JAIS-30B

EXPANDING THE HORIZON IN OPEN ARABIC NLP

THE RELEASE

Following the successful release of Jais-13B and Jais-13B-chat in August 2023, we are excited to launch the new state-of-the-art Arabic centric large language model Jais-30B and Jais-30B-chat. With more than twice the number of parameters as that in our previous release, Jais-30B and Jais-30B-chat models exhibit vastly better performance in both Arabic and English languages. Like its predecessor, the Jais-30B is the most powerful open bilingual model at its scale. Not only are the Jais-30B models the best in the world at Arabic NLP and generative AI tasks, they also are highly competitive with English language models of a similar size.

We are also proud to announce the release of Jais-30B and Jais-30B-chat model weights along with the accompanying inference code to the community. This release marks a significant milestone in our ongoing commitment to elevate Arabic language processing by positioning it at the forefront of generative AI research and development.

MODEL

Building upon our learnings from Jais-13B, we scaled up the model size while applying recent developments in model architecture. The backbone of Jais-30B is a causal decoder-only large language model. It is engineered with 48 transformer blocks, 56 attention heads, and an embedding dimension of 7168. By extending the number of transformer blocks and attention heads, we have enhanced the model's reasoning capabilities, thereby allowing it to perform deeper operations across more granular concepts per attention head. The expanded scale and architectural enhancements — like the Swiglu activation function and ALiBi positional encodings — have been carried over from Jais-13B, contributing to significant performance improvements across the board in Jais-30B.We remain dedicated to keeping Jais an Arabic-first model. For this, we used the custom Jais tokenizer, which is trained on equal proportions of English and Arabic. This tokenizer reduces over-segmentation of words which means that Jais-30B is computationally efficient in both training and inference. The reduced word segmentation also benefits cross-lingual transfer through enhanced token-level alignment. With Arabic treated equally to English, we developed a curriculum based bilingual language mix. At the

core of this is a ratio of 1:2 of Arabic to English data. Through experimentation, we found that this enables cross-lingual knowledge transfer across both languages while maintaining a highly linguistically capable Arabic LLM.

The context window is an important element of consideration in many use cases. Its length refers to the amount of text, or number of tokens, that a model can reason over while generating text. Towards this, we extended the effective context window of our model through ALiBi position encodings. ALiBi biases query and key attention values based on relative distances, meaning that Jais-30B, similar to Jais-13B, can understand and generate text well beyond its standard training window, unlocking capabilities beyond its training data.

Jais-30B is significantly larger and trained on an extensive dataset therefore necessitating the optimization of computation efficiency. In the pre-training phase, Maximal Update Parametrization enabled efficient search of hyperparameters. Rather than sweeping Jais-30B, we performed the search against smaller models and transferred these parameters to larger one. For fine-tuning, we experimented with the packing of smaller sequences into the same context length ensuring computational resources are fully utilized by avoiding padded tokens per sequence. These optimizations enabled faster training of the Jais-30B and Jais-30B-chat models.

The above model improvements, coupled with extensive efforts in curating a better pre-training and fine-tuning dataset, allows Jais-30B-chat to demonstrate significant improvements in Arabic and English understanding. Some key highlights over Jais-13B-chat include improved verbosity of answers (+160% in Arabic, +170% in English) and better summarization capabilities (+53% in Arabic, +85% In English).

