NileChat: Towards Linguistically Diverse and Culturally Aware LLMs for Local Communities

Abdellah El Mekki[λ*] **Houdaifa Atou**[δ] **Omer Nacar**[ψ]
Shady Shehata[γ] **Muhammad Abdul-Mageed**[λ,γ]

[λ]The University of British Columbia [δ]Mohammed VI Polytechnic University
[ψ]Tuwaiq Academy [γ]Invertible AI
{abdellah.elmekki,muhammad.mageed}@ubc.ca

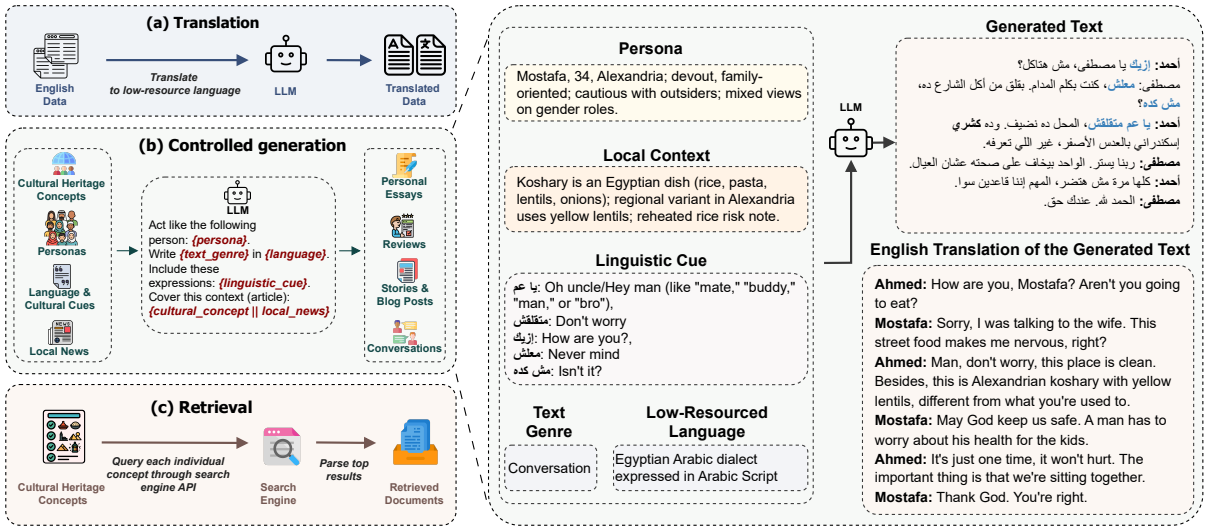

Figure 1: Our proposed framework enhances text data augmentation for low-resource local communities through a multi-stage pipeline. First, it **(a)** generates educational data using **machine translation**. Next, it **(b)** creates diverse, culturally-aware texts, such as stories and conversations, by simulating scenarios with local personas through **controlled synthetic data generation**. Finally, it **(c)** enriches the model with local knowledge by **retrieving and parsing culturally specific web content**. This entire process enables controlled text generation and retrieval-augmented pre-training, ensuring the cultural and value alignment of large language models for Arabic dialects.

Abstract

Enhancing the linguistic capabilities of Large Language Models (LLMs) to include low-resource languages is a critical research area. Current research directions predominantly rely on synthetic data generated by translating English corpora, which, while demonstrating promising linguistic understanding and translation abilities, often results in models aligned with source language culture. These models frequently fail to represent the cultural heritage and values of local communities. This work proposes a methodology to create both synthetic and retrieval-based pre-training data tailored to a specific community, considering its *(i) language*, *(ii) cultural heritage*, and *(iii) cultural values*. We demonstrate our methodology using Egyptian and Moroccan dialects as testbeds, chosen for their linguistic and cultural richness and current underrepresentation in LLMs. As a proof-of-concept, we develop *NileChat*, a 3B parameter Egyptian and Moroccan Arabic LLM adapted for Egyptian and Moroccan communities, incorporating their language, cultural heritage, and values. Our results on various understanding, translation, and cultural and values alignment benchmarks show that *NileChat* outperforms existing Arabic-aware LLMs of similar size and performs on par with larger models. This work addresses Arabic dialect in LLMs with a focus on cultural and values alignment via controlled synthetic data generation and retrieval-augmented pre-training for Moroccan Darija and Egyptian Arabic, including Arabizi variants, advancing Arabic NLP for low-resource

