Bengali and Arabic instruction tuning datasets for continual learning

Both Bengali and Arabic have strong, verified instruction-tuning datasets on Hugging Face suitable for continual learning experiments. The best Bengali option is `md-nishat-008/Bangla-Instruct` — a native, non-translated 100K-sample dataset backed by an ACL 2025 paper. For Arabic, `ClusterlabAi/InstAr-500k` leads with 481K samples and rich metadata, while `arbml/CIDAR` offers 10K human-reviewed, culturally aligned samples GitHub published at ACL 2024. Hugging Face Multilingual datasets like Bactrian-X arXiv Hugging Face and Aya provide parallel coverage across both languages with standardized formats, making them ideal for controlled cross-lingual continual learning experiments.

Bengali datasets ranked by quality and suitability

1. md-nishat-008/Bangla-Instruct ⭐ Top Pick

The strongest Bengali instruction dataset available. **100K native Bengali instruction-response pairs** generated through a self-instruct framework using GPT-4 and Claude-3.5-Sonnet as teacher models. 500 seed tasks were curated by 50 volunteers from Bangladeshi universities, Hugging Face making this genuinely native Bengali rather than translated.

- **HF path:** `md-nishat-008/Bangla-Instruct`
- **Samples:** ~100,000 (HF shows ~342K rows, likely including augmented data)
- **Columns:** `instruction` (str), `response` (str)
- **License:** MIT
- **Paper:** TigerLLM (ACL 2025 Short Paper, arXiv:2503.10995) Hugging Face
- **Last updated:** ~December 2025
- **Loading:**

```python
from datasets import load_dataset
ds = load_dataset("md-nishat-008/Bangla-Instruct")
```

- **Quirks:** Column is `response` not `output` — rename needed for standard Alpaca-style pipelines. The `response` column may contain occasional null values. No separate `input` field. Dataset description says 100K but actual row count is higher, suggesting data augmentation or versioning changes.

2. iamshnoo/alpaca-cleaned-bengali ⭐ Most Widely Used

The most popular Bengali instruction dataset with **85 likes** on HuggingFace. A Bengali translation of `yahma/alpaca-cleaned` using NLLB-1.3B. Hugging Face Standard Alpaca three-column format makes it plug-and-play for most training pipelines.

- **HF path:** `iamshnoo/alpaca-cleaned-bengali`
- **Samples:** ~51,800 (train split)
- **Columns:** `instruction` (str), `input` (str), `output` (str)
- **License:** CC BY-NC 4.0 (inherited from Alpaca)
- **Paper:** Used by multiple Bengali LLMs including Mistral-Instruct-Bangla, Hugging Face llama-3-8b-bangla, and asif00/bangla-llama
- **Last updated:** September 2023
- **Loading:**

```python
ds = load_dataset("iamshnoo/alpaca-cleaned-bengali")
```

- **Quirks:** Machine-translated via NLLB-1.3B, so complex instructions may have uneven quality. Hugging Face ~60% of entries have empty `input` fields (same as original Alpaca). Non-commercial license restricts some use cases.

3. OdiaGenAI/all_combined_bengali_252k

The **largest Bengali instruction dataset at 252K samples**, aggregating translations from Dolly, Alpaca, ChatDoctor, Roleplay, and GSM datasets. GitHub The `data_source` column enables domain-specific filtering — particularly useful for continual learning experiments that need to control for task type.

- **HF path:** `OdiaGenAI/all_combined_bengali_252k`
- **Samples:** ~252,622 Hugging Face huggingface
- **Columns:** `instruction` (str), `input` (str), `output` (str), `data_source` (str, 11 values)
- **License:** CC BY-NC-SA 4.0 huggingface
- **Last updated:** October 2024
- **Loading:**

```python
```

```
ds = load_dataset("OdiaGenAI/all_combined_bengali_252k")
# Filter by domain:
medical = ds["train"].filter(lambda x: x["data_source"] == "ChatDoctor")
```

- **Quirks:** All content is translated, not native Bengali. Includes medical and math domains. The 309MB download is manageable. Overlap with other Alpaca-based Bengali datasets exists.