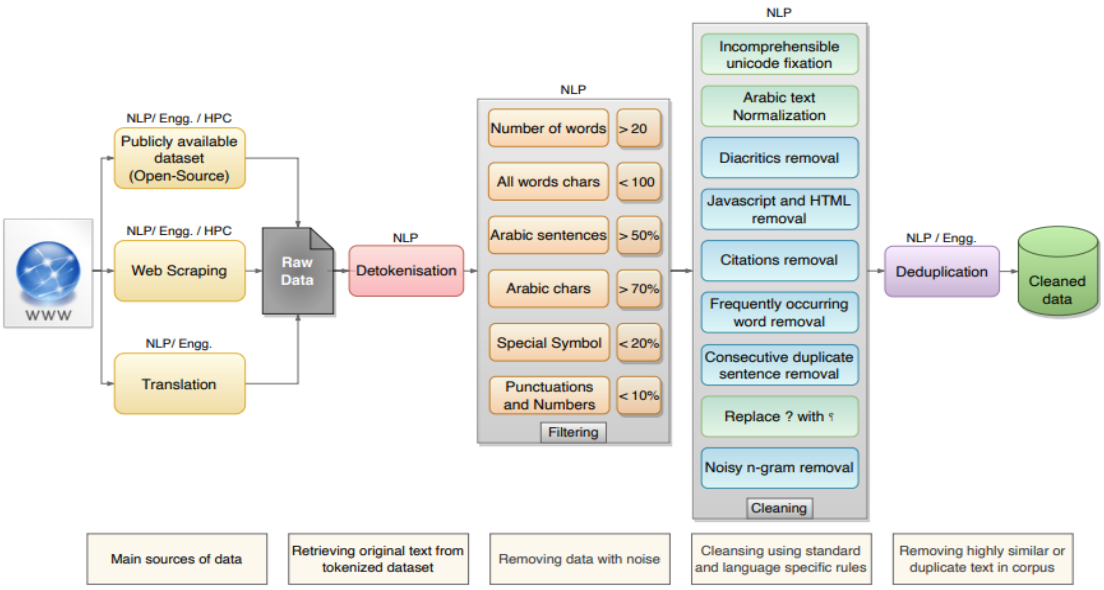
PRETRAINING:

Our training process involved enhancing the Jais 30B model by utilizing a substantial dataset consisting of 126 billion Arabic tokens, 251 billion English tokens, and 50 billion code tokens. This dataset was built upon the pre-training data that was used to train the Jais 13B model. To further augment our dataset, we embarked on an extensive data collection effort, focusing primarily on acquiring Arabic content from the open web.

The process of collecting this data involved web crawling across a variety of sources, most notably the Common Crawl, which is known for archiving web data. We gathered snapshots of data dumps spanning the time frame from 2021 to 2023. Additionally, we placed special emphasis on collecting news articles from the CCNews source, covering a broad range of time from 2018 to 2023, in both English and Arabic languages. News, books, and journals in Arabic/ English from the UAE and the region were also collected and included in the pretraining. This helps to give the Jais models more recent and locally relevant knowledge in both languages.

To ensure the data quality and filter out irrelevant information, we performed extensive processing on the Common Crawl WET files. Our efforts were concentrated on extracting

Arabic documents and disregarding unsuitable content such as brief texts and those with undesirable HTML tags. Once the data collection phase was completed, we proceeded to the initial stage of preprocessing to structure and clean the data, ensuring it was suitable for training purposes. This phase included extensive filtering and cleaning steps depicted below.



Following this, we executed an aggressive de-duplication process on the entire dataset. This de-duplication effort had two primary goals: first, it helped reduce the overall training time, and second, it improved the quality of the dataset by eliminating redundant or repeated information. To carry out this task, we utilized the SlimPajama de-duplication pipeline, which is specifically designed for normalization, cleaning, and deduplication of data.

Phased Pretraining

Jais-30B is planned to be trained on more than 1T tokens. The training is split into four phases, with checkpoints at the end of each phase where we fine-tune, evaluate, and release the model. In this release, we are sharing the model fine-tuned from the end of the first phase, after having seen 427B tokens.

The learning rate schedule for the pretraining is designed to last all 4 phases. Consequently, the checkpoint at the end of phase 1 is taken at a higher than desired learning rate. To make sure we can fine-tune a high-quality model, we forked the training process slightly before the phase 1 checkpoint and lowered the learning rate to 1/10th of its starting value following a warm-up period. This resulted in the checkpoint released as "Jais-30B," while the original training continues into phase 2. Subsequently, the Jais-30B model undergoes further fine-tuning on instructions and is released as Jais-30B-chat.

