Command R7B Arabic: A Small, Enterprise Focused, Multilingual, and Culturally Aware Arabic LLM

Yazeed Alnumay*, **Alexandre Barbet***, **Anna Bialas***, **William Darling***, **Shaan Desai***, **Joan Devassy***, **Kyle Duffy***, **Stephanie Howe***, **Olivia Lasche***, **Justin Lee***, **Anirudh Shrinivason***, **Jennifer Tracey***

Cohere

Abstract

Building high-quality large language models (LLMs) for enterprise Arabic applications remains challenging due to the limited availability of digitized Arabic data. In this work, we present a data synthesis and refinement strategy to help address this problem, namely, by leveraging synthetic data generation and human-in-the-loop annotation to expand our Arabic training corpus. We further present our iterative post training recipe that is essential to achieving state-of-the-art performance in aligning the model with human preferences, a critical aspect to enterprise use cases. The culmination of this effort is the release of a small, 7B, open-weight model that outperforms similarly sized peers in head-to-head comparisons and on Arabic-focused benchmarks covering cultural knowledge, instruction following, RAG, and contextual faithfulness.

1 Introduction

Multilingual language models are evolving rapidly (Huang et al., 2024b), yet specific languages and capabilities remain underdeveloped, particularly in enterprise applications. While state-of-the-art models continue to improve, they often struggle to adapt to linguistic and professional needs in languages like Arabic (Gabriel Nicholas, 2023), the most spoken language in Africa (Zucchet, 2024). This challenge becomes even more pronounced when additional constraints are introduced: the need to keep the model small to ensure accessibility even with limited resources, overcoming data scarcity, and accounting for linguistic nuances that do not translate well from English, all the while prioritizing rapid iteration to stay aligned with the fast-moving market. To address these issues, we developed a post-training approach that efficiently tailors cutting-edge models to specialized capabilities. This report outlines our methodology and findings, offering insights into adapting LLMs for language-specific and professional domains.

*Equal contribution. Authors appear in alphabetical order by second name.

2 Related Work

With the recent rapid development in LLMs (Zhao et al., 2024), some focus was placed on improving model multilingualism through second language acquisition techniques (Huang et al., 2024b). These techniques aim to circumvent data scarcity in languages other than English by adding other language capabilities to English models, which is more data efficient. For instance, the Llama 3 family of models adds a final pretraining stage by adding multilingual pretraining data mixed with English (Grattafiori et al., 2024). These techniques have been applied to Arabic-centric models, such as ALLaM (Bari et al., 2025), Jais (Sengupta et al., 2023; Inception, 2024), AceGPT (Huang et al., 2024a; Zhu et al., 2024; Liang et al., 2024), and Fanar (Fanar Team et al., 2025). These projects primarily focused on pretraining data mixture, staging, and tokenizer innovations, including vocabulary expansion (ALLaM), iterative vocabulary expansion (AceGPT), and morphology-based tokenization (Fanar). While they contribute strong foundational models for the community, they do not offer computationally efficient post-training methods.

Post-training has become essential for building robust models (Wei et al., 2022; Kumar et al., 2025; Ouyang et al., 2022). Many research labs have contributed to the open-source community by documenting modern post-training techniques. Notable examples include Tülu 3 (Lambert et al., 2025), which provides a comprehensive overview of general post-training methods, and Aya Expanse (Dang et al., 2024), which focuses on multilingual adaptation.

Our work builds on these efforts by developing a systematic, iterative, and comprehensive approach to efficiently adapt LLMs for languages. Specifi-

Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), pages 126–135
July 31, 2025 ©2025 Association for Computational Linguistics