4. BanglaLLM/bangla-alpaca

From the BongLLaMA project, this dataset adds a pre-formatted `text` column ready for direct SFT training, saving preprocessing time.

- **HF path:** `BanglaLLM/bangla-alpaca`
- **Samples:** ~51,800
- **Columns:** `instruction`, `input`, `output`, `text` (pre-formatted prompt), `system_prompt`
- **License:** GPL-3.0
- **Paper:** BongLLaMA (2024)
- **Loading:**

```python
ds = load_dataset("BanglaLLM/bangla-alpaca")
```

- **Quirks:** GPL-3.0 is more restrictive than MIT or Apache. The `text` column contains a complete prompt template.

5. saillab/alpaca-bengali-cleaned

Notable for having **a dedicated test split** (41.6K train + 10.4K test), useful for evaluation without needing to create your own split. Translated using Google Translate rather than NLLB.

- **HF path:** `saillab/alpaca-bengali-cleaned`
- **Samples:** ~52,000 total
- **Columns:** `instruction`, `input`, `output`
- **License:** CC BY-NC
- **Paper:** Used in the TaCo paper
- **Loading:**

```python
ds = load_dataset("saillab/alpaca-bengali-cleaned")
```

- **Quirks:** `input` field may contain literal "nan" strings instead of empty strings. Includes both Alpaca-52K and Dolly-15K translations.

Arabic datasets ranked by quality and suitability

1. ClusterlabAi/InstAr-500k ⭐ Top Pick for Scale

The **largest pure-Arabic instruction dataset** with 481K samples and exceptionally rich metadata. Each entry includes `source`, `task`, `type`, `topic`, and `system` columns — enabling fine-grained filtering by domain, task type, or data provenance. Generated using Command R+ from the 101 Billion Arabic Words Dataset. arXiv

- **HF path:** `ClusterlabAi/InstAr-500k`
- **Samples:** ~481,000
- **Columns:** `uuid`, `source` (24 values), `task` (11 types: Open QA, Summarization, Extraction, etc.), `type` ("human-crafted" or "synthetic"), `topic` (10 categories), `system` (Arabic system prompts), `instruction`, `output` huggingface
- **License:** Apache 2.0
- **Paper:** GemmAr (arXiv:2407.02147)
- **Loading:**

```python
ds = load_dataset("ClusterlabAi/InstAr-500k")
# Filter by task type:
qa_subset = ds["train"].filter(lambda x: x["task"] == "Open QA")
# Filter by data type:
human_only = ds["train"].filter(lambda x: x["type"] == "human-crafted")
```

- **Quirks:** Mix of synthetic (Command R+) and human-crafted data. System prompts are pre-written in Arabic. The rich metadata makes this particularly suited for controlled continual learning experiments.

2. arbml/CIDAR ⭐ Top Pick for Quality

10,000 human-reviewed, culturally relevant Arabic instructions published at ACL 2024. ACL Anthology

Created from ~9,109 Alpagasus samples translated via ChatGPT plus 891 native Arabic grammar instructions from the "Ask the Teacher" website. All 10K samples were reviewed by 12 Arabic-speaking annotators. Hugging Face +2

- **HF path:** `arbml/CIDAR`
- **Samples:** 10,000
- **Columns:** `instruction` (str), `output` (str), `index` (int)
- **License:** Apache 2.0 (HF metadata) / CC BY-NC 4.0 (README states non-commercial — **verify before commercial use**)
- **Paper:** CIDAR (ACL 2024 Findings, arXiv:2402.03177) GitHub
- **Last updated:** January 2026
- **Loading:**

python

```python
ds = load_dataset("arbml/CIDAR")
```

- **Quirks:** License is inconsistently listed — the HF page tags say Apache 2.0, but the dataset card explicitly states CC BY-NC 4.0. Hugging Face Small size (10K) makes it ideal as a high-quality fine-tuning or validation set rather than large-scale training. Contains genuinely Arabic-native grammar instructions alongside translated content. Hugging Face

3. FreedomIntelligence/alpaca-gpt4-arabic

50K instruction pairs where both translation and response generation used GPT-4, resulting in higher quality than GPT-3.5 or NLLB-based alternatives. Part of the AceGPT project for Arabic LLMs.

