Hala Technical Report
Building Arabic-Centric Instruction & Translation Models at Scale

Hasan Abed Al Kader Hammoud[1,*], **Mohamad Zbib**[1,*], **Bernard Ghanem**[1]

[1]King Abdullah University of Science and Technology (KAUST)

[*]Equal contribution.

Correspondence: hasanabedalkader.hammoud@kaust.edu.sa, mohamad.zbib@kaust.edu.sa

Abstract

We present HALA, a family of *Arabic-centric* instruction and translation models built with our translate-and-tune pipeline. We first compress a strong AR↔EN teacher to FP8 (yielding ∼2× higher throughput with no quality loss) and use it to create high-fidelity bilingual supervision. A lightweight language model `LFM2-1.2B` is then fine-tuned on this data and used to translate high-quality English instruction sets into Arabic, producing a million-scale corpus tailored to instruction following. We train HALA models at 350M, 700M, 1.2B, and 9B parameters, and apply `slerp` merging to balance Arabic specialization with base-model strengths. On Arabic-centric benchmarks, HALA achieves state-of-the-art results within both the "nano" (≤2B) and "small" (7–9B) categories, outperforming their bases. We are committed to release models, data, evaluation, and recipes to accelerate research in Arabic NLP. HALA models and all associated datasets are publicly released on Hugging Face.[1]

In Arabic, Hala, conveys sweetness and beauty - qualities long associated with the language itself. In this spirit, we call our models Hala.

1 Introduction

Large language models (LLMs) have rapidly advanced the state-of-the-art across general-purpose NLP, demonstrating strong capabilities in few-shot learning, instruction following, and multistep reasoning. Early milestones such as GPT-3 (Brown et al., 2020) catalyzed this progress, while more recent families (e.g., Gemini (Team et al., 2023), Claude 3) continue to expand the frontier of capability and reliability. Open-weight counterparts, including DeepSeek (Liu et al., 2024), LLaMA 3 (Grattafiori et al., 2024), Qwen (Yang et al., 2025), Gemma (Gemma Team et al., 2025), and Kimi K2 (Team et al., 2025b), have enabled broad experimentation and downstream applications, accelerating community research into scaling, alignment, and efficient deployment.

Multilingual modeling at scale. Alongside raw capability, a major thrust in recent work targets *multilinguality*: building models and resources that operate across many languages. Dataset efforts range from broad-coverage sentence-aligned corpora such as Tatoeba (Tiedemann, 2020) to large-scale conversational resources such as MASSIVE (FitzGerald et al., 2023). Engineering pipelines (e.g., warc2text extraction and parallel translation) have been used to derive multilingual corpora from web archives (de Gibert et al., 2024). Beyond data, analyses probe whether models preserve knowledge and answer consistency across languages (Ifergan et al., 2024). Model design has also embraced multilinguality from the ground up: BLOOM (Workshop et al., 2022) supports 46 languages, while Baichuan-2 (Yang et al., 2023) and other families emphasize improved performance on non-English tasks. Despite this breadth-first progress, per-language depth and cultural alignment remain uneven, especially for underrepresented languages.

Arabic LLMs and the instruction-data bottleneck. Arabic poses distinct challenges due to diglossia, rich morphology, and wide dialectal variation. A growing line of Arabic-centric work (Al-Khalifa et al., 2025) spans monolingual pretraining (e.g. AraBERT (Antoun et al., 2020)), foundation and chat models (e.g. JAIS (Sengupta et al., 2023), FANAR (Team et al., 2025a), PEACOCK (Alwajih et al., 2024), ACE-GPT (Huang et al., 2024), ALLAM (Bari et al., 2024)) and broader sovereign AI efforts such as Falcon (Almazrouei et al., 2023a). Benchmarks, including Arabic-MMLU (Koto et al., 2024) , provide initial

[1]https://huggingface.co/collections/hammh0a/hala

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 236–244
March 28, 2026.