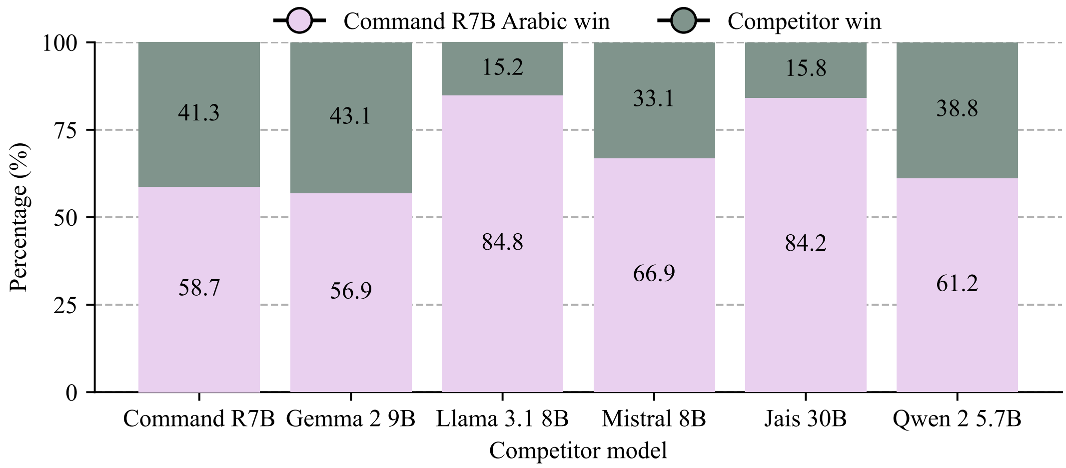

Figure 1: Evaluations on enterprise usability factors (mArenaHard, described in Section 4). Auto win-rates on Arabic version of LMSYS Arena "Hard" human preference tasks (Dang et al., 2024). Command R7B Arabic outperforms all listed similarly-sized models.

cally, we leverage iterative tuning (Grattafiori et al., 2024) methods that rely on best-of-N sampling to generate instruction and preference data via automated reward models or human preference (Yuan et al., 2024). We also further reduce compute requirements by incorporating model merging techniques (Goddard et al., 2024; Yang et al., 2024).

3 Methods

Our training procedure is illustrated in Figure 2. We start by selecting a strong starting model (Section 3.1), on which we perform three distinct training phases: *(i)* supervised fine-tuning (SFT) (Wei et al., 2022), for which we employ iterative dataset refinement techniques (Sections 3.2 and 3.3), *(ii)* off-policy (offline) preference tuning, and *(iii)* iterative preference tuning. The latter two are described in Section 3.4. After each training phase, we merge expert models into a single general model (Section 3.5).

3.1 Base Model Selection

As a starting checkpoint, we chose Command R7B (Cohere, 2024; Cohere et al., 2025) - a strong, general purpose, and open-weight model already trained on a large corpus of multilingual data, including Arabic, and specialized to enterprise use-cases. Additionally, Robinson et al. (2025) showed that Cohere models excel in dialectal Arabic compared to other open-weight models. Our primary objective was to reach state-of-the-art performance in Arabic enterprise use cases while preserving the model's performance on other core capabilities. Starting from an already polished checkpoint meant we could spend more effort on our data and training efforts that refined Arabic-specific tasks.

3.2 Multilingual Arbitrage for Capability Enhancement

Previous work by Aya (Odumakinde et al., 2024) has demonstrated that synthetic data generation is crucial for achieving state-of-the-art performance, and this is especially true for domains with limited data availability such as Arabic. However, a key challenge when training Arabic LLMs is the distinctive difference between Arabic and English. Not only do these languages differ in syntax and morphology, but there are also variations in cultural and contextual nuances that make literal translation challenging. For example, lexical control tasks such as length adherence and structured generation are awkward or nonsensical when translated to Arabic.

To address this, we implemented a human-in-the-loop approach:

- We collaborated with expert annotators to translate IFEval (Zhou et al., 2023) instructions into Arabic. Additionally, we augmented the set with two instructions specific to the Arabic language: “add N diacritics to the response” and “use a specific grammatical verb to start sentences”. This ensured better alignment with Arabic linguistic and cultural nuances.