- **HF path:** `FreedomIntelligence/alpaca-gpt4-arabic`
- **Samples:** ~50,000
- **Columns:** `id` (str), `conversations` (list of dicts: `{"from": "human"/"gpt", "value": "..."}`)
- **License:** Apache 2.0
- **Paper:** AceGPT (2024)
- **Loading:**

python

```
ds = load_dataset("FreedomIntelligence/alpaca-gpt4-arabic")
# Convert ShareGPT format to instruction/output:
def convert(example):
    convs = example["conversations"]
    return {"instruction": convs[0]["value"], "output": convs[1]["value"]}
ds = ds.map(convert)
```

- **Quirks:** Uses ShareGPT conversation format (not Alpaca format) — requires conversion. Typically single-turn (one human message + one assistant response). A sibling dataset `FreedomIntelligence/Alpaca-Arabic-GPT4` exists but appears to be a near-duplicate.

4. arbml/alpaca_arabic

Standard Alpaca-52K translated to Arabic with **parallel English-Arabic columns** — uniquely useful for bilingual experiments or translation quality verification.

- **HF path:** `arbml/alpaca_arabic`
- **Samples:** 52,002 huggingface
- **Columns:** `instruction_en`, `input_en`, `output_en`, `instruction` (Arabic), `input` (Arabic), `output` (Arabic), `index` huggingface
- **License:** CC BY-NC 4.0 (inherited)
- **Loading:**

```python
ds = load_dataset("arbml/alpaca_arabic")
```

- **Quirks:** No dataset card documentation. Parallel columns are useful for analyzing translation quality but add unnecessary size if you only need Arabic.

5. MoMonir/CohereForAI_aya_dataset_Arabic

A pre-filtered extract of the **fully human-annotated Aya dataset**, containing ~14.2K Arabic samples. The highest per-sample quality of any Arabic instruction dataset since every entry was written or verified by native speakers. Hugging Face

- **HF path:** `MoMonir/CohereForAI_aya_dataset_Arabic`
- **Samples:** ~14,200 (14K train + 250 test) Hugging Face
- **Columns:** `inputs`, `targets`, `language`, `script`, `dataset_name`

- **License:** Apache 2.0
- **Paper:** Aya Dataset (arXiv:2402.06619, ACL 2024)
- **Loading:**

```python
ds = load_dataset("MoMonir/CohereForAI_aya_dataset_Arabic")
```

- **Quirks:** Column names are `inputs`/`targets` rather than `instruction`/`output`. Includes MSA and some dialectal Arabic. Pre-filtered convenience wrapper saves downloading the full 204K multilingual Aya dataset.

Multilingual datasets covering both Bengali and Arabic

These datasets provide **parallel coverage across dozens of languages with identical instruction sources**, making them ideal for controlled continual learning experiments where you need to isolate the effect of language.

MBZUAI/Bactrian-X

67K instruction-response pairs per language across 52 languages. Instructions translated from Alpaca+Dolly via Google Translate; responses generated by GPT-3.5-turbo in each target language. GitHub +2 The consistent structure across all 52 languages makes this the single best dataset for controlled cross-lingual CL experiments.

- **HF path:** `MBZUAI/Bactrian-X`
- **Bengali subset:** `load_dataset("MBZUAI/Bactrian-X", "bn")`
- **Arabic subset:** `load_dataset("MBZUAI/Bactrian-X", "ar")`
- **Samples per language:** ~67,000
- **Columns:** `instruction`, `input`, `output`, `id` Hugging Face
- **License:** CC BY-NC 4.0
- **Paper:** arXiv:2305.15011 (2023) Hugging Face Hugging Face

CohereForAI/aya_dataset

204K human-curated instances across 65 languages. Hugging Face Hugging Face Only ~2% machine-translated; the vast majority are original annotations by native speakers. arXiv Bengali and Arabic both included. The gold standard for quality.

