The JAIS Family

Arabic-Centric Large Language Models

Table of Contents

1. Executive Summary

JAIS 30B v3 builds upon the previous success of the JAIS family, further solidifying its position as the world's leading open-source Arabic LLM. It maintains comparable performance in English compared to other open-source English models of similar size, despite employing a smaller English training dataset. This new version signifies our ongoing commitment to elevating Arabic to the forefront of generative AI research and development.

JAIS 30B v3, a 30-billion parameter GPT model, leverages a massive dataset of 1.63 trillion tokens. This model, like its predecessors, represents a collaborative effort between Core42, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), and Cerebras. The training process utilized the Condor Galaxy 1 (CG-1), a 4 exaFLOP AI supercomputer co-developed by G42 and Cerebras. The training data encompasses Arabic, English, and code, reflecting the model's multilingual capabilities.

JAIS 30B v3 surpasses all known open-source mono- and multi-lingual LLMs in Arabic language tasks. Compared to the previous model, JAIS 30B v3 demonstrates a significant leap in performance across various Arabic downstream tasks. Notably, despite utilizing a smaller English dataset, JAIS 30B v3 achieves English language performance comparable to models like LLaMA 2. This underscores JAIS' pioneering role in multilingual LLM development and its continued contribution to the field's advancement.

2. Introduction

Large language models (LLMs), trained on massive data, revolutionize technology by understanding complex instructions, facilitating problem-solving, and potentially democratizing learning for everyone. This advancement is especially critical for Arabic-centric language preservation. Arabic-centric LLMs are essential for bridging the digital divide and capturing the richness and diversity of the Arabic language. Building upon the JAIS family of models, which are infused with cultural and linguistic nuances of the Arabic language thanks to a purpose-built dataset, we present two new variants: JAIS, with 30B parameters, and JAIS-chat, specifically fine-tuned for chatbot interactions. These models were trained on the state-of-the-art AI supercomputer Condor Galaxy 1 (CG-1) co-developed by G42 and Cerebras Systems and strive to overcome data limitations and bridge the gap in technology access, ultimately aiming for a world where technology fosters connection and understanding.

This whitepaper briefly mentions prior 30B versions, but emphasizes advancements and challenges faced with the larger 30B v3 variant.

3. Model Overview

The JAIS models are based on the GPT-3 architecture. JAIS is built with the goal of enhancing Arabic language capabilities while incorporating English text and code. Trained on bilingual and code data, JAIS is able to handle code-mixed content where Arabic and English intermingle within the same context or sentence. It also allows the model to reason across languages and to leverage knowledge from both English and Arabic sources.

Unlike previous massively multilingual LLMs, such as BLOOM or mT0, which contain more than 50 languages, we do not include languages aside from Arabic and English in any significant percentage. Neither do we relegate Arabic to a minority in the pretraining dataset. Instead, Arabic data constitutes ~33% of the pretraining in all JAIS models. This choice of mixing two languages attains the best of both worlds. The LLM is highly fluent in Arabic, with linguistic capability as well as cultural awareness and sensitivity, while at the same time being on par in terms of reasoning capability and world knowledge that have been observed in recent English and code LLMs.

Building upon established advancements in model architectures and training procedures, as described in our previous whitepaper, we continue to refine the JAIS models as described in the previous release of the whitepaper.

3.1. Model Architecture:

Building upon established architectures, as detailed in our previous release, the model architecture is a causal decoder-only transformer design similar to industry-leading conversational models such as Claude, ChatGPT, and Bard. The models are trained using the standard next word prediction task on a mix of Arabic and English datasets. Table 1 summarizes the models' architecture and shape.

	JAIS-13B	JAIS-30B-v1	JAIS-30B-v3
Number of Decoder Layers	40	48	48
Attention Heads	40	56	56
Model Dimension	5120	7168	7168
Max Context Length	2048	2048	8192

Table 1: JAIS models Architecture and Shape

3.2. Tokenizer:

JAIS family of models leverages a custom multilingual vocabulary containing 84,992 unique tokens, giving equal weight to Arabic and English. This vocabulary facilitates efficient tokenization, impacting model training and inference costs.

