outperforms LLaMA-2 (13B) with a slightly updated tokenizer. The Mistral model utilizes a better representation of Arabic characters without extending the vocabulary size, as it includes only 51 Arabic tokens. For this reason, we continued training on top of Mistral without any vocabulary extension.

Up-Scaled Continually Pretraining. Our third model, ***AraMax-8B***, is based on LLaMA-3 (8B) and utilizes an innovative method called up-scaled pretraining. We fine-tune input and output embeddings, employing different learning rates for stability to ensure AraMax retains new knowledge. Rank

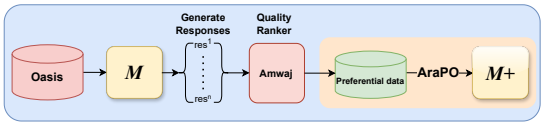

Figure 1: AraPO Methodology.

Stabilized LoRA (Kalajdzievski, 2023) further stabilizes training by adjusting the alpha parameter based on rank. We train using QLoRA, focusing on all linear layers, including embed_tokens and lm_head, with a rank of 128 on 4 A100 GPUs. This approach ensures that AraMax is adept at understanding new Arabic domains, handling long context inputs, and optimizing performance and efficiency.

5 Arabic Chat Models

This section introduces a series of chat versions of our models specifically designed to excel in Arabic language understanding and generation. These models are enhanced through various tuning and alignment techniques to improve their performance.

Instruction Tuning. To enhance the capabilities of our LLMs, we use our instruction tuning dataset described in Section 3.2 to perform instruction fine-tuning on our pre-trained model. As explained earlier, we standardize all the datasets to have the same prompting format.

Alignment with AI Feedback. We employ four alignment methods to enhance the performance of our base model: DPO (Rafailov et al., 2024), CPO (Xu et al., 2024), SimPO (Meng et al., 2024), and our newly proposed AraPO. AraPO leverages a reward model based on Amwaj embeddings, which are fine-tuned to discern and rank optimal responses. These ranked responses are then used to train the model utilizing a combination of SimPO

loss and negative likelihood loss. All experiments used the AraMax model to identify the most effective method. The AraPO loss function is formulated as follows:

$$\min_{\theta} \underbrace{\mathcal{L}(\pi_\theta, U)}_{\mathcal{L}_{\text{SimPO}}} \underbrace{-\mathbb{E}_{(x,y_w)\sim\mathcal{D}}[\log \pi_\theta(y_w|x)]}_{\mathcal{L}_{\text{NLL}}}. \quad (1)$$

In this equation, $\mathcal{L}(\pi_\theta U)$ represents the SimPO loss, which is designed to align the policy π_θ with the constant reference model U. The second term, $-\mathbb{E}_{(x,y_w)\sim\mathcal{D}}[\log \pi_\theta(y_w|x)]$, is the negative likelihood loss (NLL). This term measures the expected log probability of the correct response y_w given the input x, averaged over the dataset $\mathcal{D}$. By minimizing this combined loss, we aim to optimize the model parameters θ such that the policy not only aligns well with the utility function but also maximizes the likelihood of generating the correct responses. Since SimPO also uses length normalisation, it is more efficient to use our case.

Using the negative likelihood loss is beneficial because it encourages the model to assign higher probabilities to the correct responses. This approach helps improve the model's predictive accuracy and ensures that the generated responses align more with the expected or desired outputs. Combining SimPO loss with negative likelihood loss ensures the model is optimized for alignment with the utility function and for generating accurate responses. This approach allows us to systematically evaluate and compare the efficacy of each alignment strategy in improving model performance.

Long Context Training. Since one of the applications of AraMax we have envisioned is to be able to understand large Arabic documents, we use two different methods to make this possible. Firstly we extend the sequence length of AraMax from $8k$ to $64k$ using PoSE (Zhu et al., 2023). After pretraining, we set rope_theta to $500k$ to extend the context to 64k using the Long context instructions subset of the Oasis dataset.

6 OpenArabicEval

We propose OpenArabicEval to identify the most effective model. The OpenArabicEval benchmark comprises five main components: comprehensive multiple choice evaluation, NLU, NLG, and Long Context benchmarks. For the NLU and NLG benchmarks, we follow the same approach as (Khondaker et al., 2023).

