# ALLaM-1-13b-instruct

ALLaM is a series of powerful language models designed to advance Arabic Language Technology (ALT) developed by the National Center for Artificial Intelligence (NCAI) at the Saudi Data and AI Authority (SDAIA). These models are initialized with Llama-2 weights and are trained on both Arabic and English. To enhance their adaptation and performance to the Arabic language, we expand the vocabulary to incorporate a broader range of Arabic words and subwords.

## Intended Use

ALLaM is specifically designed to expedite the research and development of ALT through Large Language Models (LLM). It serves as one of the foundational elements for building product offerings as well as facilitating experimental initiatives.

## Model Details

ALLaM is a family of LLMs specially trained for Arabic. The main two paths followed for pretraining are:

- Continue training from open source models
- Pretraining models from scratch

For this release, we are providing our instruction-tuned 13B parameter generative model initialized from Llama 2 architecture weights.

Some parameters for this model are provided in the following table:

| Size | Context Length | Pretraining Tokens | Instructions | Preference Pairs |
|---|---|---|---|---|
| 13B parameters | 4096 tokens | 2T (from Llama-2) + 1T (En/Ar) | 7M | 260K |

## Model Description

- **Developed by:** National Center for Artificial Intelligence at SDAIA
- **Model type:** Autoregressive Transformer
- **Language(s):** Arabic, English
- **License:** Please see the LICENSE file
- **Input:** Text
- **Output:** Text

## Training Details

ALLaM-1-13b-instruct is pre-trained on a total of 3 trillion tokens in English and Arabic, including the tokens seen from its initialization. The Arabic dataset contains 500 billion tokens after cleaning and deduplication. The additional data is collected from open-source collections and our web crawls.

Our training codebase is built on NVIDIA/MegatronLM. Average MFU during training was ~58%. We trained our model using bf16-mixed precision.

## Getting started

ALLaM model checkpoints weights can be accessed via HuggingFace transformers (tested with `transformers==4.40.1`). The following code snippet demonstrates how to load the model and generate text using the `ALLaM-1-13b-instruct` model.

```
from transformers import AutoModelForCausalLM, AutoTokenizer
allam_model = AutoModelForCausalLM.from_pretrained("ALLaM-1-13b-instruct")
# Replace 'ALLaM-1-13b-instruct' with the model folder path.
tokenizer = AutoTokenizer.from_pretrained("ALLaM-1-13b-instruct") #
Replace 'ALLaM-1-13b-instruct' with the model folder path.
messages=[
    {"role": "user", "content": "كيف أجهز كوب شاهي؟"},
]
inputs = tokenizer.apply_chat_template(messages, tokenize=False)
inputs = tokenizer(inputs, return_tensors='pt',
return_token_type_ids=False)
inputs = {k: v.to('cuda') for k,v in inputs.items()}
allam_model = allam_model.to('cuda')
response = allam_model.generate(**inputs, max_new_tokens=4096,
do_sample=True, top_k=50, top_p=0.95, temperature=.6)
print(tokenizer.batch_decode(response, skip_special_tokens=True)[0])
```

Note that this model is optimized without system prompts.

## Ethical Considerations and Limitations

ALLaM is a generative model that comes with inherent uncertainties. Trials cannot encompass every possible use case. Hence, predicting ALLaM's responses in every context is not possible, leading on occasion to incorrect or biased outputs. Developers must conduct thorough safety evaluations and make specific adjustments to ensure the model is suitable for the intended purposes.

*The output generated by this model is not considered a statement of NCAI, SDAIA, or any other organization.*

## Evaluation

### Automatic Benchmarks

**Massive Multitask Language Understanding** (MMLU) is a collection of many multiple-choice evaluation questions sourced from various academic levels (elementary to college level). These questions are typically related to humanities, STEM, or social sciences. It was originally an English dataset, but other variants were developed for Arabic:

- Original English MMLU (MMLU-en): A collection of 14,079 original English questions spanning 57 domains.

- Translated Arabic MMLU (MMLU-ar-trans): An English to Arabic machine translation of the original English MMLU.
- Natural Arabic MMLU (Arabic MMLU): A collection of 14,575 original Arabic questions spanning 40 domains.

**Exams Arabic** (Exams Ar): A multiple choice question dataset with 537 samples, covering several domains e.g., Islamic studies, science, humanities, and physics.

**Arabic Cultural Alignment** (araCA): This dataset was generated by `gpt-3.5-turbo` and contains 8,710 True and False questions from 58 different areas.

**Education and Training Evaluation Commission** (ETEC): An Arabic multiple choice questions evaluation dataset collected by ALLaM team in collaboration with Saudi ETEC. It covers different levels of education (from elementary to after-college level) with a total of 1,188 test samples. This dataset is not publically available and only accessible to our evaluation team to prevent accidental contamination.

We evaluated all models using our own evaluation pipeline to ensure fair comparison.

| Model | MMLU-en (0-shot) | MMLU-ar-trans (0-shot) | Arabic MMLU (0-shot) | Exams Ar (5-shot) | araCA (5-shot) | ETEC (0-shot) |
|---|---|---|---|---|---|---|
| Llama2 13B chat | 53.8 | 28.7 | 35.8 | 22.9 | 60.1 | 30.4 |
| AceGPT 13B chat | 54.6 | 37.2 | 52.6 | 42.6 | 67.7 | 37.3 |
| Jais 13B chat | 50.5 | 41.0 | 54.8 | 46.9 | 70.7 | 48.7 |
| Jais 30B chat (v1) | 54.5 | 44.0 | 60.4 | 48.6 | 71.1 | 48.5 |
| Jais 30B chat (v3) | 59.1 | 30.2 | 62.3 | 51.2 | 70.0 | 38.5 |
| ALLaM-1-13b instruct | **63.4** | **51.0** | **68.1** | **54.9** | **78.6** | **65.6** |

## MT-bench

**Multi-turn bench** (MT-bench): A challenging multi-turn benchmark that uses GPT-4 as a judge. MT-bench comprises 80 questions from 8 domains. Each question is presented to the model and the responses are submitted to GPT-4 to assign scores to each response. The judge returns a score for the first and second turn separately.

This dataset was automatically translated to Arabic and manually verified and culturally aligned.

| Model | AR Average | AR Turn 1 | AR Turn 2 | EN Average | EN Turn 1 | EN Turn 2 |
|---|---|---|---|---|---|---|
| AceGPT 13B chat | 5.53 | 6.76 | 4.12 | 6.36 | 7.01 | 5.64 |
| Jais 13B chat | 4.4 | 4.77 | 3.96 | 4.71 | 5.07 | 4.36 |
| Jais 30B chat (v1) | 3.57 | 4.13 | 3.64 | 3.57 | 4.13 | 2.95 |
| Jais 30B chat (v3) | 5.92 | 6.25 | 5.47 | 6.28 | 6.78 | 5.78 |
| ALLaM-1-13b instruct | **7.38** | **7.67** | **7.01** | **7.57** | **7.9** | **7.23** |

## Citation

If you found this work helpful or used any part of this work, please include the following citation:

```
@misc{allam2024,
  title = "ALLaM: A Series of Large Language Models for Arabic and
English",
  author = "{NCAI, SDAIA}",
  howpublished = "\url{https://eu-
de.dataplatform.cloud.ibm.com/wx/samples/models/sdaia/allam-1-13b-
instruct?context=wx}",
  year = 2024,
}
```