*Work initiated during a stay at MBZUAI.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 10979–11003
November 4-9, 2025

communities. We share our methods, data, and models with the community to promote the inclusion and coverage of more diverse communities in cultural LLM development.[1]

1 Introduction

Large Language Models (LLMs) have advanced rapidly, enabling remarkable proficiency across many tasks. Yet, this success is unevenly distributed across languages, with substantial performance disparities observed for non-English languages, particularly low-resource languages and dialectal variants (Navigli et al., 2023). A primary factor underlying this discrepancy is the limited representation of diverse multilingual data within the foundational pre-training corpora of these models which favors high-resource languages spoken in regions with high economic influence (Bender, 2011; Joshi et al., 2020).

Beyond linguistic limitations, a more profound challenge is the inherent risk of *cultural encapsulation* (Wrenn, 1962) in LLMs. Cultural encapsulation refers to an unconscious tendency to operate within one's own cultural lens, leading to misunderstanding or avoidance of differing perspectives and values. As LLMs are optimized to replicate patterns in their training data—predominantly sourced from specific cultural contexts (e.g., Western, English-speaking)—they risk internalizing and propagating these dominant perspectives as the norm (Dwivedi et al., 2023; Tao et al., 2024; Wang et al., 2024; Naous et al., 2024). The significance of cultural context cannot be overstated. As Edward Sapir noted:

> *"No two languages are ever sufficiently similar to be considered as representing the same social reality. The worlds in which different societies live are distinct worlds, not merely the same world with different labels attached." - Sapir (1929)*

This cultural bias is compounded by a fundamental mismatch: LLMs typically process data through a language-centric lens, whereas human communities are structured around shared social ties, perspectives, and values (MacQueen et al., 2001). Current LLMs adaptation techniques for new languages or communities (Gurgurov et al., 2024; Joshi et al., 2025) often fall short in bridging this cultural divide, especially for low-resource communities (Naous et al., 2024). For instance, machine translation, while useful for generating synthetic data to boost linguistic coverage (Joshi et al., 2025; Shang et al., 2025; Wang et al., 2025), primarily addresses the linguistic deficit. The translated content often retains the source language's cultural perspective, failing to incorporate authentic local nuances crucial for genuine interaction. Supervised fine-tuning (SFT) on target language data (Gala et al., 2024; Shang et al., 2025) can align models to specific tasks, but small datasets may not reshape deep-seated cultural biases from pre-training (Rystrøm et al., 2025) and can encourage hallucination with new factual data (Gekhman et al., 2024). While continued pre-training with culturally rich data could mitigate these issues, it faces a critical bottleneck for low-resource contexts: the scarcity of such high-quality digital texts.

This paper addresses the critical need to adapt multilingual LLMs to low-resource language communities by jointly considering their linguistic characteristics and cultural heritage & values. We propose a novel pipeline (illustrated in Figure 1) focused on data augmentation for continued pre-training. Our approach combines *controlled synthetic data generation* (Section 3.1.2) with *retrieval* (Section 3.1.3) methods. To address linguistic adaptation, we translate English pre-training data into the target local language focusing only on high-quality data from the educational domain (Section 3.1.1). Crucially, to imbue cultural relevance, we generate diverse texts reflecting specific cultural heritage concepts (e.g., food, celebrations, proverbs) using local persona descriptions (Section 3.1.2) reflecting the local cultural values. We demonstrate our method on the *Moroccan* and *Egyptian* Arabic dialects as low-resource testbeds. We further pre-train a multilingual LLM on a curated mix of real and synthetic data, evaluating its performance on tasks involving language understanding, translation, and alignment with cultural knowledge and values. Our findings show that the adapted model significantly outperforms baseline and existing models that are even bigger in size on most evaluation tasks.