Phase 1 of the training lasted approximately 45 days on 16 CS2 nodes within the Condor Galaxy supercomputer.

From our experience training Jais-13B and smaller models under a data constrained environment for Arabic, we learnt that a mixture of 1x Arabic, 2x English, and 0.4x Code works well, allowed us to address the data scarcity while maintaining the performance in both Arabic and English. We therefore followed the same data mix in this model training. Moreover, Jais-13B was trained on its English and code pretraining corpus for only one epoch, while at the same time training on its Arabic pretraining corpus for 1.6 epochs. This means each example in the Arabic corpus was seen on an average of 1.6 times by the model. We observed continued performance improvements even as the Arabic content was repeated in the training process (>1 epochs).

Being a larger model, the data scarcity issue is more serious for Jais-30B. To address this, we will extend the above methodology by repeating (most of) the Arabic data during phases 2-4. This approach ensures that while we continue to inject new knowledge and reasoning prowess through abundant English and Code data, we will repeat the Arabic content to maintain the model's capability to generate and understand Arabic.

FINETUNING

Jais-30B is pre-trained to complete the next token over a large text corpus. We fine-tuned this model to better follow instructions by using a dataset of instructions along with the ideal responses. Building on the instruction dataset used by Jais-13B-chat, we covered a wide range of common tasks including question answering, code generation, and reasoning over textual content. To enhance performance in Arabic, we developed an in-house Arabic dataset as well as translating some open-source English instructions into Arabic.

While these open-source datasets improved performance in Arabic and English NLP benchmarks, we wish to further focus on key areas such as longer conversations, and summarization. State of the art open-source models such as Vicuna and WizardLM have demonstrated success with instruction tuning using generations of GPT-3.5 and other larger models. We experimented along this direction by using a dataset of human-prompted GPT-3.5 conversations. These conversations were further augmented with Arabic conversations extracted from other open-source GPT conversations such as Orca.

A typical instruction dataset consists of prompt-response pairs that can vary greatly in length. During training, however, sequence lengths must be the same in a batch. During instruction tuning of a pretrained model, it is common to pad each example till a fixed context length to create batches of data for training. While this ensures that the response tokens in any given example do not attend to unrelated prompts, this can make the fine-tuning process inefficient. As the vast majority of prompt-response

pairs are smaller than the full context length, it results in many pad tokens in the model input.

In Jais-30B-chat, rather than padding short sequences, we packed multiple sequences together into the same example and used an end of sequence token to separate them. This is similar to the packing strategy during pretraining, with a special token to separate documents. However, unlike in pretraining, an instruction response pair is not broken to fill up a sequence till the last token. This means that while packing a sequence if the next available prompt response pair is collectively longer than the number of tokens left in the sequence for packing, the remaining tokens are padded and moved to the next sequence. Another key difference from pretraining is the use of loss masking: the loss at each step is computed only on the response tokens of all examples. The losses on the prompt tokens, and any pad tokens, are masked.

Packing significantly improves the efficiency of training; we observed a speedup of 8x when compared to unpacked training. However, it introduces one confounding factor when compared to packed training. Due to the causal attention masks, the response tokens of examples later in the packed sequence attend to prompts in earlier, unrelated examples. For the attention relationships to be equivalent to the unpacked case, the causal attention masks must be adjusted to not cross the document boundaries. However, we leave this for future work and do not adjust attention masks in this version.

Surprisingly, with packed training, we noted an improvement in performance on downstream benchmarks and GPT-4 based evaluations. This improvement was observed across different datasets and model sizes we experimented with.

Training with packed sequences, coupled with improved conversation data and model scale means Jais-30b-chat performs much better in conversations and handles a wider variety of productive tasks.

BENCHMARKS

Building upon our Jais-13B evaluations, we conducted a comprehensive evaluation of Jais-30B and Jais-30B-chat and benchmarked them against leading language models, focusing on both English and Arabic.