- These instructions were used as seeds to

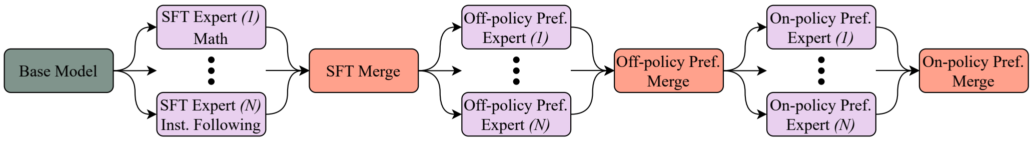

Figure 2: Outline of Command R7B Arabic's training processes with three training stages, each training multiple experts that are merged into a single general model. For instance, in the SFT stage, multiple SFT expert models are trained to excel in specific domains, such as mathematics or instruction following. These experts are subsequently merged to create a generalist SFT model via parameter-wise linear interpolation of the experts' weights.

synthetically generate instruction following prompts in Arabic and subsequently the corresponding completions.

- In accordance with the work done in Aya's Multilingual Arbitrage (Odumakinde et al., 2024), we scored and filtered completions using a reward model, a panel of LLM judges for Arabic natural language quality, and max reward difference for preference pair dataset creation.

This targeted approach ensured that the model learned to follow instructions naturally in Arabic, which is apparent in arena style win-rates where our model is consistently favored over other competitor models, as shown in Figure 1.

3.3 Dataset Curation and Iterative Supervised Refinement

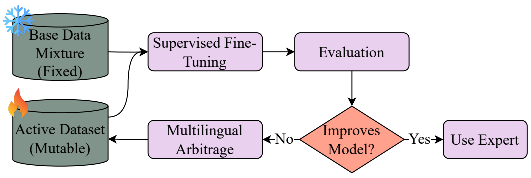

Figure 3: Flowchart for our iterative supervised refinement approach. It ensures that all datasets used improve targeted model performance by mixing a base data mixture with a targeted dataset that is iteratively improved via multilingual arbitrage.

The availability of high-quality Arabic datasets is a well-documented challenge (Gabriel Nicholas, 2023). We aimed to incorporate both publicly available datasets, including ArMATH (Alghamdi et al., 2022), ArabicaQA (Abdallah et al., 2024), and synthetically generated datasets, while enforcing a high-quality data standard. With this in mind, we defined the Iterative Supervised Refinement during Supervised Fine-Tuning (SFT) training phase as a process to optimize our dataset composition. The steps are illustrated in Figure 3 and are as follows:

1. Define a base data mix consisting of high-quality instruction-tuning data.
2. For each new dataset in consideration, add it to the base data mixture and fine-tune the model.
3. Evaluate the resulting model using a benchmark evaluation harness to measure the impact of the new dataset.
4. If the dataset improves performance in any critical capability, retain it for the next iteration.
5. If no improvement was observed, apply Multilingual Arbitrage, refining the prompts before re-running the process.

This approach enabled us to design an optimal dataset mixture that maximized the model's instruction-following capabilities while maintaining a high standard for data quality.

3.4 Preference Tuning for Final Model Optimization

Since we initialized from a strong Command R7B model, it was essential to ensure that enhancements in Arabic did not degrade performance on other benchmarks. Similar to the methodology described by Aya (Üstün et al., 2024), we used two stages of preference tuning as final polishing steps to improve model performance and align it with human preferences. In the first phase, we performed offline preference training on general preference datasets to refine the model's conversational fluency. In the second phase, we ran iterative preference training, incorporating an Arabic-translated reasoning and math-focused dataset (Alghamdi et al., 2022), which proved particularly beneficial for maintaining high performance across diverse enterprise use cases. Both preference tuning stages utilize the direct preference optimization (DPO) (Rafailov et al., 2024) algorithm.