The main contributions of this work are: **(i)** A novel framework for augmenting pre-training corpora tailored to local communities. This framework considers their unique linguistic features, cultural heritage, and values by leveraging a teacher LLM. **(ii)** The public release of new datasets, representing the largest publicly available corpora for Egyptian

[1] https://github.com/UBC-NLP/nilechat.

and Moroccan Arabic dialects. These resources are intended to foster further research in these under-resourced languages. **(iii)** The development and public release of *NileChat*, a robust 3-billion parameter LLM. This model demonstrates proficiency in both Egyptian and Moroccan dialectal Arabic (using Arabic script and Arabizi) while maintaining strong performance in Modern Standard Arabic, French, and English.

2 Related Work

Adaptation of LLMs. LLMs, despite general strengths, often require adaptation for specific languages, domains, or cultures (Bang et al., 2023; AlKhamissi et al., 2024; Naous et al., 2024; Song et al., 2025). Adaptation techniques include prompt engineering (Shen et al., 2024), SFT on culturally specific datasets (Huang et al., 2024), and continued pre-training on target-specific data (Fujii et al., 2024; Huang et al., 2024). A key challenge, especially for SFT-based cultural adaptation, is the scarcity of comprehensive cultural datasets, hindering alignment with under-represented communities (Ahmad et al., 2024; Shen et al., 2024).

Synthetic Data Augmentation for LLMs. To address data limitations, synthetic data augmentation has shown promise in improving LLM performance (Ge et al., 2024; Li et al., 2024; Joshi et al., 2025). Machine-translated data, for instance, can enhance capabilities in new languages (Joshi et al., 2025; Shang et al., 2025), and persona-driven synthetic data generation has also yielded performance gains (Ge et al., 2024) and aided in tasks like assessing LLM political alignment (Bernardelle et al., 2024). However, synthetic data can sometimes degrade performance (Seddik et al., 2024), necessitating best practices for its use (Liu et al., 2024).

Arabic LLMs. In Arabic LLM development, models are either trained from scratch (Billah Nagoudi et al., 2023; Sengupta et al., 2023) or adapted from existing ones (Huang et al., 2024; Bari et al., 2025; Team et al., 2025a). A common method involves translating English data to Arabic, which, however, can introduce cultural biases from the source language (Sengupta et al., 2023; Naous et al., 2024). Recent work on dialectal Arabic, such as translating instructions into Moroccan dialect for SFT, has improved generation tasks (Shang et al., 2025). Yet, enhanced performance on standard tasks does not guarantee cultural awareness. While models like AceGPT (Huang et al., 2024) and Fanar (Team et al., 2025a) aim for cultural cognizance, our work uniquely focuses on adapting existing LLMs to a local community by deeply integrating its specific linguistic features, cultural heritage, and values, building upon these prior advancements.

3 Methodology

In this work, we investigate the potential of pre-training data to imbue LLMs with the specific local characteristics of under-represented communities. We conceptualize these characteristics along three primary dimensions (Geertz, 1977; Anderson, 1991; Bourdieu and Thompson, 1991; Higgins and Douglas, 2020; Stanlaw and Adachi, 2025): *(i) Language:* Encompassing dialectal nuances, idiomatic expressions, and linguistic structures unique to the community. *(ii) Cultural Heritage:* Reflecting the customs, traditions, social norms, historical context, and common knowledge prevalent within the community. *(iii) Cultural Values:* Capturing the ethical standpoints, belief systems, and societal priorities that define the community. We refer to these three dimensions as Language-Heritage-Values dimensions, *LHV* for short. While we do not posit these as exhaustive of the attributes of a given community, we employ them as a vehicle to approximate the LLM communication and information needs at local levels. To ground our investigation, we focus on two low-resource varieties of Arabic—The Egyptian Arabic (EGY) and Moroccan Arabic (MOR)—. These dialects serve as our primary case studies for evaluating the methods proposed herein.