6.1 Comprehensive Benchmark

The comprehensive benchmark is designed to evaluate the performance of models in the Arabic language across various tasks, focusing on *Commonsense Reasoning* and *Multiple-Choice Question Answering*. The benchmark includes tasks such as HellaSwag (Zellers et al., 2019) and PiQA (Bisk et al., 2020), which test the model's ability to complete sentences and choose plausible physical solutions, respectively. It also features the AI2 Reasoning Challenge (ARC) (Clark et al., 2018) and BoolQ (Clark et al., 2019), which assess scientific reasoning and comprehension skills. The benchmark incorporates the Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2021), which spans 57 different tasks, and TruthfulQA (Zellers et al., 2021), which evaluates the model's ability to generate truthful responses without prior examples. The datasets above are all translated into Arabic using GPT-4 (OpenAI, 2023b). Including Arabic culture-specific tasks like AlGhafa (Almazrouei et al., 2023) and ACVA (Huang et al., 2023) ensures that the benchmark tests models on culturally relevant contexts and vernacular nuances, making it a critical tool for advancing Arabic NLP research.

6.2 NLU Benchmark

This benchmark consists of 17 datasets across three clusters: **(1) Social Meaning Analysis**, cluster includes nine datasets focusing on various tasks such as sentiment analysis (Abdul-Mageed et al., 2021a), hate and offensive language (Mubarak et al., 2020), dangerous speech (Alshehri et al., 2020), sarcasm (Farha and Magdy, 2020), adult content (Mubarak et al., 2021), irony (Ghanem et al., 2019), and emotion, age, and gender (Mohammad et al., 2018; Abdul-Mageed et al., 2020c). **(2) Dialect Identification (DI)**, which spans three classification levels: binary (MSA vs. DI), country, and city level. This cluster employs various datasets, including $\text{ArSarcasm}_{\text{Dia}}$(Farha and Magdy, 2020), AOC(Zaidan and Callison-Burch, 2014), NADI-2020 (Abdul-Mageed et al., 2020a), MADAR (Bouamor et al., 2019), QADI (Abdelali et al., 2020), and Habibi (El-Haj, 2020). **(3) Fact Checking.** In this cluster, we investigate two tasks. First **(1) Claim Prediction**, utilizing the ANS-claim dataset (Khouja, 2020). Second **(2) Machine-Generated Text Detection (MGTD)**, which uses

Comprehensive Benchmark (CB)		
ACV (Huang et al., 2023)	ARC (Clark et al., 2018)	HellaSwag (Zellers et al., 2019)
AlGhafa(Almazrouei et al., 2023)	BloolQ (Clark et al., 2019)	MMLU (Hendrycks et al., 2021)
PiQA (Bisk et al., 2020)	TruthfulQA (Zellers et al., 2021)	
Natural Language Understanding (NLU) Benchmark		
Social Meaning	**Dialect Identification**	**Fact Chgecking**
Sentiment Analysis (Abdul-Mageed et al., 2021a)	$\text{ArSarcasm}_{\text{Dia}}$ (Farha and Magdy, 2020)	Claim Prediction (Khouja, 2020)
Hate and Offensive (Mubarak et al., 2020)	AOC (Zaidan and Callison-Burch, 2014)	Machine-Generated Detection (Nagoudi et al., 2020)
Dangerous (Alshehri et al., 2020)	NADI-2020 (Abdul-Mageed et al., 2020a)	
Sarcasm (Farha and Magdy, 2020)	MADAR (Bouamor et al., 2019)	
Adult (Mubarak et al., 2021)	QADI (Abdelali et al., 2020)	
Irony (Ghanem et al., 2019)	Habibi (El-Haj, 2020)	
Emotion (Mohammad et al., 2018)		
Age and Gender (Abdul-Mageed et al., 2020c)		
Natural Language Generation (NLG) Benchmark		
MSA and Dialectal MT	**Code-Switching**	**DI→ English Machine Translation**
UNP Corpus (Ziemski et al., 2016)	DZ-FR → FR and JO-EN → EN (Nagoudi et al., 2023)	MDPC (Bouamor et al., 2014)
Long Context Benchmark (Our paper)		
Arabic NarrativeQA (**Our paper**)	Arabic HotpotQA (**Our paper**)	Arabic TriviaQA (**Our paper**)
Arabic PassageDomainQA (**Our paper**)	Arabic MultiFieldQA (**Our paper**)	

Table 2: OpenArabicEval benchmark.