Model	Vocabulary Size	English Tokens/Word (Avg)	Arabic Tokens/Word (Avg)
GPT-2	50,257	1.095	4.171
BERT Arabic	32,000	1.632	1.125
BLOOM	250,000	1.083	1.195
JAIS v3	84,992	1.010	1.050

Table 2: Tokens/word of various tokenizers in English and Arabic

3.3. Positional Embeddings:

We use the ALiBi positional encodings, inspired by the LLaMA family of models. ALiBi encodes the position of a word relative to the other words in the context, and so the model can be used with longer text sequences than ever seen during training. This enables training to be faster and less memory-intensive, while unlocking the power of larger contexts during inference.

3.4. Context Length:

JAIS leverages ALiBi positional encodings, as detailed in our previous whitepaper, to effectively extend the context window beyond the standard training window. This capability allows the model to understand and generate text while considering longer sequences of information. Similarly to previous versions, JAIS-30B v3 has been pre-trained and fine-tuned with a context length of 8192 tokens. The foundation laid by previous advancements continues to empower the model for handling longer contexts.

3.5. SwiGLU Activation:

Activation functions play a pivotal role in Large Language Models (LLMs), allowing the model to grasp complex linguistic patterns for a nuanced understanding of language. We use the SwiGLU activation functions, as in the LLaMA model family.

3.6. Maximal Update Parameterization:

Building upon the computational efficiency benefits of Maximum Update Parameterization (muP), as detailed in our previous whitepaper, the JAIS models leverage this technique to efficiently tune hyperparameters across different model sizes.

4. Pretraining

Building upon the Arabic-centric focus of previous JAIS models, v3 further prioritizes Arabic proficiency. This is achieved by significantly increasing the Arabic pretraining data to 475 billion tokens (compared to 140billion in v1, and 267 billion tokens in v2), while maintaining the English and code data at 1.16 trillion tokens, totaling 1.63T tokens.

4.1. Pretraining Data:

Building upon the comprehensive data collection and processing methods established in our previous whitepaper, JAIS v3 leverages an enhanced Arabic pretraining corpus. This corpus incorporates data from various sources, including web pages, Arabic books, and social media content, with a significant increase in Arabic data compared to previous versions.

Table 3 summarizes the final Arabic and English/Code data volumes used in training our JAIS models. The latest version of JAIS is trained on a total of 1.15T tokens, resulting in ~38.3 tokens per parameter.

	English/Code	Arabic	Total
JAIS30B-v1	301B	126B	427B
JAIS30B-v2	654B	267B	921B
JAIS30B-v3	1.16T	475B	1.63T

Table 3: Pretraining data volume in JAIS models

4.2. Training:

JAIS 30B

Because Arabic data is much scarcer than the abundantly available English datasets, the JAIS models employ repeated epochs to bridge the gap between the volume of Arabic data needed to train Arabic centric models of such scale, and that available.

JAIS-13B already uses the above approach to some extent by repeating Arabic data for a total of 1.6 epochs, while at the same time conducting only 1 epoch over English and Code data.

The data scarcity issue is more serious for JAIS-30B, since it is an even larger model. We therefore extend the above methodology and repeat (most of) the Arabic data for a total of 4 phases. This means that while we continue to inject new

knowledge and reasoning prowess through English and Code data, which are abundant, we will repeat the Arabic content to maintain the model's capability to generate and understand Arabic.

The latest JAIS-30B v3 is trained on more than 1.63T tokens, with 475B of these tokens being the Arabic data, much of which is repeated 4 times. The remaining 1.16T tokens are all distinct and deduplicated English + Code. The training is split into three phases, with checkpoints at the end of each phase where we fine-tune, evaluate, and release the model. Phase 1 ended at 427B tokens, while Phase 3 ended at 1.63T tokens of training.

JAIS 30B v3 was trained within 26 days of 48-nodes in CG-1 time. At the end of each phase the checkpoint of JAIS-30B is thus fine-tuned further on instructions and is released as JAIS-30B-chat.

Training Setup: The JAIS models were trained on the powerful Condor Galaxy-1 supercomputer utilizing Cerebras Wafer-Scale Engines and Weight Streaming execution for efficient training.