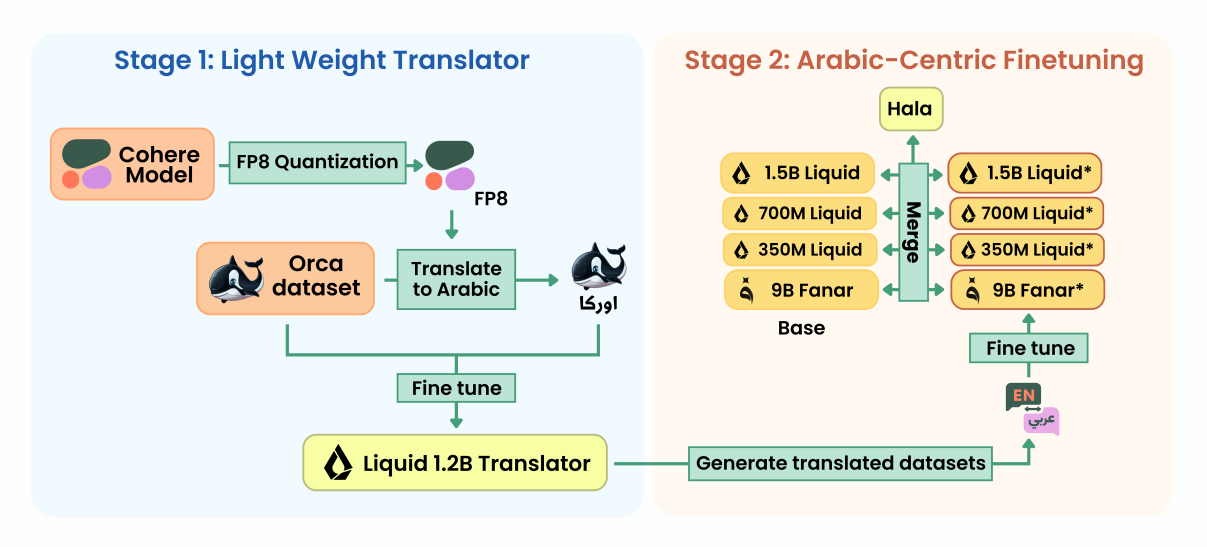

Figure 1: Cross-lingual translation and fine-tuning pipeline for Liquid 1.2B. In the teacher phase, the Cohere model with FP8 inference is used to translate the Orca dataset, which is then used to fine-tune Liquid 1.2B. In the bootstrapped translator phase, Liquid 1.2B translates datasets, producing group of arabic dataset. Liquid models and FANAR were then further fine-tuned on the combined translated datasets, yielding the final instruction-tuned models.

evaluation scaffolding, although coverage and difficulty remain limited relative to English. A persistent bottleneck is the scarcity of *high-quality Arabic instruction data*, which constrains both instruction tuning and scaling. Previous works document the underrepresentation of non-English languages in pretraining corpora and their impact on downstream performance (Lin et al., 2022; Xue et al., 2021; Touvron et al., 2023). In parallel, the community has explored the paradigms of 'AI trains AI', e.g., self-instruction and synthetic supervision, to overcome data scarcity (Xu et al., 2023; Mukherjee et al., 2023; Achiam et al., 2023; Wang et al., 2023). However, in Arabic, the volume and fidelity of the instruction data still lag behind.

Language-centric vs. multilingual. We adopt the term *language-centric* to denote models whose *primary optimization target* is depth of capability in a specific language (here, Arabic), rather than uniform breadth across many languages. A language-centric approach can better capture linguistic nuance (e.g. morphology, orthography)(Conneau et al., 2020), dialectal variation, and cultural/safety alignment, while still benefiting from cross-lingual transfer when appropriate. In practice, this requires (i) reliable translation pipelines to convert strong English supervision into Arabic *without* eroding instruction fidelity, and (ii) training strategies that scale across model sizes while preserving Arabic fluency and task competence.

LLMs as translators: opportunities for Arabic data bootstrapping. LLMs have recently emerged as strong machine translation engines (Lyu et al., 2024), capable of long-document and stylistic translation, interactive workflows, and even domain-preserving scientific translation (Kleidermacher and Zou, 2025). Creative strategies, such as searching for keywords / topics with multiple generations of candidates and selection (He et al., 2024), further improve quality. Broad evaluations in 120+ languages (Zhu et al., 2024) suggest that carefully managed LLMs can serve as reliable translators. These developments make *translation-first* bootstrapping especially attractive for Arabic instruction tuning: If we can (1) compress a capable translator for efficient, scalable inference and (2) preserve instruction semantics during translation, we can unlock large Arabic corpora suitable for high-quality tuning.