Method	Model	ARC	Hellaswag	ExamsQA	MMLU	Truthfulqa	ACVA	AlGhafa	CB Avg.
BL	Llama2-7B	32.42	42.76	31.39	35.34	45.90	53.22	29.42	38.64
	Mistral-7B	38.53	45.24	32.44	34.53	49.23	63.89	29.95	41.97
	Llama3-8B	40.33	51.90	39.43	42.55	51.45	68.35	30.12	46.30
Arabic BL	AceGPT-7B-base	37.50	48.90	35.75	29.70	43.04	68.96	33.11	42.42
	AceGPT-13B-base	39.90	51.30	39.48	40.50	46.73	75.29	30.37	46.22
	Jais-13b-base	39.60	50.30	39.29	36.90	50.59	68.09	30.07	44.98
VE	**AraLLama-7B-base**	38.40	50.12	38.43	40.23	45.32	69.42	31.52	44.78
CP	**AraMistral-7B-base**	41.50	52.50	38.92	37.50	51.27	69.64	30.24	45.94
UPCP	**AraMax-8B-base**	**44.32**	**52.54**	**40.90**	**43.02**	**50.34**	**75.34**	**34.52**	**48.71**

Table 3: Results of Base Model Evaluations. Here, BL stands for Baselines, Arabic BL stands for Arabic Baselines, VE shows our vocabulary extension strategy, CP denotes continual pretraining, and UPCP denotes scaled continual pretraining.

dataset from (Nagoudi et al., 2020).

6.3 NLG Benchmark

For NLG, we create a benchmark using a collection of 11 datasets from different sources organized into 3 clusters. This benchmark aims to evaluate various aspects of NLG performance, ensuring robust and diverse assessments across multiple domains and tasks: **(1) X→ MSA Machine Translation**. This cluster tests translation from four foreign languages into MSA using the United Nations Parallel Corpus (Ziemski et al., 2016). **(2) DI→ English Machine Translation** translation from five Arabic dialects into English using the Multi-dialectal Parallel Corpus (MDPC) (Bouamor et al., 2014). The **(3) Code-Switching** task translates code-switched Arabic dialectal text into a foreign language using datasets like DZ-FR → FR and JO-EN → EN (Nagoudi et al., 2023).

6.4 Long Context Benchmark

The **LongContext** evaluation involves two settings: one with the entire context and another with retrieved-context using an embedding model. This evaluation includes several datasets to assess the model's comprehension of extensive context. The datasets in this cluster are NarrativeQA, HotpotQA, TriviaQA, MultiFieldQA, and PassageRetrieval, all of which are translated datasets using Google Translate and shown by Alwajih et al. (2024). It is one of the best methods to translate data. These datasets are tested in zero-shot scenarios. NarrativeQA and HotpotQA focus on reading comprehension and multi-hop reasoning, respectively, both benefiting from Retrieval-Augmented Generation

	Task	AraMax	+SFT	+DPO	+CPO	+SimPO	+AraPO
CB	**ARC**	44.32	45.24	48.24	43.89	48.33	**48.44**
	Hellaswag	52.54	54.33	**55.43**	50.42	52.53	53.52
	Exams	40.90	40.23	42.56	41.44	42.03	**45.53**
	MMLU	43.02	45.98	45.3	45.53	45.40	**46.53**
	Truthfulqa	50.34	49.44	53.42	52.55	54.69	**54.7**
	ACVA	75.34	78.23	78.44	75.93	**79.44**	76.55
	AlGhafa	34.52	33.53	35.55	32.89	36.9	**37.66**
NLG	**MT**	20.44	22.54	20.98	**26.35**	24.66	26.24
NLU	**Classification**	45.44	46.90	47.53	49.24	45.30	**50.55**
	Average	45.21	46.27	47.49	46.47	47.70	**48.64**

Table 4: Results of AraMax-8B-base model showing the potency of our **AraPO**'s alignment method.