5. Instruction Tuning

Building upon the successful instruction tuning approach used in JAIS-30B v2, we continue to focus on enhancing capabilities in longer conversations and summarization for this version. Like the previous approach, we leverage a dataset of human-prompted GPT-3.5 conversations, further augmented with Arabic conversations from open-source sources like Orca.

The full instruction dataset comprises more than 10 million examples, with 6 million in English, and 4 million in Arabic. We use packed finetuning, i.e. where prompt and response pairs are packed into a single sequence up to 8,192 tokens. Tokens in the prompt are loss-masked – i.e. the model does not learn to generate tokens as in the prompt. Rather, it is taught to generate tokens as in the target given the corresponding prompt.

6. Performance Evaluations

6.1 LM Harness Evaluations

We conducted a comprehensive evaluation of JAIS-30B-chat-v3 and benchmarked it against other leading base and instruction finetuned language models, focusing on both English and Arabic. Benchmarks used have a significant overlap with the widely used OpenLLM Leaderboard tasks. The evaluation criteria span various dimensions, including:

- Knowledge: How well the model answers factual questions.
- Reasoning: The model's ability to answer questions that require reasoning.
- Misinformation/Bias: Assessment of the model's susceptibility to generating false or misleading information, and its neutrality.

The following results report F1 or Accuracy (depending on the task) of the evaluated models on benchmarked tasks. Both metrics are higher the better.

Arabic Benchmark Results

Model	Avg	Knowledge				Reasoning					Misinformation/ Bias	
		Exams	MMLU	LitQA	Hellaswag	PIQA	BoolQA	Situated QA	ARC-C	OpenbookQA	TruthfulQA	CrowS-Pairs
JAIS-30b-chat-v3	**51.3**	**40.7**	**35.1**	**57.1**	59.3	64.1	**81.6**	**52.9**	39.1	29.6	**53.1**	52.5
JAIS-chat (13B)	49.22	39.7	34	52.6	**61.4**	**67.5**	65.7	47	**40.7**	31.6	44.8	**56.4**
acegpt-13b-chat	45.94	38.6	31.2	42.3	49.2	60.2	69.7	39.5	35.1	**35.4**	48.2	55.9
BLOOMz (7.1B)	43.65	34.9	31	44	38.1	59.1	66.6	42.8	30.2	29.2	48.4	55.8
acegpt-7b-chat	43.36	37	29.6	39.4	46.1	58.9	55	38.8	33.1	34.6	50.1	54.4
aya-101-13b-chat	41.92	29.9	32.5	38.3	35.6	55.7	76.2	42.2	28.3	29.4	42.8	50.2
mT0-XXL (13B)	41.41	31.5	31.2	36.6	33.9	56.1	77.8	44.7	26.1	27.8	44.5	45.3
LLama2-70b-chat	39.4	29.7	29.3	33.7	34.3	52	67.3	36.4	26.4	28.4	46.3	49.6
Llama2-13b-chat	38.73	26.3	29.1	33.1	32	52.1	66	36.3	24.1	28.4	48.6	50
Mixtral 8x7B instruct	43.54	32	34.2	35.4	43.4	58.5	66.2	39.2	32.9	31.6	53.7	51.9

For evaluations, we focus on instruction tuned LLMs that are multilingual or Arabic centric, except for Llama2 13B-chat and Mixtral models. Among Arabic centric models like AceGPT and multilingual models like Aya, both JAIS models outperform all other models by 4+ points. JAIS models outperforming English only LLMs such as Llama2-13B/70B-chat Mixtral 8x7B instruct demonstrates the obvious - though these models are trained on more tokens (2T) and in one case is much larger, JAIS' Arabic centric training gives it a dramatic advantage in Arabic linguistic tasks. Note that Llama or Mixtral's pretraining may include traces of Arabic as evidenced by their limited yet observable capability to understand Arabic, but it is insufficient to obtain an LLM capable of conversing in Arabic, as is expected.