Our approach and contributions. In this report, we introduce HALA, a family of Arabic *language-centric* instruction and translation models built around an efficient translate-and-tune pipeline. Our contributions are as follows:

- **Lightweight AR↔EN translator.** We compress a strong multilingual translator to FP8 with dynamic scaling using *LLM Compressor*

(AI and vLLM Project, 2024) and fine-tune `LiquidAI/LFM2-1.2B` to serve as a fast, robust AR↔EN engine. This translator is used to construct Arabic instruction data at scale while maintaining fidelity to the source instructions.

- **Million-scale bilingual supervision.** We build a 1.25M AR↔EN bilingual corpus by pairing translated and original texts (e.g., from Open-Orca (Mukherjee et al., 2023)) and a filtered subset of OPUS-100 (Zhang et al., 2020), enabling stable training of lightweight translation models and consistency checks.

- **Large Arabic instruction corpus.** Using our translation stack, we convert several high-quality English instruction datasets into Arabic, including Hermes 3 (Teknium et al., 2024), SCP-116K (Lu et al., 2025), ReAlign-Alpaca (Fan et al., 2024), LaMini (Wu et al., 2024), Tulu 3 (Lambert et al., 2024), and Synthetic Instruct-GPT-J Pairwise (Havrilla, 2023), alongside Open-Orca (Mukherjee et al., 2023). The resulting Arabic corpus (millions of pairs) emphasizes instruction following, reasoning, and alignment.

- **Arabic-centric models across scales.** We release HALA models at 350M, 700M, and 1.2B parameters (based on Liquid checkpoints) as well as a 9B model built on the FANAR architecture (Team et al., 2025a). To combine complementary strengths from English- and Arabic-tuned checkpoints, we employ *MergeKit* (Goddard et al., 2024) with spherical linear interpolation.

- **Open releases and recipes.** We release models, data, and training/evaluation scripts to facilitate reproducibility and further research on Arabic instruction tuning.

Summary. By coupling an efficient AR↔EN translator with million-scale data construction, HALA advances Arabic instruction tuning under constrained compute budgets. Our results (Section 3) indicate that HALA models achieve competitive performance within their parameter classes on Arabic-centric benchmarks (Koto et al., 2024), supporting the view that *language-centric* modeling is a practical and effective complement to breadth-first multilingual scaling.

2 Methodology

2.1 Quantizing the main translator to FP8

We begin with a high-capacity multilingual translator (`CohereLabs/command-a-translate-08-2025`) and compress it to FP8 (Kuzmin et al., 2022) with *dynamic scaling* using *LLM Compressor* (AI and vLLM Project, 2024), releasing the FP8 artifact as `our/command-a-translate-FP8-Dynamic`. The FP8 conversion reduces memory footprint and improves inference throughput (empirically $\approx 2\times$ faster than the non-quantized counterpart) while preserving translation quality on our evaluation sets. We follow the official `llm-compressor` recipe (per-tensor dynamic scaling and post-conversion validation) to ensure stability.

2.2 Bootstrapping bilingual supervision from Open-Orca

To construct high-quality AR↔EN supervision aligned with instruction-tuning style data, we translate the *first 405K* instruction–response pairs from `Open-Orca/OpenOrca` (Mukherjee et al., 2023) *into Arabic*, covering both the user questions and assistant responses. The quantized FP8 translator is prompted with a minimal instruction:

```
Translate from English to Arabic: {x}
```

For each example, we *pair* the Arabic translations with their original English counterparts, yielding bilingual tuples of the form $\langle \text{instr}_{\text{en}}, \text{instr}_{\text{ar}}, \text{resp}_{\text{en}}, \text{resp}_{\text{ar}} \rangle$. This produces an instruction-focused bilingual set mirroring the semantics and difficulty of `Open-Orca`, with substantial coverage of reasoning-heavy queries.

2.3 Quality filtering of OPUS-100 with a strict bilingual judge

We augment the above with a large parallel corpus drawn from the `Helsinki-NLP/opus-100` (Zhang et al., 2020) `ar-en` subset. From 1M candidate pairs, we filter for fidelity using a compact judge model (`Qwen2.5-3B-Instruct` (Yang et al., 2025)) prompted to emit a binary verdict (`accept`/`reject`):

```
prompt = f"""
You are a strict bilingual judge.
You will be given a translation pair.
Arabic: {ar_text}
English: {en_text}
```

```
If the English is a correct and natural
translation of the Arabic, output only:
accept
Otherwise, output only:
reject
"""
```

Pairs marked `accept` are retained; this procedure yields **439,592** accepted pairs out of ∼1M candidates, providing a clean AR↔EN signal complementary to the instruction-style data above.