	Task	AceGPT-7B	AceGPT-13B	Jais-13B	AraLLama	AraMistral	AraMax
CB	**ARC**	38.50	43.80	41.10	39.45	43.20	**48.44**
	Hellaswag	49.80	52.70	57.70	50.23	**55.53**	53.52
	Exams	37.62	42.09	46.74	38.24	45.54	**45.63**
	MMLU	34.30	41.10	42.80	41.03	43.50	**46.53**
	Truthfulqa	49.85	49.96	47.48	50.44	52.44	**54.7**
	ACVA	71.81	78.42	72.56	70.45	**77.06**	76.55
	AlGhafa	31.83	31.95	34.42	32.54	35.57	**35.66**
NLG	**MT**	18.55	23.95	13.56	22.78	23.91	**26.24**
NLU	**Classification**	42.50	45.66	42.45	45.65	46.35	**50.55**
	Average	41.64	45.51	44.31	43.42	47.01	**48.65**

Table 5: Comparison of Chat models on OpenArabicEval Benchmark.

(RAG) to handle complex queries (Kočiský et al., 2017; Yang et al., 2018). TriviaQA, which tests the model's ability to answer trivia questions, also employs RAG for improved performance (Joshi et al., 2017). MultiFieldQA, designed for understanding and answering questions across various fields, and our dataset PassageDomainQA (PDQA), which involves retrieving relevant passages from a large multi-domain corpus, domains include News, Finance, Legal, Medicine and Politics (Karpukhin et al., 2020; Bai et al., 2023). PDQA dataset is built synthetically using Command-R+ as the question generator. We randomly sample 30 passages for each domain and select one for generating questions using Command-R+. The task asks the model to identify the original paragraph to which the crafted summary corresponds. The evaluation in both settings allows for a comprehensive analysis of the model's capabilities with and without an embedding-based retrieval process.

7 Experiments

We train three models of varying architectures: LLama2 (Touvron et al., 2023a), Mistral (Jiang et al., 2023) and LLama3 (AI@Meta, 2024). Our implementation has three stages: Base model pretraining, Instruction alignment for Chat models and Long context extension.

Pre-trained Models Evaluation. In Table 3, we show the results for evaluation of the pre-trained base models we have considered LLama2-7B (Touvron et al., 2023b), Mistral-7B (Jiang et al., 2023) and LLama3-8B (AI@Meta, 2024) as our baselines along with them we also use AceGPT-7B and 13B (Huang et al., 2024) and Jais 13B (Sengupta et al., 2023) as our Arabic baselines. We use different pretraining methods for other models, depending on which strategy suits them the best. For LLama2, we use the vocabulary extension method because its tokenizer is incapable of understanding complex Arabic texts. Hence, we extend the vocabulary to $60k$. We continually pre-train from the available checkpoint using our high-quality dataset for Mistral. Finally, we also train the embedding layers for LLama3 as it has a significantly bigger tokenizer of $128k$ tokens to ensure the models learn the new distribution of the tokens used for the Arabic language. As seen from Table 3, we see that our AraMax model significantly outperforms all the other models with an average Comprehensive benchmark score of 48.71, AraMistral follows it with an average score of 45.94, and finally, our vocabulary extended model AraLLama has an average score of 44.78.

Chat Models Evaluation. In Table 5, we present the performance of various chat models on the OpenArabicEval benchmark, focusing on their average scores across multiple tasks. These models are AraLLama, AraMistral, and AraMax, alongside AceGPT-7B, AceGPT-13B, and Jais-13B. Here, we compare all the instructed models to their own, ensuring a fair comparison. The results show that AraMax consistently achieves the highest average score of 48.65, indicating its superior overall performance. This is followed by AraMistral, with an average score of 47.01, demonstrating its robustness and effectiveness. AceGPT-13B comes next with an average score of 45.51, showing competitive performance, albeit slightly behind AraMax and AraMistral. Jais-13B has an average score of 44.31, while AraLLama follows with an average of 43.42. AceGPT-7B, has the lowest average score of 41.64. Moreover, these average scores highlight the effectiveness of AraMax in handling a diverse range of tasks within the OpenArabicEval benchmark, underscoring the advantages of the model enhancements and alignment strategies employed in its development.