Downstream Evaluation

We follow standard lm-evaluation-harness setup to evaluate models in zero-shot setting for each task and report accuracy. Our evaluation criteria spanned following dimensions:

- Knowledge: How well the model answers factual questions.
- Reasoning: The model's ability to answer questions requiring reasoning.

- Misinformation/Bias: Assessment of the model's susceptibility to generating false or misleading information, and its neutrality.

Arabic Evaluations

Models	Tuned	Knowledge	Commonsense	Misinformation/ Bias	Avg
Jais-30b-chat	tuned	46.6	56.6	53.2	51.7
Jais-13b-chat	tuned	42.1	52.3	50.6	48.4
Jais-30b	-	43	52.1	49.8	47.8
Jais-13b	-	42.9	49.8	49.8	46.5
acegpt-13b-chat	tuned	37.4	48.2	52.1	44.72
BLOOMz (7.1B)	tuned	36.6	44.3	52.1	42.9
acegpt-13b	-	35.6	45.4	50.8	42.54
acegpt-7b	-	36.9	44.7	50.4	42.39
acegpt-7b-chat	tuned	35.3	44.4	52.3	42.23
BLOOM (7.1B)	-	33.1	42.3	49	40.9
mT0-XXL (13B)	tuned	33.1	44.4	44.9	40.9
LLaMA (30B)	-	29.7	40.1	48.5	38.8
LLaMA2 (13B)	-	29.9	40.3	47.7	38.1
LLaMA2-Chat (13B)	tuned	29.5	39.8	47.9	38.1
falcon-40b_instruct	tuned	28.4	39.6	48.4	37.33
llama-30b_instruct	tuned	29.2	39.1	46.3	37.03

English Evaluations

Models	Tuned	Knowledge	Commonsense	Misinformation/ Bias	Avg
falcon-40b_instruct	tuned	43.2	69	63.65	63.35
llama-30b_instruct	tuned	42.8	66.8	56.3	60.49
OPT-30b	-	41.9	65.4	55.7	59.4
Jais-30b-chat	tuned	41.9	65.3	55.3	59.23
acegpt-13b-chat	tuned	38.6	62.9	59.55	57.84
Jais-13b-chat	tuned	39.3	63.7	53.85	57.45
MPT-30b	-	39.3	63.4	53.8	57.3
acegpt-13b	-	37.2	62	56.55	56.51
Jais-30b	-	37.2	62	54.9	56.22
Llama-30b	-	38.6	61.7	50.25	55.4
falcon-40b	-	34.2	61.3	53.3	54.89

acegpt-7b-chat	tuned	35.5	58.2	59.3	54.25
acegpt-7b	-	35.7	59.3	54.75	54.21
Jais-13b	-	34.9	59.5	53.45	53.92

Rows in the tables are arranged in descending order of average score.

Cultural evaluations (UAE datasets)

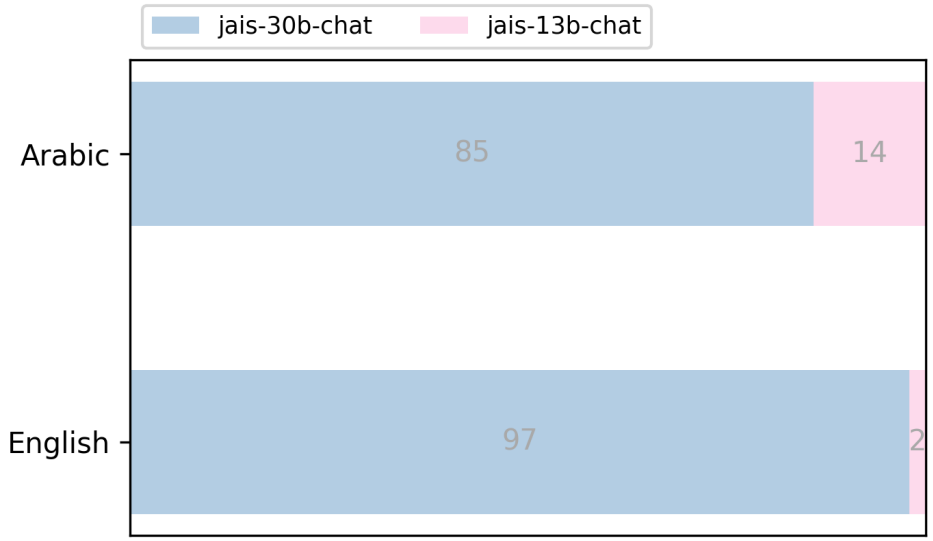
One of the key motivations to train an Arabic LLM is to include knowledge specific to the local context. In training Jais-30B, we have invested considerable effort to include data that reflects high quality knowledge in both languages in the UAE and regional domains. To evaluate the impact of this training,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