English Benchmark Results

Model	Avg	Knowledge		Reasoning							Misinformation / Bias	
		MMLU	RACE	Hellaswag	PIQA	BoolQA	SIQA	ARC-Challenge	OpenBookQA	Winogrande	TruthfulQA (mc-2)	CrowS-Pairs
JAIS-30b-chat-v3	59.59	36.5	45.6	78.9	73.1	90	56.7	51.2	44.4	70.2	42.3	66.6
JAIS-chat (13B)	57.45	37.7	40.8	77.6	78.2	75.8	57.8	46.8	41	68.6	39.7	68
acegpt-13b-chat	57.84	34.4	42.7	76	78.8	81.9	45.4	45	41.6	71.3	45.7	**73.4**
BLOOMz (7.1B)	57.81	36.7	45.6	63.1	77.4	**91.7**	**59.7**	43.6	42	65.3	45.2	65.6
acegpt-7b-chat	54.25	30.9	40.1	67.6	75.4	75.3	44.2	38.8	39.6	66.3	49.3	69.3
aya-101-13b-chat	49.55	36.6	41.3	46	65.9	81.9	53.5	31.2	33	56.2	42.5	57
mT0-XXL (13B)	50.21	34	43.6	42.2	67.6	87.6	55.4	29.4	35.2	54.9	43.4	59
LLama2-70b-chat	61.25	43	45.2	80.3	80.6	86.5	46.5	49	43.8	**74**	**52.8**	72.1
Llama2-13b-chat	58.05	36.9	45.7	77.6	78.8	83	47.4	46	42.4	71	44.1	65.7
Mixtral 8x7B instruct	**66.03**	**48.5**	**47.3**	**84.3**	**84.3**	88	49.2	**61.3**	**47.6**	**77.6**	**64.9**	73.3

JAIS-30b-chat-v3 outperforms the best other multilingual/ Arabic centric model in English language capabilities by ~2 points. Note that the best model among other Arabic centric models is AceGPT, which finetunes from Llama2-13B. Llama2 models (13B and 70B) are both pre-trained on far more English tokens (2T) vs those that were used for the pretrained JAIS-30B-v3 (0.97T). At less than half the model and pretraining data size, JAIS models reach within 2 points of the English capabilities of Llama2-70B-chat.