Alignment Methods Evaluation. AraPO is a new alignment method specifically designed for Arabic LLMs. Table 4 shows the results of evaluating our chat models using different alignment methods. We compare our base model (i.e., AraMax) with several alignment techniques: SFT,DPO, CPO, SimPO, and our novel AraPO. The results demonstrate that AraPO is an all-around alignment method that boosts performance across different tasks. Notably, DPO performs better at commonsense reasoning tasks like Hellaswag and ARC, whereas DPO leads to a decrease in score for tasks like machine translation. CPO excels at crosslingual tasks like machine translation and natural language understanding. Also, AraPO demonstrates

	Model	Seq Len	NarrativeQA	HotpotQA	TriviaQA	MultiFieldQA	PDQA	Average
LongContext	Jais-13b-chat	2048	9.67	12.42	78.29	33.66	43.56	35.52
	AceGPT-7B-chat	2048	16.54	10.54	74.33	25.66	53.99	36.21
	AraLLama-7B-Chat	2048	15.52	9.54	72.56	45.56	45.63	37.76
	AceGPT-13B-chat	2048	18.96	13.90	75.66	23.92	58.24	38.14
	ArMistral-7B-Chat	4096	20.53	20.53	76.66	43.12	51.55	42.48
	AraMax-8B-Chat	8192	25.55	22.67	77.64	50.98	**63.80**	48.13
	GPT-3.5-Turbo-16k	16385	31.52	**55.35**	91.55	**74.35**	56.77	61.91
	AraMax-8B-Chat-64K	65536	**35.66**	49.54	**94.98**	68.92	62.25	**62.27**
LongContext+RAG	Jais-13b-chat	2048	10.90	17.30	78.29	35.99	46.43	37.78
	AceGPT-7B-chat	2048	21.35	11.49	75.75	26.70	56.90	38.44
	AceGPT-13B-chat	2048	19.53	16.64	76.45	28.86	61.32	40.56
	AraLLama-7B-Chat	2048	20.25	13.58	76.65	49.86	48.03	41.67
	ArMistral-7B-Chat	4096	25.06	21.57	78.95	45.01	55.50	45.22
	AraMax-8B-Chat	8192	27.58	24.44	79.11	45.18	**65.46**	48.35
	GPT-3.5-Turbo-16k	16385	33.90	51.58	93.06	74.63	58.46	62.33
	AraMax-8B-Chat-64K	65536	**37.07**	**56.14**	**93.48**	**75.68**	62.32	**64.94**

Table 6: Results of Long context benchmark.

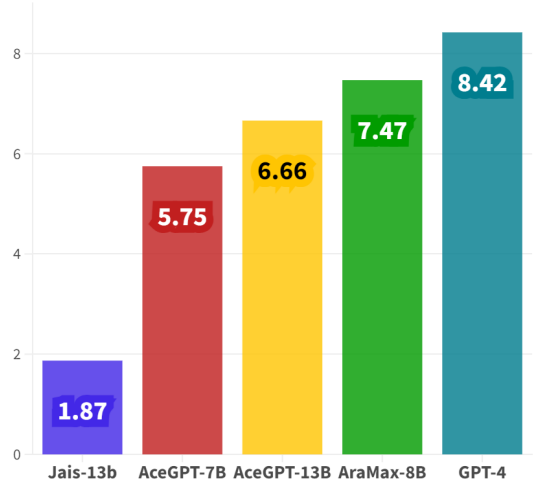

Figure 2: Human Evaluation Results.

superior performance with an average score of 48.64, indicating its effectiveness as an alignment method, particularly for Arabic language tasks.

Long Context Results. We evaluate all models on the LongContext dataset. As shown in Table 6, the instructed version of our model, AraMax-8B, outperforms all other models as well as ChatGPT-3 Turbo by an average of 2.61 points. Notably, both versions of AraMax-chat are strong performers on the Long Context benchmark. By extending the context length of AraMax from $8K$ to $64K$, we significantly improve the performance on the Long context benchmark. Additionaly, our setting using embedding models to retrieve only relevant data from the provided context. There, we see that using RAG improves performance by an average of 3 points for every model. However, this does not discount the importance of long-context models, as our AraMax-64K performs even better with limited context. Models with smaller sequence lengths, like AceGPT and Jais, struggle with answering questions correctly, even when given a limited and relevant context.