Benchmark	R7B Arabic	R7B (Cohere et al., 2025)	Gemma 9B (Gemma Team et al., 2024)	Llama 3.1 8B (Grattafiori et al., 2024)	Qwen 2.5 7B (Yang et al., 2025)	Ministral 8B (Mistral, 2024)
AlGhafa-Native	**82.2**	81.5	81.3	80.1	80.2	76.6
ArabicMMLU	60.9	59.7	**62.4**	56.6	61.2	53.6
IFEval AR	**69.0**	57.8	67.8	48.4	62.4	49.3
TyDIQA-GoldP Arabic	**83.0**	79.9	76.4	65.9	60.9	57.7
FaithEval Arabic	**51.6**	49.9	47.0	40.9	49.9	25.5
Average	**69.3**	65.8	67.0	58.4	62.9	52.5

Table 1: Full performance comparison against competitor models on Arabic-specific benchmarks. The highest score in each row is in **bold**. Command R7B Arabic is best-in-class compared to similarly sized models on all Arabic benchmarks, with the exception of ArabicMMLU.

3.5 Expert Model Merging

After completing the iterative supervised refinement procedure described in Section 3.3 to create multiple expert models from various datasets, one path forward is to retrain a new generalist model by combining appropriate datasets based on the insights obtained from these experiments. However, we can eliminate computational redundancy by merging various expert models. This is a common practice with mature frameworks (Goddard et al., 2024). The literature lacks conclusive theoretical foundations for the effectiveness of model merging, but extensive experimentation has shown it is a successful strategy in practice (Yang et al., 2024).

To reduce the expert merge search space, we only considered linear merges (Utans, 1996) of the expert models. We tested several weighting schemes based on the importance of each capability and the size of each expert's training data. In the end, our best model was obtained by assigning equal weight to each expert.

In practice, model merging reduces computational cost. However, it complicates replication and adds an additional source of potential errors.

4 Experiments and Results

4.1 Arabic Language

To measure the performance of various models in general Arabic language generation and understanding, as well as enterprise use-cases, such as grounding model generation with enterprise-specific data via RAG and precise instruction following, we utilized the following evaluation suite:

- **IFEval AR**: An internal Arabic translation of the original English dataset (Zhou et al., 2023) with 541 test samples. It measures a model's precise instruction following ability, with instructions such as "use at least 300 words" or "do not use commas."

- **AlGhafa-Native**: The subset[1] of AlGhafa (Almazrouei et al., 2023) tasks which were curated by native Arabic speakers, which encapsulates the following:
  - MCQ Exams AR (562 samples) (Hardalov et al., 2020).
  - Belebele AR Dialects (5,400 samples) and Belebele AR MSA (900 samples) (Bandarkar et al., 2024).
  - AraFacts balanced (80 samples) (Sheikh Ali et al., 2021).
  - SOQAL (155 samples) (Mozannar et al., 2019).
  - XGLUE (155 samples) (Liang et al., 2020).
  - Rating sentiment no neutral (8,000 samples) and rating sentiment (6,000 samples) from the HARD-Arabic-Dataset (Elnagar et al., 2018).
  - Sentiment (1,725 samples) (Abu Farha et al., 2021).

  We report the unweighted average percentage performance across all tasks.

- **TyDiQA-GoldP Arabic**: The 921 samples in Arabic from the original TyDiQA (Clark et al., 2020) golden passage (GoldP) secondary task, in which models are provided with a question and a single passage that contains the question's answer. Models are prompted to determine the substring in the passage that answers the question.

[1] https://huggingface.co/datasets/OALL/AlGhafa-Arabic-LLM-Benchmark-Native

Benchmark	R7B Arabic	R7B (Cohere et al., 2025)	Gemma 9B (Gemma Team et al., 2024)	Llama 3.1 8B (Grattafiori et al., 2024)	Qwen 2.5 7B (Yang et al., 2025)	Ministral 8B (Mistral, 2024)
BBH (Suzgun et al., 2022)	36.2	36.0	**42.1**	29.9	34.9	25.8
MuSR (Sprague et al., 2024)	**11.9**	10.2	9.7	8.4	8.5	8.4
GPQA (Rein et al., 2023)	7.9	7.8	**14.8**	2.4	5.5	4.5
MMLU Pro (Wang et al., 2024)	29.4	28.6	32.0	30.7	**36.5**	30.7
IfEval (Zhou et al., 2023)	**83.3**	77.1	74.4	78.6	75.9	59.0
MATH* (Hendrycks et al., 2021b)	19.6	29.9	19.1	19.3	50.0	19.6
Average	31.4	31.6	32.1	28.2	**35.2**	22.0

* The MATH benchmark used in this leaderboard changed in early January due to a DMCA takedown notice for the original benchmark.