3.1 Data Augmentation

The construction of linguistically-rich and culturally-rich LLMs that can serve a specific population fundamentally depends on the availability of representative data. Recognizing the acute scarcity of publicly available pre-training corpora for many low-resource languages, including EGY and MOR, we propose a novel data production method encapsulating the *LHV* dimensions of a given country-level population. As depicted in Figure 1, our approach leverages three complementary strategies intended to collectively capture the *LHV* dimensions: *(a) machine translation (MT)*, *(b) controlled synthetic data generation* and *(c) retrieval*. We explain these next.

3.1.1 MT for Knowledge and Fluency

To ensure linguistic fluency and coherence, we translate structured educational content from English into the target low-resource language using a specialized teacher model. Our pipeline preserves original formatting and includes filtering to remove unreliable translations identified by repetitive n-grams. We use educational materials for their topical breadth (including subjects such as education, history, health, medicine, and biology).

3.1.2 Controlled Synthetic Data Generation for Cultural Heritage and Cultural Values

Linguistic fluency, while a foundational capability for LLMs, does not inherently guarantee their awareness of, or alignment with, the culture and values of a specific target community (Naous et al., 2024). To bridge this gap, we employ *controlled synthetic data generation*. For controlled generation, we use the teacher LLM to generate diverse texts in the target language. These texts are specifically designed to discuss local topics, which are identified from articles sourced from local news websites or the target country's Wikipedia portal. Furthermore, the generated content is crafted to reflect distinct personas, each defined by a profile encompassing specific moral values, demographic characteristics, and socioeconomic attributes. Our approach integrates four key components to achieve this:

Local Contextual Information. We ground our synthetic data by incorporating local context drawn from news websites within the target communities. These sources provide relevant contextual information and do not necessarily need to be in the target local language.

Core Cultural Heritage Concepts. We integrate key local cultural elements, such as cuisine, landmarks, and celebrations, by extracting relevant articles from country-specific Wikipedia portals.

Linguistic and Cultural Expressions. To authentically capture local idiomatic styles, we collect common expressions, proverbs, idioms, dialogues from TV programs, and local terminology, pairing each with English translations for accuracy.

Representative Personas. We develop representative personas reflecting local moral, demographic, and socioeconomic attributes by leveraging data from the World Values Survey (WVS) (EVS/WVS, 2024). Selected survey responses are transformed into textual descriptions, which are further refined by an LLM to create concise and coherent persona profiles (see Figure 2).

To produce diverse text genres for pre-training, we combine data points from the four listed components into a unified prompt to guide the teacher LLM. This prompt instructs the LLM to generate varied text outputs in the target low-resource language, explicitly integrating the selected persona's values, the specified cultural concepts, and provided linguistic cues. Specifically, we focus on generating the following genres: *stories*, *personal essays*, *blog posts*, *reviews*, and *conversations*. An example of this process is depicted in Figure 1 (b).

3.1.3 Retrieval for Local Cultural Heritage

This method involves querying a search engine using a pre-defined list of cultural concepts that span multiple cultural categories. For each concept, we extract the top 20 search results, systematically excluding social media platforms. The textual content from the retrieved web pages is then parsed and extracted using Trafilatura (Barbaresi, 2021).

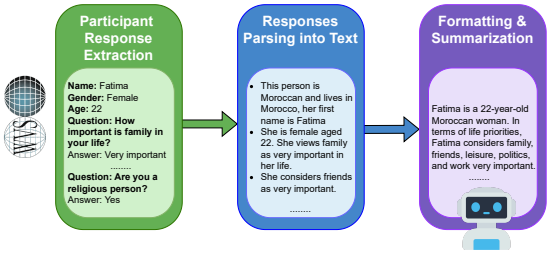

Figure 2: Pipeline for generation of persona descriptions using the WVS.

3.2 Model Training

We evaluate our proposed method on Egyptian (EGY) and Moroccan (MOR) Arabic dialects. Despite their large speaker populations, these dialects remain low-resource, underscoring the need for specialized language models. We select Command R+ (Cohere Labs, 2024) (104B) as our teacher model, as it demonstrates reasonable text-generation capabilities in both target dialects. Additionally, Command R+ provides open weights, enabling us to efficiently generate or translate extensive datasets without incurring API costs.

3.2.1 Continued Pre-Training

Data. We generate pre-training data for EGY and MOR using the methods outlined in Section 3.1. Our approach involves three main components: **(i) MT Data.** We employ our teacher model to translate English educational content into both dialects.