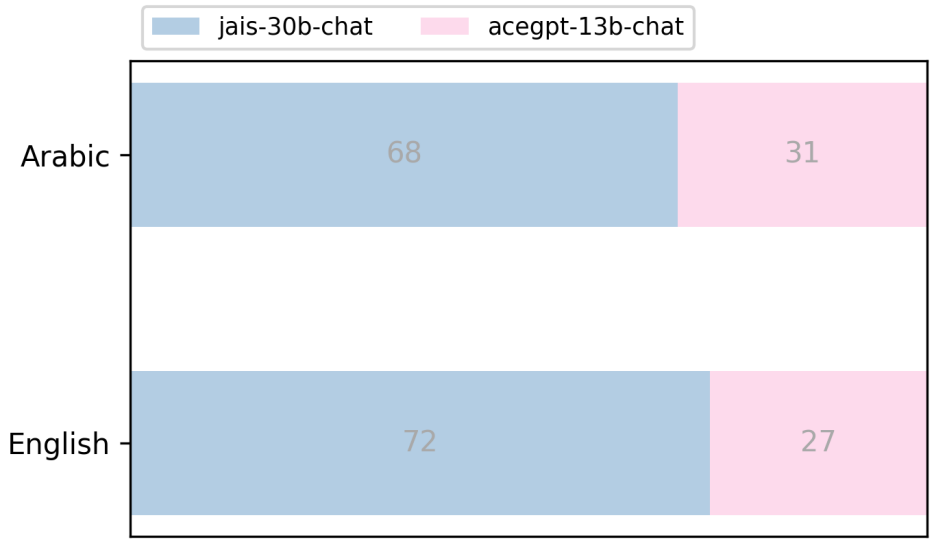
Model	Arabic	English
Jai-30b-chat	62.3	60.4
jais-13b	54.1	45.3
Jais-30b	53.8	48.2
acegpt-7b	53.5	56.8
jais-13b-chat	51.9	55.3
acegpt-13b	51.3	58.9
acegpt-13b-chat	50.9	59.5
acegpt-7b-chat	50.6	53.8

Generation Evaluation (GPT4 evals)