Table 2: Performance comparison of R7B Arabic against similarly sized models on multiple benchmarks. The highest score in each row is in **bold**. Command R7B Arabic retains most of the general and English capabilities of its base model, Command R7B, as indicated by the similar average scores.

- **ArabicMMLU** (Koto et al., 2024): Inspired by the original MMLU (Hendrycks et al., 2021a) in English, ArabicMMLU is a collection of 14,575 native Arabic multiple choice questions focusing on knowledge and reasoning. It covers 40 tasks at various education levels (elementary to college) and regions (North Africa, Levant, and Gulf).

- **FaithEval Arabic**: An internal Arabic translation of a 500 sample subset of the original English dataset (Ming et al., 2024). It measures the model's RAG performance when provided with unanswerable, inconsistent, or counterfactual contexts.

- **Multilingual ArenaHard** (Dang et al., 2024): A machine translation of 500 questions from the original English LMArena (formerly LMSYS) Arena-Hard-Auto (Li et al., 2024) prompts into various other languages. We limit our evaluation to the Arabic subset. The evaluation uses GPT-4o as a judge to compare completions from two different models.

Table 1 shows results compared to other models in the same size category. The Command R7B Arabic model outperforms all baselines across key Arabic benchmarks, achieving an average score of 69.3, surpassing Command R7B (65.8) and Gemma 9B (67.0). It performs at the top of its size class in the following benchmarks: Cultural Knowledge (AlGhafa-Native), Instruction Following (IFEval AR) validating our human-in-the-loop data strategy, RAG Question Answering (TyDiQA-GoldP Arabic), and RAG Faithfulness (FaithEval Arabic). In General Knowledge (ArabicMMLU), Command R7B Arabic scores third, while staying competitive with Gemma 9B and Qwen 2.7.

4.2 General Capabilities

Retaining general capabilities is essential for the model to be helpful in enterprise settings. We thoroughly measured our model's performance and present the results of the standardized Hugging Face Open LLM Leaderboard benchmarks (Fourrier et al., 2024; Gao et al., 2021). Table 2 shows that our model excels in IfEval and MuSR, achieving the highest scores among similarly sized models. Notably, it outperforms the initial checkpoint on all benchmarks except for MATH, possibly due to the change in methodology.

These benchmark results (Table 1 and Table 2), coupled with auto win-rate data (Figure 1), validate that our approach effectively enhances Arabic language capabilities while maintaining robust performance in enterprise applications.

5 Conclusion

In this work, we rapidly iterated to develop Command R7B Arabic, a small, yet competent Arabic LLM optimized for enterprise applications. By leveraging synthetic data generation, multilingual arbitrage, and human-in-the-loop interventions, we significantly improved instruction following, retrieval-augmented generation (RAG), and question answering capabilities in Arabic. However, transferring knowledge from English-centric datasets to Arabic remains an open challenge. Future work should explore more effective adaptation strategies, ensuring higher linguistic and factual alignment across languages.

Limitations

Our work focuses on Modern Standard Arabic (MSA), which is widely used in formal and professional settings but differs significantly from spoken dialects across the Arabic-speaking world. While

MSA provides a strong foundation for enterprise applications, real world use cases often involve dialectal Arabic, which varies by region and context. Future work should explore dialect adaptation strategies to improve robustness across diverse Arabic varieties.

We adapted Faithfulness (FaithEval Arabic), Question Answering (TyDi QA Arabic), and Instruction Following (IFEval AR) to measure enterprise-relevant capabilities. Still, these benchmarks remain proxies rather than direct tests of real-world deployment challenges. The effectiveness of our model in enterprise workflows can only be fully validated through real-world deployment and user feedback.