Human Evaluation Results. To further assess the performance of our models, we conduct a human evaluation using 80 questions from various Arabic domains. Annotators were asked to rate from 1-10 the generated answers by different models. The results are illustrated in Figure 2. The scores indicate that Jais-13b achieved a score of 1.87, AceGPT-7B scored 5.75, AceGPT-13B scored 6.66, AraMax-8B scored 7.47, and GPT-4 achieved the highest score of 8.42. These findings highlight the superior performance of GPT-4, followed closely by our model AraMax-8B.

8 Conclusion

In this paper, we introduce Eagle, a suite of cutting-edge Arabic LLMs, built upon advanced English and multilingual models like Mistral-7B, LLaMA2-7B, and LLaMA3-7B, and trained on the diverse 35B tokens from Oasis dataset. Enhancements through instruction fine-tuning along with novel algorithms like AraPO, and the development of Amwaj, an Arabic embedding model, significantly boosted the models' performance. We also propose OpenArabicEval, a comprehensive benchmark for evaluating Arabic LLMs across multiple NLP tasks. Our flagship model, AraMax, outperform existing models in various benchmarks, and the novel AraPO alignment method improves the models performance, highlighting the potential of these advancements to enhance Arabic NLP applications such as MT, sentiment analysis, and QA.

9 Limitations

We identify the following limitations in our work:

1. Despite our efforts to include extensive dialectal texts in our pretraining data, our automated analysis indicates that the dataset still lacks broad coverage of certain dialects, such as Algeria, Iraqi, Sudanese, Syrian, and Yemeni.

2. While some studies in the literature employ word lists to filter out toxic and hateful language from pretraining data, we do not adopt this practice. Our goal is to develop models capable of detecting toxic and hateful language as few-shot learners. Additionally, we believe that using word lists, though potentially effective in removing some antisocial content, may only provide a superficial level of data cleaning. Nonetheless, we emphasize that our models should be used with caution, and strategies to mitigate social risks, biases, and toxicities should be meticulously applied.

3. One significant disadvantage of CLMs is their potential misuse for generating fake content or spreading misinformation at scale, which is one of the most dangerous applications of these models. Consequently, we believe that all necessary measures should be taken to regulate their use, and ourmodels are no exception. This may include implementing regulations and policies that restrict these models to pro-social applications, such as in education, travel, and recreation. Due to these concerns, we will release our models responsibly.

10 Ethics Statement

Energy Efficiency. Our models models, like many CLMs, required substantial pretraining time and are not energy efficient. We recognize this important issue and believe that efforts to develop energy-efficient models should continue to receive scholarly attention.

Data. Our pretraining datasets are sourced from the public domain and encompass diverse genres, communities, and varieties of Arabic. As we have demonstrated, our models have the potential to support applications across various Arabic dialects and serve a wide range of populations. We emphasize that all the datasets we use are collected from publicly available sources, ensuring that our data collection process does not violate any copyrights

Human Annotation. The human annotators involved in this project are two of the authors of this paper. Both annotators are native Arabic speakers with Ph.D. degrees and extensive experience in NLP. They are full-time employees of the research group responsible for this work, with data annotation included in their job duties. No Institutional Review Board (IRB) review or approval was required for this project since we only use publicly available data, which does not require access to any social networking accounts or passwords. Additionally, no external annotators were involved in this work.

References

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2020. Arabic Dialect Identification in the Wild. *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Muhammad Abdul-Mageed. 2015. *Subjectivity and sentiment analysis of Arabic as a morophologically-rich language*. Ph.D. thesis, Indiana University.

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

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020a. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

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

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020b. Toward micro-dialect identification in diaglossic and code-switched environments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5855–5876, Online. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, Azadeh Hashemi, and El Moatez Billah Nagoudi. 2020c.

AraNet: A Deep Learning Toolkit for Arabic Social Media. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 16–23, Marseille, France. European Language Resource Association.

AI@Meta. 2024. Llama 3 model card.

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022. User-Centric Gender Rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 618–631, Seattle, United States. Association for Computational Linguistics.

Marwah Alian, Arafat Awajan, Ahmad Al-Hasan, and Raeda Akuzhia. 2019. Towards building arabic paraphrasing benchmark. In *Proceedings of the Second International conference on Data Science E-learning and Information Systems (DATA' 2019)*, pages 1–5.

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.

Manel Aloui, Hasna Chouikhi, Ghaith Chaabane, Haithem Kchaou, and Chehir Dhaouadi. 2024. 101 billion arabic words dataset.