2.4 Training a lightweight AR↔EN translator

We combine the translated `Open-Orca` set (405K × 2 = 810K) with the filtered `OPUS-100` pairs (440K), totaling ∼1.26M bilingual examples, and fine-tune `LiquidAI/LFM2-1.2B` into a fast, stable AR→EN translator specialized for instruction-style inputs (instructions and responses). We use simple chat-style prompting during training (for E→A) and standard supervised fine-tuning with cross-entropy. This lightweight translator serves as *workhorse* for the construction of large-scale Arabic data in the next stage.

2.5 Building the Arabic instruction corpus via translation

Using the above translator, we convert multiple high-quality English instruction datasets into Arabic, preserving formatting and answer style:

- **Open-Orca/OpenOrca** (Mukherjee et al., 2023): 405K (first subset), covering multi-step, reasoning-heavy queries.
- **NousResearch/Hermes-3-Dataset** (Teknium et al., 2024): filtered to remove all code-related samples to avoid translation artifacts.
- **EricLu/SCP-116K** (Lu et al., 2025): instructional and conversational pairs.
- **GAIR/ReAlign-Alpaca** (Fan et al., 2024): realigned version of Alpaca instructions.
- **Dahoas/synthetic-instruct-gptj-pairwise** (Havrilla, 2023): synthetic paired preference-style instructions.
- **MBZUAI/LaMini-instruction** (Wu et al., 2024): lightweight instruction data, translated fully.
- **allenai/tulu-3-sft-mixture** (Lambert et al., 2024): we keep only English subsets and translate them.

The resulting corpus emphasizes instruction following, reasoning, and alignment, providing broad coverage for Arabic-centric instruction tuning. We collect a total of roughly 4.5M samples.

2.6 Arabic instruction fine-tunes and model merging

We fine-tune models across scales on the translated Arabic instruction mix, then apply merging to balance Arabic gains with base-model strengths:

- **350M**: fine-tune `LiquidAI/LFM2-350M`, then merge with its base to obtain HALA-350M.
- **700M**: fine-tune `LiquidAI/LFM2-700M`, then merge with its base to obtain HALA-700M.
- **1.2B**: fine-tune `LiquidAI/LFM2-1.2B`, then merge with its base to obtain HALA-1.2B.
- **9B**: fine-tune on top of `QCRI/Fanar-1-9B-Instruct` (Team et al., 2025a), then merge with its base to obtain HALA-9B.

Merging is performed with *MergeKit* (Goddard et al., 2024) using spherical linear interpolation (`slerp`) at t=0.5, which we found to preserve general capability while boosting Arabic instruction-following performance. The overall translate–and–tune pipeline is illustrated in Fig. 1.

3 Evaluation

Benchmarks and protocol. We evaluate on a suite of Arabic-centric tasks following the *Open-Arabic-LLM-Leaderboard (OALL)* task selection where feasible. Concretely, we report results on: **AlGhafa** (Almazrouei et al., 2023b), **AraTrust** (Alghamdi et al., 2024), **ArabicMMLU** (Koto et al., 2024), **ArbMMLU-HT** (Koto et al., 2024), **EXAMS** (Hardalov et al., 2020), and **MadinahQA** (Koto et al., 2024). We *exclude* **Alrage** (present in some OALL variants) because it requires an LLM-as-a-judge setup. All evaluations are conducted with `LightEval` (Habib et al., 2023) using `vLLM` (Kwon et al., 2023) as the backend for efficient, reproducible inference. We will release exact *LightEval* command lines, task definitions, and configuration files in the accompanying GitHub repository.

Table 1: Arabic-centric evaluation across six benchmarks following the OALL task suite (excluding *Alrage*); higher is better. Columns 4–9 report task scores (%). **Average** is the unweighted mean across the six tasks. Best **Average** within each size bucket is **bold**; second-best is underlined. All runs use `LightEval` with `vLLM`; exact commands are released in the repo.