6.2 Cultural/ Local Context Knowledge

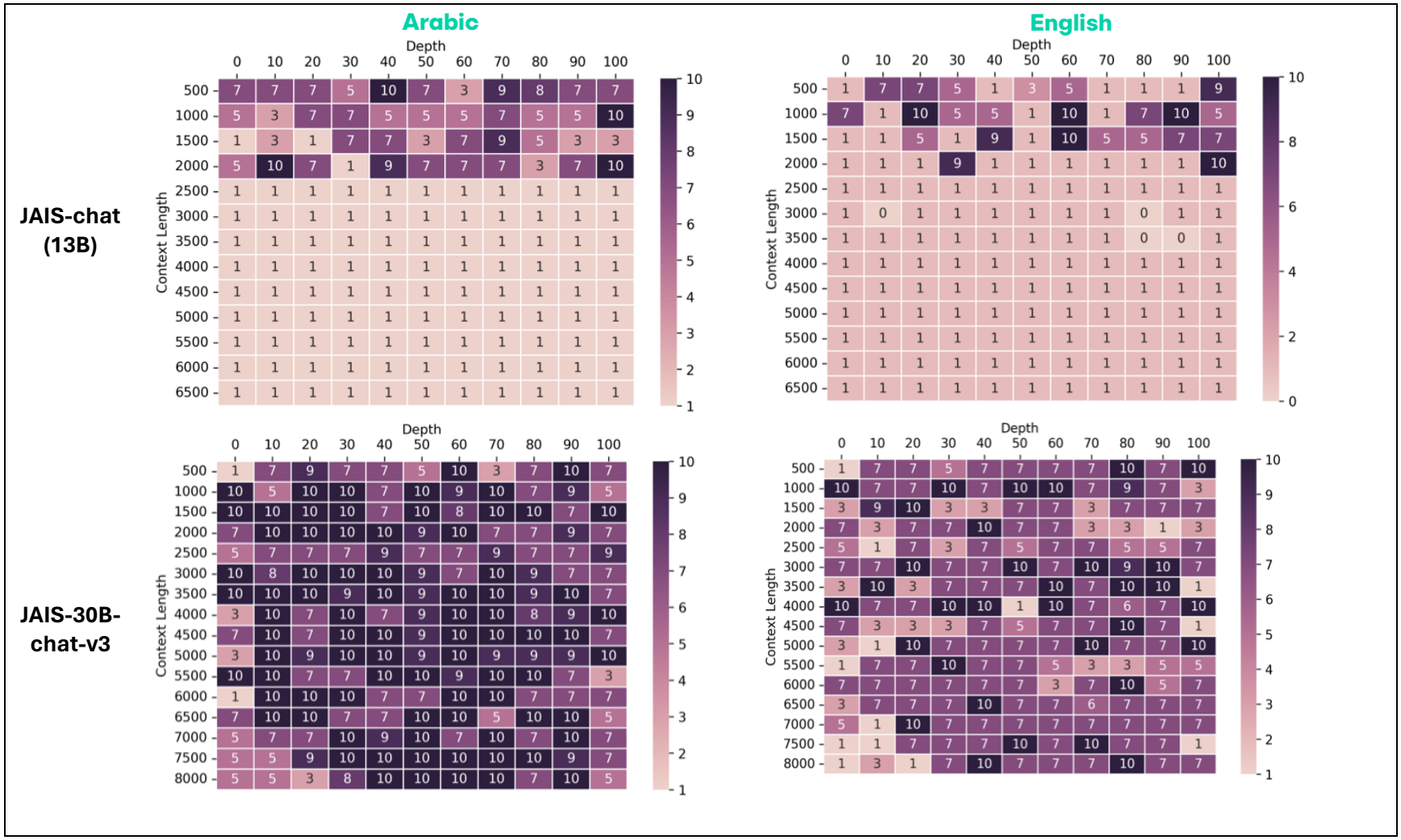
One of the key motivations to train an Arabic LLM is to include knowledge specific to the local context. In training JAIS-30B-chat-v3, we have invested considerable effort to include data that reflects high quality knowledge in both languages in the UAE and regional domains. To evaluate the impact of this training, in addition to LM harness evaluations in the general language domain, we also evaluate JAIS models on a dataset testing knowledge pertaining to the UAE/regional domain. We curated ~320 UAE + Region specific factual questions in both English and Arabic. Each question has four answer choices, and like in the LM Harness, the task for the LLM is to choose the correct one. The following table shows Accuracy for both Arabic and English subsets of this test set.

Model	Arabic	English
JAIS-30b-chat-v3	57.2	55
JAIS-chat (13B)	54.7	45

6.3 Long Context Evaluations

JAIS-30B-chat-v3 offers a context length of 8k tokens. While both JAIS models (13B and 30B) utilize Alibi positional embeddings giving the model an extensible context length, going beyond 1.25 times the training context length results in some decline of accuracy. We explicitly pretrain and finetune JAIS-30B-chat-v3 for 8k tokens in the context.

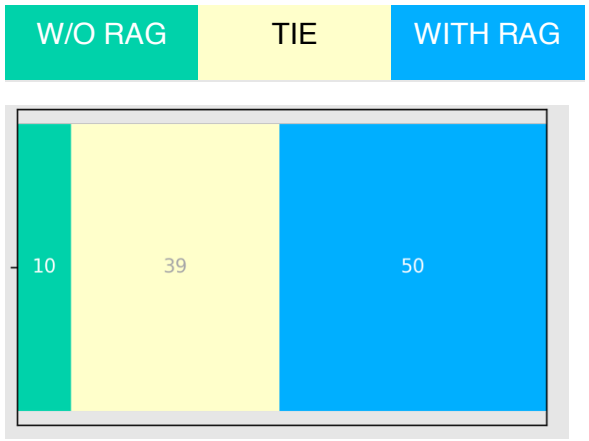
We use the needle-in-a-haystack approach to assess the model's capability of handling long contexts. In this evaluation setup, we input a lengthy irrelevant text (the haystack) along with a required fact to answer a question (the needle), which is embedded at a randomly selected position within this text. The model's task is to answer this question by locating and repeating the relevant phrase (needle) from the text (haystack).