Acknowledgments

This work was a collaboration between many teams in Cohere. We would like to particularly acknowledge the following people who supported the project through advice and maintenance of our core infrastructure:

Modeling Team: Théo Dehaze, Jesse Willman, Lewis Stott, Florian Strub, Jay Alammar, Matthias Gallé, Samuel Cahyawijaya, Alexandre Bérard, Wei-Yin Ko, Kocmi Tom, Dennis Aumiller, Nathan Grinsztajn, Phil Blunsom, Jon Ander Campos, Yi Chern Tan, Sander Land, Nithya Govindarajan, Nick Jakobi, Adrien Morisot, Olivia Markham;

C4AI: Sungjin Hong, Alejandro Salamanca, Marzieh Fadaee, Ahmet Üstün, Sara Hooker;

Infrastructure: Cécile Robert-Michon, Jessica Xie, Adi Bongale, Ace Eldeib, Sudip Roy, Manoj Govindassamy, Maxime Brunet, Jeremy Pekmez, Terrence Zhao, Renjie Huang;

Applied ML Team: Neeral Beladia, Gokce Keskin, Utsav Garg, Jason Jung, Hemangani Nagarajan, Sanal Shivaprasad, Sam Passaglia, Edmond Wen, Trushant Kalyanpur, Vivek Muppalla, Evren Tumer, Harri Bell-Thomas;

Annotators: Arwa Alaya, Noha Shehata , Eyas Shanaah , Abdullah Omran, Nermeen Isaac, Izzat Homsi, Mahmoud Mansour, Mayar Soliman, Israr Wahid, Vanessa Choueiry, Mona Knobloch, Fatima Zahra Zyad;

Annotator Operations: Claire Cheng, Trisha Starostina, Brenda Malacara Lopez;

Leadership: Aidan Gomez, Martin Kon, Saurabh Baji, Phil Blunsom;

External partners: Neha Sengupta, Ali El Filali.

References

Abdelrahman Abdallah, Mahmoud Kasem, Mahmoud Abdalla, Mohamed Mahmoud, Mohamed Elkasaby, Yasser Elbendary, and Adam Jatowt. 2024. Arabicaqa: A comprehensive dataset for arabic question answering. *Preprint*, arXiv:2403.17848.

Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy. 2021. Overview of the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 296–305, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Reem Alghamdi, Zhenwen Liang, and Xiangliang Zhang. 2022. ArMATH: a dataset for solving Arabic math word problems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 351–362, Marseille, France. European Language Resources Association.

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

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 749–775. Association for Computational Linguistics.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, Ahmed Abdelali, Jeril Kuriakose, Abdalghani Abujabal, Nora Al-Twairesh, Areeb Alowisheq, and Haidar Khan. 2025. ALLam: Large language models for arabic and english. In *The Thirteenth International Conference on Learning Representations*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Preprint*, arXiv:2003.05002.

Cohere. 2024. Introducing command r7b: Fast and efficient generative ai.