Specifically, we sample 5.5 million texts from the Fineweb-edu dataset (Penedo et al., 2024) and translate them into EGY and MOR. **(ii) Controlled Synthetic Generation Data.** We craft tailored prompts incorporating personas, local cultural contexts, dialectal glosses, expressions, and utterances to instruct the LLM in generating diverse genres of text. For *persona* descriptions, we generate 1,200 descriptions based on data from Egyptian and Moroccan participants in the WVS. For *local news context*, we leverage an in-house corpus comprising approximately 1.5 million Egyptian and 800,000 Moroccan news articles, originally published in MSA by local news websites. Additionally, we include 25,000 Egyptian and 49,000 Moroccan Wikipedia articles. For *dialectal glosses, expressions, and utterances*, we draw from publicly available resources on EGY and MOR proverbs and idiomatic expressions, each accompanied by English explanations. We further augment this with an in-house dataset of 600 dialectal utterances from Egyptian and Moroccan television shows paired with English translations, as well as 4,000 dialect-to-English word pairs for each dialect from the Gatitos dictionary (Jones et al., 2023). **(iii) Retrieval Data.** For information retrieval, we query the Brave Search API[2] using 6,500 cultural concepts from Morocco and 4,500 cultural concepts from Egypt. These concepts represent the ten cultural heritage categories in the set *{food, clothes, landmarks, festivals & celebrations, geography, handicrafts, architecture, fauna, flora, music}*.

The generated dataset comprises approximately 5.5 million educational articles for both EGY and MOR. Additionally, for EGY, it includes approximately 300,000 samples for each category of conversations, personal essays, blog posts, reviews, and stories. For MOR, there are approximately 150,000 samples for each of these same categories. These latter categories represent the *LHV* dimensions (§3.1). Table A.3 presents a sample of these texts.

A filtering process using a repetitive n-gram filter removed 3.97% of the data. We also conducted a dialectness check on the generated data using ALDi (Keleg et al., 2023). The average dialectness scores for the EGY and MOR educational articles are 0.45 and 0.32, respectively. In contrast, for the texts focused on cultural heritage and values, the average dialectness scores are higher, at 0.84 for EGY and 0.72 for MOR. We attribute the lower dialectness levels in the educational articles to the prevalence of scientific terms that often lack direct equivalents in EGY and MOR, and were therefore retained in MSA. We convert 1.5M EGY and 0.5M MOR samples from the generated data to Arabizi. For retrieval, we collect 110,000 and 30,000 articles about cultural heritage for both EGY and MOR.

Our final pre-training dataset is a mixture of our generated and retrieved data, combined with pre-existing publicly available data for these dialects, MSA, English, French, Math, and Code. Our objective is to preserve the data distribution of the base model's pre-training data to mitigate catastrophic forgetting (Luo et al., 2025). The resulting pre-training dataset comprises 98.57 billion words, and its composition is detailed in Table A.1.

Compute. We used a cluster of 4×A100 80GB GPUs for 1,096 hours to create our augmented pre-training dataset using the listed inputs.

Continued Pre-training. Rather than pretraining an LLM from scratch, we continue pretraining Qwen-2.5-3B (Qwen et al., 2025) with our data. We select this model due to its competitive performance and good tokenizer compression ratio on MSA. We continue pretraining the full model (3.1B parameters) for one whole epoch, which took 750 hours on 4×A100 80GB GPUs. More details about the base model selection and the training are in Appendix B.1.

3.2.2 Supervised Fine-Tuning

To adapt our pre-trained model for instruction following, we perform supervised fine-tuning (SFT).