Arabic
English
JAIS-chat (13B)
JAIS-30B-chat-v3
Depth
Context Length

In the above results, we see that the ability of JAIS-chat (13B) to retrieve the relevant fact declines sharply after 2048 tokens. In contrast, JAIS-30B-chat-v3 can retrieve the fact up to 8k tokens with good accuracy. Moreover, performance is generally better in Arabic than in English.

Retrieval-Augmented Generation (RAG) Evaluation

In our RAG experiments, we focused on "recency" as well as long-context capability of JAIS. Recency, in context of LLMs, refers to the model's capability to answer questions related to recent topics which were not present in the model's pretraining or fine-tuning data. In such scenarios, we provide the relevant context as form of text in the prompt itself. The LLM then utilizes this information to answer the query. RAG evaluation focuses on the model's capability to utilize in-context learning and to handle long-contexts, as the context is typically long pieces of text. We use a publicly available dataset called FreshQA for our evaluations. FreshQA has 560 human curated questions focusing on recent events. The dataset provides questions along with correct answer keys. We evaluate JAIS generations with reference to the answer key based on factuality and completeness. We compare vanilla JAIS's performance against JAIS with additional context. As we see from the below figure, adding context in the prompt improves model's factualness by a significant margin.



Win Rates

In summary, our Arabic-English Bilingual LLMs demonstrate the effectiveness of purposeful design and careful training. JAIS LLMs bridge the gap created by limited Arabic data availability (in comparison to English) and highlight their strength against both Arabic and English monolingual models. JAIS LLMs establish that a focused bilingual model, which includes only two major languages, outperforms a highly multilingual model. Although including English datasets have been shown to improve Arabic performance, this behavior may not extend to training and instruction-tuning on several languages together - as illustrated by the large margin by which our models outperform the BLOOM family of models on Arabic tasks. The success achieved at the 13B/30B scale opens a promising path forward for future work in this direction.

7. Safety and Alignment

With the continued improvement in LLM capabilities, and the exceptional growth and solutions to complex problems offered by these models, there comes an intrinsic need to ensure these models are safe and fully aligned with human values and societal norms. Ensuring the safety and alignment of large language models requires robust and reliable systems, careful design and implementation, rigorous testing and validation, and ongoing monitoring. One important way to deal with these requirements is through continuous learning. Indeed, safety and alignment pose a significant challenge for AI systems, and as AI models learn and evolve, they must be constrained by ethical guardrails that prevent drift from originally intended purposes. For that, AI systems are trained based on human preferences and feedback loops, ensuring that they remain aligned with human values as they adapt and grow.

Safety datasets and evaluations

We embed built-in safeguards on the model output during the supervised finetuning process for both JAIS-13B-chat and JAIS-30B-chat.

During instruction-tuning, we added examples containing potentially malicious prompts paired with desirable and safe responses. These taught JAIS-30b-chat to (1) refrain from generating discriminatory or toxic language; (2) never attempt to generate sensitive or private information; (3) respond with caution on domains where inaccurate information could lead to material harm, for instance medicine or law; (4) reject to answer queries about unethical or illegal activities; (5) indicate that it is a chatbot and not a human, particularly when there is a discernible overreliance on its responses; and (6) avoid engaging in discussions on sensitive topics, particularly those related to certain aspects of religion and politics.

We included 21,709 and 22,474 examples of prompt and response pairs of the above form in English and Arabic, respectively. Some of the included datasets already contained relevant and appropriate safe responses. For the datasets that did not include such safe responses, we sampled a response from a collection of pre-constructed safe responses or each prompt.

Responsible Use

The introduction of JAIS, the world's most sophisticated Arabic language model, ushers in a transformative era for the Arabic language and computational linguistics. Developed by Core42 in the UAE, JAIS showcases cutting-edge advancements in artificial intelligence. Capable of generating human-like text, translating between Arabic and English, answering inquiries, and even writing code, JAIS redefines what is possible with language processing. However, as JAIS's capabilities expand, so does the responsibility associated with its usage. Our team at Core42 remains devoted to advancing models while prioritizing ethical considerations. We recognize the need for responsible implementation, particularly with JAIS, to address potential risks and ensure fair and equitable use. This involves devising strategies aligned with ethical principles, preventing discrimination, and avoiding harm. By carefully managing misinformation, protecting privacy, and refraining from causing harm, we enable the beneficial integration of JAIS into society.