Team Cohere, :, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, Alexandre Bérard, Andrew Berneshawi, Anna Bialas, Phil Blunsom, Matt Bobkin, Adi Bongale, Sam Braun, Maxime Brunet, Samuel Cahyawijaya, David Cairuz, Jon Ander Campos, Cassie Cao, Kris Cao, Roman Castagné, Julián Cendrero, Leila Chan Currie, Yash Chandak, Diane Chang, Giannis Chatziveroglou, Hongyu Chen, Claire Cheng, Alexis Chevalier, Justin T. Chiu, Eugene Cho, Eugene Choi, Eujeong Choi, Tim Chung, Volkan Cirik, Ana Cismaru, Pierre Clavier, Henry Conklin, Lucas Crawhall-Stein, Devon Crouse, Andres Felipe Cruz-Salinas, Ben Cyrus, Daniel D'souza, Hugo Dalla-Torre, John Dang, William Darling, Omar Darwiche Domingues, Saurabh Dash, Antoine Debugne, Théo Dehaze, Shaan Desai, Joan Devassy, Rishit Dholakia, Kyle Duffy, Ali Edalati, Ace Eldeib, Abdullah Elkady, Sarah Elsharkawy, Irem Ergün, Beyza Ermis, Marzieh Fadaee, Boyu Fan, Lucas Fayoux, Yannis Flet-Berliac, Nick Frosst, Matthias Gallé, Wojciech Galuba, Utsav Garg, Matthieu Geist, Mohammad Gheshlaghi Azar, Ellen Gilsenan-McMahon, Seraphina Goldfarb-Tarrant, Tomas Goldsack, Aidan Gomez, Victor Machado Gonzaga, Nithya Govindarajan, Manoj Govindassamy, Nathan Grinsztajn, Nikolas Gritsch, Patrick Gu, Shangmin Guo, Kilian Haefeli, Rod Hajjar, Tim Hawes, Jingyi He, Sebastian Hofstätter, Sungjin Hong, Sara Hooker, Tom Hosking, Stephanie Howe, Eric Hu, Renjie Huang, Hemant Jain, Ritika Jain, Nick Jakobi, Madeline Jenkins, JJ Jordan, Dhruti Joshi, Jason Jung, Trushant Kalyanpur, Siddhartha Rao Kamalakara, Julia Kedrzycki, Gokce Keskin, Edward Kim, Joon Kim, Wei-Yin Ko, Tom Kocmi, Michael Kozakov, Wojciech Kryściński, Arnav Kumar Jain, Komal Kumar Teru, Sander Land, Michael Lasby, Olivia Lasche, Justin Lee, Patrick Lewis, Jeffrey Li, Jonathan Li, Hangyu Lin, Acyr Locatelli, Kevin Luong, Raymond Ma, Lukáš Mach, Marina Machado, Joanne Magbitang, Brenda Malacara Lopez, Aryan Mann, Kelly Marchisio, Olivia Markham, Alexandre Matton, Alex McKinney, Dominic McLoughlin, Jozef Mokry, Adrien Morisot, Autumn Moulder, Harry Moynehan, Maximilian Mozes, Vivek Muppalla, Lidiya Murakhovska, Hemangani Nagarajan, Alekhya Nandula, Hisham Nasir, Shauna Nehra, Josh Netto-Rosen, Daniel Ohashi, James Owers-Bardsley, Jason Ozuzu, Dennis Padilla, Gloria Park, Sam Passaglia, Jeremy Pekmez, Laura Penstone, Aleksandra Piktus, Case Ploeg, Andrew Poulton, Youran Qi, Shubha Raghvendra, Miguel Ramos, Ekagra Ranjan, Pierre Richemond, Cécile Robert-Michon, Aurélien Rodriguez, Sudip Roy, Sebastian Ruder, Laura Ruis, Louise Rust, Anubhav Sachan, Alejandro Salamanca, Kailash Karthik Saravanakumar, Isha Satyakam, Alice Schoenauer Sebag, Priyanka Sen, Sholeh Sepehri, Preethi Seshadri, Ye Shen, Tom Sherborne, Sylvie Shang Shi, Sanal Shivaprasad, Vladyslav Shmyhlo, Anirudh Shrinivason, Inna Shteinbuk, Amir Shukayev, Mathieu Simard, Ella Snyder, Ava Spataru, Victoria Spooner, Trisha Starostina, Florian Strub, Yixuan Su, Jimin Sun, Dwarak Talupuru, Eugene Tarassov, Elena Tommasone, Jennifer Tracey, Billy Trend, Evren Tumer, Ahmet Üstün, Bharat Venkitesh, David Venuto, Pat Verga, Maxime Voisin, Alex Wang, Donglu Wang, Shijian Wang, Edmond Wen, Naomi White, Jesse Willman, Marysia Winkels, Chen Xia, Jessica Xie, Minjie Xu, Bowen Yang, Tan Yi-Chern, Ivan Zhang, Zhenyu Zhao, and Zhoujie Zhao. 2025. Command a: An enterprise-ready large language model. *Preprint*, arXiv:2504.00698.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Ashraf Elnagar, Yasmin S Khalifa, and Anas Einea. 2018. Hotel arabic-reviews dataset construction for sentiment analysis applications. *Intelligent natural language processing: Trends and applications*, pages 35–52.