Size	Model Name	Params	Arabic-centric Benchmarks (%)						Average
			AlGhafa	ArabicMMLU	EXAMS	MadinahQA	AraTrust	ArbMMLU-HT	
Nano (≤2B)									
≤2B	meta-llama/Llama-3.2-1B	1B	33.9	26.5	21.2	25.7	37.1	23.9	28.0
≤2B	Qwen/Qwen2-1.5B-Instruct	1.5B	53.1	49.2	35.2	45.5	68.9	37.4	48.2
≤2B	Qwen/Qwen2.5-1.5B-Instruct	1.5B	48.4	43.5	31.8	38.2	70.8	35.9	44.8
≤2B	Sakalti/Saka-1.5B	1.5B	51.4	40.0	31.3	31.5	47.5	33.5	39.2
≤2B	Qwen/Qwen3-1.7B-Base	1.7B	56.8	49.7	38.2	40.0	75.6	43.9	50.7
≤2B	Qwen/Qwen1.5-1.8B	1.8B	32.7	26.7	23.8	26.0	31.5	23.6	27.4
≤2B	silma-ai/SILMA-Kashif-2B-Instruct-v1.0	2B	59.7	45.6	33.1	38.8	73.3	35.8	47.7
≤2B	google/gemma-2-2b-it	2B	34.1	30.1	23.6	20.1	31.2	23.4	27.1
≤2B	LiquidAI/LFM2-350M	350M	39.0	35.2	30.9	28.3	43.3	29.1	34.3
≤2B	HALA-350M	350M	51.4	41.2	36.9	34.5	52.1	35.4	41.9 (+7.6)
≤2B	LiquidAI/LFM2-700M	700M	50.1	38.3	34.3	32.5	56.3	37.2	41.4
≤2B	HALA-700M	700M	55.5	45.9	40.6	34.7	65.2	39.4	46.9 (+5.5)
≤2B	LiquidAI/LFM2-1.2B	1.2B	53.8	45.2	35.0	34.7	65.6	43.4	46.3
≤2B	HALA-1.2B	1.2B	59.2	48.6	43.4	41.6	71.7	44.2	**51.4** (+5.1)
Small (7B–9B)									
7B–9B	CohereForAI/c4ai-command-r7b-arabic-02-2025	7B	74.8	59.3	65.0	63.8	80.5	50.1	65.6
7B–9B	JasperV13/Yehia-7B-DPO-Reasoning-preview	7B	75.1	66.3	51.8	54.9	81.9	55.1	64.2
7B–9B	Navid-AI/Yehia-7B-preview	7B	70.8	64.9	52.1	54.4	87.5	53.4	63.9
7B–9B	JasperV13/Yehia-7B-Reasoning-preview	7B	75.2	66.3	52.7	55.0	80.8	55.2	64.2
7B–9B	ALLaM-AI/ALLaM-7B-Instruct-preview	7B	69.5	64.9	51.6	54.2	86.9	52.8	63.3
7B–9B	Qwen/Qwen2-7B-Instruct	7B	73.2	60.0	47.3	59.5	82.8	51.3	62.4
7B–9B	Qwen/Qwen3-8B-Base	8B	74.8	65.0	52.5	52.2	83.4	61.5	64.9
7B–9B	QCRI/Fanar-1-9B-Instruct	9B	76.4	65.8	52.7	73.3	88.3	58.6	69.2
7B–9B	HALA-9B	9B	78.3	65.6	53.8	70.4	89.6	61.4	**69.9** (+0.7)

Model families. To contextualize HALA within the broader landscape, we include models spanning both multilingual and Arabic-centric families: LLaMA (Grattafiori et al., 2024), Qwen (Yang et al., 2025), Gemma (Gemma Team et al., 2025), SILMA (silma-ai, 2024; SILMA-AI, 2025), Saka, FANAR (Team et al., 2025a), Yehia (Navid-AI, 2025), ALLAM, Command-R, and LiquidAI. We report our HALA models at 350M, 700M, 1.2B, and 9B parameters alongside their corresponding bases (LiquidAI checkpoints and FANAR), and representative competitive baselines (e.g., Command-R-7B Arabic).

Main results. The aggregated results across the six benchmarks are summarized in Table 1. In the *nano* regime (≤2B), HALA-1.2B improves substantially over its base (LiquidAI/LFM2-1.2B), achieving the best average within the size bucket (cf. Table 1). Similarly, HALA-350M and HALA-700M consistently outperform their Liquid bases across most tasks, indicating that our translate–and–tune pipeline yields *consistent Arabic gains* even at very small scales. In the *small* regime (≤9B), HALA-9B consistently outperforms the previous state-of-the-art `QCRI/Fanar-1-9B-Instruct` baseline on the average metric, while maintaining competitive scores on individual tasks. These trends support our central claim: *language-centric* tuning on high-fidelity Arabic instruction data improves Arabic capability across scales, and merging (Sanyal et al., 2023) (`slerp`, t=0.5) preserves general competence while enhancing Arabic instruction-following.