Data. Due to the scarcity of SFT datasets for EGY and MOR, we construct a comprehensive training set. This process involves several key steps: (i) translation of SmolTalk dataset (Allal et al., 2025) into MOR, EGY, French, and MSA using the teacher LLM; (ii) synthetic generation of dialectal question-answer pairs using our retrieved dataset of local Egyptian and Moroccan cultural heritage;[3] (iii) incorporation of the Darija-SFT-Mixture MOR dataset provided by Shang et al. (2025); and (iv) translation of TULU-V2-mix dataset (Ivison et al., 2023) into EGY. Finally, (v) this consolidated SFT

[2] https://brave.com/search/api/

[3] This data is initially created in Arabic script, and a portion is subsequently converted to Arabizi.

dataset is augmented by converting understanding and generation tasks from the training sets of the ORCA (Elmadany et al., 2023) and Dolphin (Nagoudi et al., 2023) benchmarks into instruction-response formats. The final composition of our instruction dataset is in Table A.2.

Fine-Tuning. For model SFT, we follow recent approaches (Ramé et al., 2024; Dang et al., 2024) that leverage model merging techniques to produce models effective across multiple languages or tailored for particular tasks. Specifically, we fully fine-tuned two separate variants of the base model—one specialized for MOR and the other for EGY—each trained on its respective dialectal data in both Arabic script and Arabizi (plus an amount of shared data between the two variants from the other languages; see B.2). We fine-tune each dialect-specific model for two epochs and employ weighted linear averaging (Aakanksha et al., 2024) for merging, dubbing our merged model *NileChat*.

More information about our model merging is in B.2 and the prompts used for generating and translating our pre- and fine-tuning datasets is in E.

4 Experiments

4.1 Evaluation Tasks

We employ a comprehensive evaluation framework to measure the performance of *NileChat* for EGY and MOR. This framework enables comparison with our baseline and other LLMs across multiple capability dimensions: Understanding, cultural knowledge, translation, and value alignment.

Understanding. We evaluate understanding capabilities using MMLU (Hendrycks et al., 2021), HellaSwag (Zellers et al., 2019), and Belebele (Bandarkar et al., 2024) benchmarks, each adapted to both EGY and MOR dialects. For MOR, we directly employ the MMLU and HellaSwag versions provided by Shang et al. (2025). For EGY, we follow the translation pipeline described in Shang et al. (2025), translating the English and MSA MMLU tasks and the English HellaSwag dataset into EGY using our teacher model [4]. A careful verification of the translation quality for the generated EGY MMLU and EGY HellaSwag shows that the average correctness is 3.85 on a scale from 1-5 and the average dialectness is approximately 4 on a scale from 1-5. Further details are provided in Appendix C.1. For the Belebele benchmark, we utilize the official Moroccan and Egyptian dialect sets. Evaluations are conducted in both zero-shot and 3-shot scenarios, using accuracy as our performance metric.

Cultural Knowledge. To assess cultural knowledge specific to Morocco and Egypt, we utilize the publicly available test set from the Palm benchmark (Alwajih et al., 2025), focusing on these two countries only. We adopt an LLM-as-Judge methodology (Zheng et al., 2023), employing Gemma-3-27b (Team et al., 2025b) to rate the correctness of model-generated responses compared to ground-truth answers on a scale from 0 to 10. The final evaluation score is calculated as the average correctness across all responses.

Translation. We evaluate the translation performance across multiple directions: *dialect*↔*dialect* (i.e., Moroccan↔Egyptian), *dialect*↔*MSA*, *English*↔*dialect*, and *French*↔*dialect*. Our primary benchmark is the Flores-200 dataset (Team et al., 2022), comprising $1,012$ test examples per translation direction. Additionally, we introduce an in-house, human-curated dataset consisting of 300 authentic EGY and MOR utterances transcribed from local television programs then translated to MSA and English. This dataset provides a more accurate reflection of natural, colloquial language usage compared to Flores-200, which primarily contains Wikipedia-based sentences. We conduct evaluations in both zero-shot and 4-shot settings, reporting results using ChrF++ (Popović, 2015) and spBLEU scores (Goyal et al., 2022).

Value Alignment. To assess alignment with societal values, we adapt WVS questions into a multiple-choice format (expressed in the local language). The questions are categorized into 13 dimensions such as *Economic Values (EcoV)*, *Ethical Values (EthV)*, and *Happiness and Wellbeing (HW)*.[5] We use the Social Value Alignment (SVA) metric (Lee et al., 2024), which measures alignment using the distribution of survey responses. A model's alignment score for each question corresponds to the proportion of participants who chose

[4] We have publicly released the EgyMMLU and EgyHellaswag benchmarks for evaluation on the LM Evaluation Harness Framework at https://github.com/EleutherAI/lm-evaluation-harness/tree/main/lm_eval/tasks/egymmlu and https://github.com/EleutherAI/lm-evaluation-harness/tree/main/lm_eval/tasks/egyhellaswag.