Ali Alshehri, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2020. Understanding and detecting dangerous speech in social media. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 40–47, Marseille, France. European Language Resource Association.

Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. Peacock: A family of arabic multimodal large language models and benchmarks. *arXiv preprint arXiv:2403.01031*.

Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran A. Q. Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, Yousef Ali, and Maged S. Al-Shaibani. 2024. Cidar: Culturally relevant instruction dataset for arabic.

Mohamed Seghir Hadj Ameur, Farid Meziane, and Ahmed Guessoum. 2019. Anetac: Arabic named entity transliteration and classification dataset. *arXiv preprint arXiv:1907.03110*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress.

MS Badawi. 1973. Levels of contemporary arabic in egypt. *Cairo: Dâr al Ma'ârif*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2023. Longbench: A bilingual, multitask benchmark for long context understanding.

Gagan Bhatia. 2023. PolyDeDupe.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong bin Kang, and Rifat Shahriyar. 2021. Crosssum: Beyond english-centric cross-lingual abstractive text summarization for 1500+ language pairs. *CoRR*, abs/2112.08804.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. `https://leaderboard.allenai.org/piqa`.

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

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020b. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

ChatGPT. 2024. Openai. www.openai.com. Version GPT-4.

Zhihong Chen, Shuo Yan, Juhao Liang, Feng Jiang, Xiangbo Wu, Fei Yu, Guiming Hardy Chen, Junying Chen, Hongbo Zhang, Li Jianquan, et al. 2023. Multilingualsift: Multilingual supervised instruction fine-tuning, july 2023b. *URL https://github.com/FreedomIntelligence/MultilingualSIFT. git*.

Amina Chouigui, Oussama Ben Khiroun, and Bilel Elayeb. 2021. An arabic multi-source news corpus: Experimenting on single-document extractive summarization. *Arabian Journal for Science and Engineering*, 46:3925–3938.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. https://github.com/google-research-datasets/boolean-questions.

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. https://leaderboard.allenai.org/arc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, et al. 2020. Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Mahmoud El-Haj. 2020. Habibi: a multi Dialect multi National Arabic Song Lyrics Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1318–1326.

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

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

Ali Fadel, Ibraheem Tuffaha, Bara' Al-Jawarneh, and Mahmoud Al-Ayyoub. 2019. Arabic text diacritization using deep neural networks.

Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic Sentiment Analysis to Sarcasm Detection: The ArSarcasm Dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39.

Kamel Gaanoun, Abdou Naira, Anass Allak, and Imade Benelallam. 2022. *Automatic Text Summarization for Moroccan Arabic Dialect Using an Artificial Intelligence Approach*, pages 158–177.

Bilal Ghanem, Jihen Karoui, Farah Benamara, Véronique Moriceau, and Paolo Rosso. 2019. IDAT@FIRE2019: Overview of the Track on Irony Detection in Arabic Tweets. . In *Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. In: CEUR-WS.org, Kolkata, India, December 12-15*.

Nizar Habash and David Palfreyman. 2022. ZAEBUC: An annotated Arabic-English bilingual writer corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 79–88, Marseille, France. European Language Resources Association.

Nizar Y Habash. 2010. Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

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.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. https://github.com/hendrycks/mmlu.

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. 2023. Acegpt, localizing large language models in arabic.

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.

Walaa Ismail and Masun Nabhan Homsi. 2018. Dawqas: A dataset for arabic why question answering system. *Procedia Computer Science*, 142:123–131.

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

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension.

Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with lora.

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

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp.

Jude Khouja. 2020. Stance prediction and claim verification: An Arabic perspective. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 8–17, Online. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2017. The narrativeqa reading comprehension challenge.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating Cross-lingual Extractive Question Answering. pages 7315–7330.

Shenggui Li, Hongxin Liu, Zhengda Bian, Jiarui Fang, Haichen Huang, Yuliang Liu, Boxiang Wang, and Yang You. 2023. Colossal-ai: A unified deep learning system for large-scale parallel training. In *Proceedings of the 52nd International Conference on Parallel Processing*, pages 766–775.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward.