The model is trained as an AI assistant for Arabic and English speakers. The model is limited to producing responses for queries in these two languages and may not produce appropriate responses to queries in other languages.

JAIS models must be used with safety guardrails protecting users or systems consuming its output from incorrect, misleading and/or offensive information or content. Information generated by JAIS models is not intended as advice and should not be relied upon in any way, nor are we responsible for any of the content or consequences resulting from its use. We are continuously working to progressively improve the capabilities of JAIS and welcome feedback on the models.

8. Impact

The development and deployment of a bilingual Arabic-English LLM holds the promise of far-reaching implications across linguistic, cultural, and technological dimensions, with a strategic impact that positions governmental and commercial organizations at the forefront of the digital revolution. Our endeavor is a journey towards a future where the power of cutting-edge natural language processing (NLP) not only bridges language barriers, but also fuels advancements in understanding, generation, and deployment of Arabic language applications in diverse contexts.

Empowering the Arabic NLP Community:

The introduction of a powerful, competitive bilingual Arabic-English LLM opens doors to unprecedented advances in Arabic language understanding and generation within the region's Arabic NLP community. Harnessing the model's capabilities, researchers, educators, and innovators are empowered to explore novel use cases. The possibilities range from creative content generation to virtual assistants, and integration into more complex systems such as digital avatars. This empowerment drives innovation and strategically positions the Arabic NLP community as a key player in the global NLP landscape.

Sovereign LLM Implementation:

The inherent sovereignty of this LLM allows organizations across the Middle East to leverage and to deploy the model within their own infrastructures. Our fully in-house implementation ensures complete control over the model's usage and fine-tuning and inference, promoting self-reliance while reducing dependency on external resources. By implementing the LLM locally, governmental and commercial entities can strategically position themselves as technological leaders, driving innovation and digital transformation in their respective domains.

Privacy-Enhanced On-Premise Deployment:

A significant outcome of this endeavor is the capability for local players to fine-tune and to deploy the model on-premise, ensuring complete data privacy and security. The protection of sensitive personal information not only engenders trust, but also strategically positions organizations to excel in today's increasingly privacy-conscious environment. This enables the development of diverse applications, positioning governmental and commercial entities as pioneers in safeguarding individual privacy while delivering advanced Arabic language solutions.

Catalyzing Arabic-Centric Downstream Applications:

A robust Arabic-English LLM will ignite interest within the community, sparking a surge of enthusiasm for Arabic-focused LLMs. This renewed focus on linguistic and cultural nuances stimulates the creation of a myriad of downstream applications that cater to Arabic-speaking populations. By strategically leveraging these applications, governmental and commercial organizations can position themselves as thought leaders, driving innovative solutions aligned with the region's cultural heritage and linguistic diversity.

9. Conclusion

We have introduced here the latest development in JAIS, a family of state-of-the-art Arabic-English bi-lingual large language models (LLM) capable of performing various generative and downstream language tasks in both languages, including language understanding and generation. JAIS-30b-v3 outperforms all existing open Arabic models and matches state-of-the-art open English models trained on larger datasets.

Models are licensed under Apache 2.0 and available on HuggingFace, along with a conversational interface for testing. Researchers, hobbyists, and enterprises are encouraged to experiment with and build upon the model, especially those working on multilingual or non-English applications.

JAIS is testament to the G42-Cerebras Systems partnership aimed at advancing AI research, building access to powerful computing resources, supporting open-source communities, and fostering innovative enterprise app development. Additionally, it represents a set of major milestones for the NLP AI landscape in the Middle East, positioning the UAE at the forefront of the digital revolution while promoting digital and AI transformation, cultural awareness, and linguistic inclusion. Finally, we express gratitude towards the Arabic NLP community for their valuable feedback and involvement in improving the JAIS models.