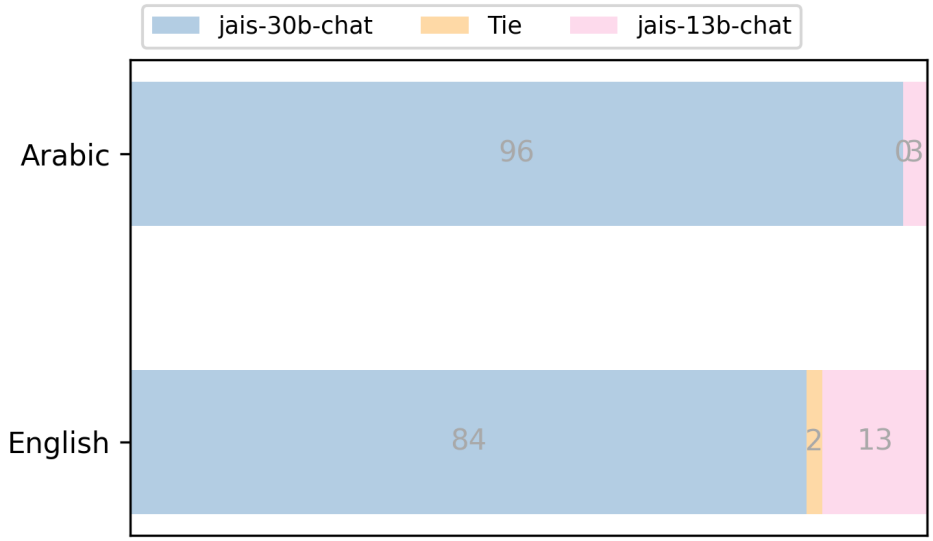
We conducted GPT-4 based evaluations to measure models' ability to follow instructions and respond to prompts across different domains, using Vicuna-80 questions benchmark. We compared Jais-30b-chat against Jais-13b-chat and acegpt-13b-chat, asking GPT-4 to compare and score the generated outputs of a pair of models at a time. Jais-30b-chat has a winning rate of about 85% against Jais-13b-chat and 68% against accept-13b-chat in Arabic.

The following figures show win rate (%) of the models:





Summarization Evaluation



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 the use of 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 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

Responsibility to Ethical AI

The creation of, Jais the world's most powerful Arabic language model, represents the start of a new age in computational linguistics for the Arabic language. It is an inflection point in how individuals, businesses, and governments can and will interact with technology, and each other in the UAE, the Arab region, and globally. Jais' ability to generate human-like text, translate between Arabic and English, answer questions, and even write code, represent a significant leap in artificial intelligence. However, with great power comes great responsibility.

Actions taken

With our team's continued commitment to create more powerful models and improved their capabilities, we cannot stress enough on the importance of ethical considerations and responsible use and deployment of LLMs, in general, and Jais, in particular, and the strategies to mitigate such risks. Indeed, the responsible design and deployment of LLMs must consider the adherence to ethical principles, ensuring that LLM outputs do not propagate biases, preventing discrimination and harm, and managing misinformation, while ensuring the safeguarding of privacy, and the commitment to avoid harm.

Call to action for users of model

Furthermore, while striving to improve Jais' ability to generate persuasive and coherent text, we also aim to reduce the risks of misusing the model such as creating fake news or deepening existing biases. To mitigate the risks associated with the training and fine-

tuning of Jais, the team's utmost priority was to improve the datasets the model was trained and fine-tuned on. While ensuring diverse and well-curated content is used, we conducted regular audits for bias to detect and correct unfair model outputs.

Additionally, the responsible use of these tools requires a commitment to ethical principles, vigilance against misuse, and the establishment of governance structures that ensure transparency and accountability.

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.

Future actions towards AI safety

By adopting a holistic and proactive approach to the development and deployment of LLMs, we can harness their potential while safeguarding our social fabric, maintaining trust in digital ecosystems, and upholding the principles of a just and equitable society.

The Jais models are trained on publicly available data which was in part curated in house at Core42. While efforts have been made to minimize biases, it is likely that the model, as with all LLM models, will exhibit some bias.

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.

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

PARTNERS

Jais-30B and Jais-30B-chat were co-developed by Core42 and Cerebras on the Condor galaxy supercomputer. It is the second milestone in our journey to advance Arabic NLP with large scale bilingual models.

Additionally, we would like to acknowledge the contributions of the Arabic NLP community whose feedback and active participation in the usage of the Jais-13B models helped us to

address several shortcomings and improve our models' outputs. A special acknowledgment to Mohamed bin Zayed University of Artificial Intelligence who supported the training, testing and evaluation of Jais since its initial launch.

Disclaimer: By using Jais, you acknowledge and accept that, as with any large language model, it may generate incorrect, misleading and/or offensive information or content. The information 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 develop models with greater capabilities, and as such, welcome any feedback on the model.