- **HF path:** `CohereForAI/aya_dataset`

- **Bengali:** filter by `language == "ben"`
- **Arabic:** filter by `language == "arb"` (also `arz`, `ary` for dialects)
- **Columns:** `inputs`, `targets`, `language`, `script`, `dataset_provenance`
- **License:** Apache 2.0
- **Paper:** arXiv:2402.06619 (ACL 2024)

bigscience/xP3

~81M prompt-completion pairs across 46 languages, arXiv used to train BLOOMZ and mT0. Hugging Face Task-oriented (NLI, QA, summarization) rather than free-form instruction following. Hugging Face

- **HF path:** `bigscience/xP3`
- **Bengali:** `load_dataset("bigscience/xP3", "bn", trust_remote_code=True)`
- **Arabic:** `load_dataset("bigscience/xP3", "ar", trust_remote_code=True)`
- **Columns:** `inputs`, `targets`
- **License:** Apache 2.0
- **Paper:** arXiv:2211.01786 (ACL 2023) Hugging Face arXiv

Papers backing these datasets in continual learning contexts

Five key papers from 2023–2025 directly validate using these datasets for continual learning and multilingual instruction tuning. **TigerLLM** (ACL 2025) demonstrated continual pretraining + instruction fine-tuning using Bangla-Instruct, achieving state-of-the-art Bengali performance. Hugging Face **CIDAR** (ACL 2024) showed that models fine-tuned on just 10K culturally relevant Arabic instructions outperformed models trained on 30× more generic data. ACL Anthology **GemmAr** (2024) used InstAr-500k to fine-tune Gemma-7B and Llama-8B via LoRA. arXiv **Bactrian-X** (2023) trained LoRA adapters on LLaMA and BLOOM across 52 languages. Hugging Face Hugging Face **EMMA-500** (arXiv:2409.17892) performed continual pre-training of Llama 2 7B across 546 languages, GitHub directly demonstrating the continual learning paradigm these datasets support. arXiv Cool Papers

Two additional papers are especially relevant for the continual learning framing: **TL-CL** (EMNLP 2024) studies task and language incremental continual learning, GitHub and **CrossAlpaca** showed that combining instruction and translation data during continual adaptation significantly improves multilingual QA performance. OpenReview

---

Alternative low-resource languages if needed

Arabic is **not** limited — it has excellent coverage arXiv with InstAr-500k (481K), CIDAR (10K), and alpaca-gpt4-arabic (50K). However, if additional low-resource languages are needed for broader multilingual continual learning experiments, these alternatives have strong instruction datasets:

- **Hindi:** ai4bharat/indic-instruct-data-v0.1 — **385K samples** across 8 diverse subsets (Dolly, Flan, OASST, etc.), conversational format, backed by the Airavata paper Hugging Face (arXiv:2401.15006). arXiv The best-resourced Indic language for instruction tuning.
- **Tamil:** abhinand/tamil-alpaca — ~52K samples, standard Alpaca format, used in Tamil-LLaMA (arXiv:2311.05845). Hugging Face
- **Vietnamese:** Active ecosystem with Bactrian-X subset (65K) plus PhoGPT-related datasets. Multiple models available. GitHub
- **Thai:** Available through Bactrian-X (65K), Aya, and the Sailor/SeaLLM projects for Southeast Asian languages. GitHub
- **Swahili:** Most limited standalone resources; best obtained via Bactrian-X (~65K) or Aya subsets.

Recommended experimental configuration

For a continual learning experiment comparing language adaptation across Bengali and Arabic, the following configuration balances quality, scale, and format consistency:

Role	Bengali Dataset	Arabic Dataset	Samples
Primary training	md-nishat-008/Bangla-Instruct	ClusterlabAi/InstAr-500k	100K / 481K
Controlled comparison	MBZUAI/Bactrian-X (bn)	MBZUAI/Bactrian-X (ar)	67K / 67K
High-quality evaluation	CohereForAI/aya_dataset (ben)	arbml/CIDAR	~3K / 10K
Fallback / Alpaca-format	iamshnoo/alpaca-cleaned-bengali	FreedomIntelligence/alpaca-gpt4-arabic	52K / 50K

Using Bactrian-X for both languages ensures identical instruction sources, isolating language as the only variable. The primary training sets provide scale and native-language quality. Aya and CIDAR serve as high-quality evaluation benchmarks with human verification. All datasets load cleanly with datasets.load_dataset() and have permissive licenses (MIT or Apache 2.0) except where noted as CC BY-NC.