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, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, Mus'ab Husaini, Soon-Gyo Jung, Ji Kim Lucas, Walid Magdy, Safa Messaoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan Naeem, Mourad Ouzzani, Dorde Popovic, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, Yifan Zhang, Ahmed Ali, Yassine El Kheir, Xiaosong Ma, and Chaoyi Ruan. 2025. Fanar: An arabic-centric multimodal generative ai platform. *Preprint*, arXiv:2501.13944.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open llm leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`.

Aliya Bhatia Gabriel Nicholas. 2023. Lost in translation: Large language models in non-english content analysis. *Center for Democracy & Technology*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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. *arXiv preprint arXiv:2011.03080*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024a. Acegpt, localizing large language models in arabic. *Preprint*, arXiv:2309.12053.

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, et al. 2024b. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936*.

Inception. 2024. Jais family model card. *Hugging Face*.

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*.

Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip H. S. Torr, Salman Khan, and Fahad Shahbaz Khan. 2025. Llm post-training: A deep dive into reasoning large language models. *Preprint*, arXiv:2502.21321.

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, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *Preprint*, arXiv:2406.11939.

Juhao Liang, Zhenyang Cai, Jianqing Zhu, Huang Huang, Kewei Zong, Bang An, Mosen Alharthi, Juncai He, Lian Zhang, Haizhou Li, Benyou Wang, and Jinchao Xu. 2024. Alignment at pre-training! towards native alignment for arabic LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2024. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". *Preprint*, arXiv:2410.03727.

Mistral. 2024. Ministral 8b instruct model card. *Hugging Face*.

Hussein Mozannar, Karl El Hajal, Elie Maamary, and Hazem Hajj. 2019. Neural arabic question answering. *Preprint*, arXiv:1906.05394.

Ayomide Odumakinde, Daniel D'souza, Pat Verga, Beyza Ermis, and Sara Hooker. 2024. Multilingual arbitrage: Optimizing data pools to accelerate multilingual progress. *Preprint*, arXiv:2408.14960.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *Preprint*, arXiv:2311.12022.

Nathaniel R. Robinson, Shahd Abdelmoneim, Kelly Marchisio, and Sebastian Ruder. 2025. Al-qasida: Analyzing llm quality and accuracy systematically in dialectal arabic. *Preprint*, arXiv:2412.04193.

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, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Zien Sheikh Ali, Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2021. AraFacts: The first large Arabic dataset of naturally occurring claims. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 231–236, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *Preprint*, arXiv:2310.16049.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *Preprint*, arXiv:2210.09261.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Joachim Utans. 1996. Weight averaging for neural networks and local resampling schemes. In *Proc. AAAI-96 Workshop on Integrating Multiple Learned Models. AAAI Press*, pages 133–138. Citeseer.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *Preprint*, arXiv:2408.07666.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *Preprint*, arXiv:2401.10020.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A survey of large language models. *Preprint*, arXiv:2303.18223.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

Jianqing Zhu, Huang Huang, Zhihang Lin, Juhao Liang, Zhengyang Tang, Khalid Almubarak, Abdulmohsen Alharthik, Bang An, Juncai He, Xiangbo Wu, Fei Yu, Junying Chen, Zhuoheng Ma, Yuhao Du, He Zhang, Emad A. Alghamdi, Lian Zhang, Ruoyu Sun, Haizhou Li, Benyou Wang, and Jinchao Xu. 2024. Second language (arabic) acquisition of llms via progressive vocabulary expansion. *Preprint*, arXiv:2412.12310.

Elinor Zucchet. 2024. African languages: The top 70 most spoken languages in Africa — berlitz.com. `https://www.berlitz.com/blog/what-are-most-spoken-languages-africa`. [Accessed 27-04-2025].