Translator quality: EN→AR MMLU question translation. We assess translation fidelity in an instruction-style regime by constructing a controlled, reference-based evaluation using `cais/mmlu` (English questions) and `openai/mmmlu` (Arabic questions). We uniformly sample n=500 English questions from `cais/mmlu` with a fixed random seed, translate each to Arabic using the system under test, and align it to its ground-truth Arabic counterpart from the `openai/mmmlu` Arabic subset (same subject and item ID). We report *BLEU* (SacreBLEU, 13a tokenization), *ROUGE-L* (F1, `rouge-score`), and

Table 2: **EN→AR translation quality on 500 sampled MMLU questions.** References come from the Arabic subset of `openai/mmmlu`. Higher is better. Values in (·) denote absolute deltas vs. the reference system within each block (FP8 vs. FP16 for the teacher translator; HALA vs. `LFM2-1.2B` base for the lightweight translator). Prompts are fixed as specified above.

System	BLEU ↑	ROUGE-L ↑	chrF++ ↑
Teacher translator			
CohereLabs/command-a-translate-08-2025 (FP16)	53.1	26.0	68.6
CohereLabs/command-a-translate-08-2025 (FP8)	53.5 (+0.3)	26.0 (+0.0)	68.9 (+0.3)
Lightweight translator (LFM2-1.2B family)			
LiquidAI/LFM2-1.2B (base)	16.0	19.3	43.2
HALA-LFM2-1.2B Translator (ours)	48.2 (+32.1)	25.1 (+5.9)	64.2 (+21.0)

chrF++ (SacreBLEU) between the system output and the reference Arabic question. Exact sampling seeds, preprocessing, and metric commands will be released in the accompanying repository.

Prompting setup (fairness control). To ensure comparability across systems, we use fixed prompts:

- **LiquidAI/LFM2-1.2B (specialized translator) prompt:**

```
You are a professional translation
engine. Translate English to
Modern Standard Arabic. Reply ONLY
with the Arabic translation—no quotes,
notes, or explanations.
Translate everything that follows
into Arabic: {text}
```

- **All other models (teacher FP16/FP8 and baselines) prompt:**

```
Translate everything that follows
into Arabic: {text}
```

Here, `{text}` is replaced verbatim by the English question from `cais/mmlu`. Outputs are evaluated directly against the paired Arabic reference from `openai/mmmlu` without post-processing beyond each metric's built-in normalization.

Compute and cost. All models were trained within a budget of $1,000 on 8×H100-SXM GPUs, and dataset translation was performed on 12×A100 GPUs at an additional cost of roughly $500.

4 Limitations

In this work, we focus exclusively on nanoscale and small-scale models. Extending our investigation to larger models could provide further insight into whether similar improvements can be obtained at scale. However, this exploration is omitted because of the substantial computational cost associated with training and evaluating larger models.

5 Conclusion

We presented HALA, a family of *language-centric* Arabic models that leverage an efficient translate–and–tune pipeline: compress a capable AR↔EN translator to FP8, bootstrap million-scale Arabic instruction data from high-quality English sources, and fine-tune compact and small models with `slerp`-based merging. HALA delivers consistent improvements over base Liquid and FANAR checkpoints, achieving state-of-the-art averages in both the *nano* (≤2B) and *small* (≤9B) categories on a diverse Arabic benchmark suite. We release models, data, and recipes to catalyze further research on Arabic instruction tuning and to encourage *language-centric* approaches as a complement to breadth-first multilingual scaling.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Red Hat AI and vLLM Project. 2024. LLM Compressor.

Shahad Al-Khalifa, Nadir Durrani, Hend Al-Khalifa, and Firoj Alam. 2025. The landscape of Arabic large language models (ALLMs): A new era for Arabic language technology. *Preprint*, arXiv:2506.01340.