Yuval Merhav and Stephen Ash. 2018. Design Challenges in Named Entity Transliteration. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 630–640, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Behrang Mohit, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, and Ossama Obeid. 2014. The first QALB shared task on automatic text correction for Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 39–47, Doha, Qatar. Association for Computational Linguistics.

Hussein Mozannar, Karl El Hajal, Elie Maamary, and Hazem Hajj. 2019. Neural Arabic Question Answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, Florence, Italy. Association for Computational Linguistics.

amina Mubarak. 2018. Dial2msa: A tweets corpus for converting dialectal arabic to modern standard arabic. In *OSACT 3: The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools*, page 49.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020. Overview of OSACT4 Arabic Offensive Language Detection Shared Task. In *Proceedings of the 4th*

Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pages 48–52, Marseille, France. European Language Resource Association.

Hamdy Mubarak, Sabit Hassan, and Ahmed Abdelali. 2021. Adult content detection on Arabic Twitter: Analysis and experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 136–144, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, AbdelRahim Elmadany, Alcides Alcoba Inciarte, and Md Tawkat Islam Khondaker. 2022a. Jasmine: Arabic gpt models for few-shot learning. *arXiv preprint arXiv:2212.10755*.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022b. TURJUMAN: A public toolkit for neural Arabic machine translation. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 1–11, Marseille, France. European Language Resources Association.

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

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

Lorenz Nigst, Maxim Romanov, Sarah Bowen Savant, Masoumeh Seydi, and Peter Verkinderen. 2020. Openiti: A machine-readable corpus of islamicate texts. *nd http://doi. org/10.5281/zenodo*, 4075046.

OpenAI. 2023a. Gpt-4: Technical report. Technical report.

OpenAI. 2023b. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Aissam Outchakoucht and Hamza Es-Samaali. 2021. Moroccan dialect -darija- open dataset.

Jupinder Parmar, Shrimai Prabhumoye, Joseph Jennings, Mostofa Patwary, Sandeep Subramanian, Dan Su, Chen Zhu, Deepak Narayanan, Aastha Jhunjhunwala, Ayush Dattagupta, Vibhu Jawa, Jiwei Liu, Ameya Mahabaleshwarkar, Osvald Nitski, Annika Brundyn, James Maki, Miguel Martinez, Jiaxuan You, John Kamalu, Patrick LeGresley, Denys Fridman, Jared Casper, Ashwath Aithal, Oleksii Kuchaiev, Mohammad Shoeybi, Jonathan Cohen, and Bryan Catanzaro. 2024. Nemotron-4 15b technical report.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5919–5930, Online. Association for Computational Linguistics.

Alla Rozovskaya, Houda Bouamor, Nizar Habash, Wajdi Zaghouani, Ossama Obeid, and Behrang Mohit. 2015. The second QALB shared task on automatic text correction for Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 26–35, Beijing, China. Association for Computational Linguistics.

Yves Scherrer. 2020. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France. European Language Resources Association.

Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, and Abhishek Srivastava. 2020. Building a user-generated content North-African Arabizi treebank: Tackling hell. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1139–1150, Online. Association for Computational Linguistics.

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

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2023. mgpt: Few-shot learners go multilingual.

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

Bashar Talafha, Analle Abuammar, and Mahmoud Al-Ayyoub. 2021. Atar: Attention-based lstm for arabizi transliteration. *International Journal of Electrical and Computer Engineering*, 11:2327–2334.

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, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

Teknium. 2023. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants.

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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023a. Llama 2: Open foundation and fine-tuned chat models.

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

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 6000–6010.

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, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette

Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

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

M. Zellers, N. Holtzman, M. Rashkin, A. Bisk, L. Bosselut, and Y. Choi. 2021. Truthfulqa: Measuring how models mimic human falsehoods. `https://truthfulqa.github.io`.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? `https://leaderboard.allenai.org/hellaswag`.

Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. 2019. OSIAN: Open Source International Arabic News Corpus - Preparation and

Integration into the CLARIN-infrastructure. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 175–182, Florence, Italy. Association for Computational Linguistics.

Peitian Zhang, Ninglu Shao, Zheng Liu, Shitao Xiao, Hongjin Qian, Qiwei Ye, and Zhicheng Dou. 2024. Extending llama-3's context ten-fold overnight.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2023. Pose: Efficient context window extension of llms via positional skip-wise training.

Michal Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *Lrec*.