Emad A Alghamdi, Reem I Masoud, Deema Alnuhait, Afnan Y Alomairi, Ahmed Ashraf, and Mohamed Zaytoon. 2024. Aratrust: An evaluation of

trustworthiness for llms in arabic. *arXiv preprint arXiv:2403.09017*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023a. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

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. 2023b. AlGhafa evaluation benchmark for Arabic language models. In *Proceedings of ArabicNLP 2023*, pages 244–275, Singapore (Hybrid). Association for Computational Linguistics.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12753–12776, Bangkok, Thailand. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

M. Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, Ahmed Abdelali, and 5 others. 2024. ALLaM: Large language models for Arabic and english. *Preprint*, arXiv:2407.15390.

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, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

Run-Ze Fan, Xuefeng Li, Haoyang Zou, Junlong Li, Shwai He, Ethan Chern, Jiewen Hu, and Pengfei Liu. 2024. Reformatted alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 574–597, Miami, Florida, USA. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Jean-bastien Grill, Geoffrey Cideron, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 14 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786. Author list abbreviated.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's MergeKit: A toolkit for merging large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485, Miami, Florida, US. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, ..., Xiaolan Ma, Xilun Wang, and Yossi Adi. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783. Author list abbreviated.

Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. 2023. Lighteval: A lightweight framework for llm evaluation.

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.

Alex Havrilla. 2023. synthetic-instruct-gptj-pairwise (revision cc92d8d).

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, 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. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.

Maxim Ifergan, Leshem Choshen, Roee Aharoni, Idan Szpektor, and Omri Abend. 2024. Beneath the surface of consistency: Exploring cross-lingual knowledge representation sharing in LLMs. *Preprint*, arXiv:2408.10646.

Hannah Calzi Kleidermacher and James Zou. 2025. Science across languages: Assessing LLM multilingual translation of scientific papers. *arXiv preprint arXiv:2502.17882*.

"Fajri Koto, Haonan Li, Sara Shatanawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin". 2024. Arabicmmlu: Assessing massive multitask language understanding in arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Andrey Kuzmin, Mart Van Baalen, Yuwei Ren, Markus Nagel, Jorn Peters, and Tijmen Blankevoort. 2022. Fp8 quantization: The power of the exponent. *Advances in Neural Information Processing Systems*, 35:14651–14662.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2024. Tülu 3: Pushing frontiers in open language model post-training. *https://arxiv.org/abs/2411.15124*.

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, and 2 others. 2022. Few-shot learning with multilingual language models. *Preprint*, arXiv:2112.10668.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Dakuan Lu, Xiaoyu Tan, Rui Xu, Tianchu Yao, Chao Qu, Wei Chu, Yinghui Xu, and Yuan Qi. 2025. Scp-116k: A high-quality problem-solution dataset and a generalized pipeline for automated extraction in the higher education science domain. *arXiv preprint arXiv:2501.15587*.

Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. 2024. A paradigm shift: The future of machine translation lies with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352, Torino, Italia. ELRA and ICCL.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of GPT–4. *Preprint*, arXiv:2306.02707.

Navid-AI. 2025. Yehia 7b preview. `https://huggingface.co/Navid-AI/Yehia-7B-preview`.

Sunny Sanyal, Atula Neerkaje, Jean Kaddour, Abhishek Kumar, and Sujay Sanghavi. 2023. Early weight averaging meets high learning rates for llm pre-training. *Preprint*, arXiv:2306.03241.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos

Mahmoud Bsharat, and 9 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

silma-ai. 2024. Silma 9b instruct v1.0. https://huggingface.co/silma-ai/SILMA-9B-Instruct-v1.0.

SILMA-AI. 2025. Silma kashif 2b instruct v1.0. https://huggingface.co/silma-ai/SILMA-Kashif-2B-Instruct-v1.0.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, and 22 others. 2025a. Fanar: An Arabic-centric multimodal generative AI platform. *Preprint*, arXiv:2501.13944.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025b. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Ryan Teknium, Jeffrey Quesnelle, and Guang Chen. 2024. Hermes 3 technical report. *Preprint*, arXiv:2408.11857.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

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, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2024. LaMini-LM: A diverse herd of distilled models from large-scale instructions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 944–964, St. Julian's, Malta. Association for Computational Linguistics.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

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.

Ai Ming Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Hai Zhao, Hang Xu, Hao-Lun Sun, and 36 others. 2023. Baichuan 2: Open large-scale language models. *ArXiv*, abs/2309.10305.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.