[5] See Appendix C.1 for the full list.

the model-predicted option, averaged across all questions for the final score.

Baseline Models. We compare *NileChat* against a set of 17 instruction-tuned LLMs known for their strong capabilities in Arabic, capped at 13B parameters (see full list in Table C.2 and Appendix C.2 for details).[6]

4.2 Results and Discussion

Understanding. As Table 1 shows, *NileChat* demonstrates SoTA performance on the MMLU, HellaSwag, and Belebele benchmarks for both EGY and MOR when compared to similar size models. Specifically, *NileChat* surpasses its baseline model, Qwen2.5-3B-instruct, by a significant margin of $\sim$10 points across the majority of these tasks. Notably, *NileChat* also outperforms larger Arabic-focused models such as AceGPT-13B and Jais-13B. Furthermore, it achieves on-par performance with recent leading Arabic LLMs like ALLaM-7B, with a performance gap of less than 3 points on most tasks, and even surpasses it on certain benchmarks, including Belebele. Results for 3-shot are presented in Table D.1 and they show a similar trend to the zero-shot ones.

Cultural Knowledge. As shown in Table 1, our approach significantly enhances cultural knowledge (Palm), enabling *NileChat* to achieve scores of 5.72 (EGY) and 5.86 (MOR), compared to baseline Qwen2.5-3B-instruct scores of 2.86 and 2.31, respectively. Among similarly sized models, ours achieves the highest performance on MOR and ranks second only to Gemma-3-4B for EGY. Although larger models such as Gemma-3-12B exhibit superior overall scores (EGY: 8.71, MOR: 7.09), *NileChat* notably surpasses AceGPT-7B and -13B on Moroccan cultural knowledge, despite their claimed alignment with Arabic cultures. Additionally, it outperforms Atlas-chat-2B and -9B, models specifically fine-tuned for Moroccan dialects. *These results support our claim that linguistic fluency alone—gained through supervised fine-tuning or pre-training on potentially biased, translated datasets—is insufficient for genuine cultural alignment with local communities.*

Translation. Table 2 summarizes the spBLEU scores from our zero-shot translation. Overall, *NileChat* achieves the highest average translation quality (spBLEU: 21.32), outperforming all evaluated models, including larger alternatives such as ALLaM-7B (20.60) and NLLB-200-3.3B (18.29). Specifically, on the Flores benchmark, *NileChat* demonstrates comparable performance to the similarly-sized NLLB-200-3.3B, with only a marginal 1-point spBLEU difference aggregated across MOR and EGY. Notably, *NileChat* surpasses even larger competitors in all translation directions, except when translating into MOR, where its performance matches that of Atlas-Chat-9B—a larger, single-dialect-focused model that is 3X larger.

On our in-house, human-curated dataset—which closely represents authentic speech patterns from local populations—*NileChat* significantly outperforms all baselines, including NLLB-200-3.3B, in all translation directions for both EGY and MOR. This real-world evaluation emphasizes the effectiveness of our strategy to incorporate local linguistic and cultural elements into synthetic data generation, enriching the pre-training data with diverse dialectal expressions and vocabulary. Detailed results for both zero-shot and 4-shot translation experiments are provided in Table D.2.

Value Alignment. Figure 3 illustrates the results of value alignment evaluation based on the WVS. *NileChat* demonstrates substantial improvements over the baseline across most societal-value dimensions for both Moroccan and Egyptian contexts. Specifically, for Morocco, *NileChat* surpasses the baseline in all dimensions except Religious Values and the Index of Postmaterialism. Similarly, for Egypt, it outperforms the baseline across all dimensions except Political Interest and Political Participation, and the Index of Postmaterialism. These findings indicate that our approach—where a teacher LLM engages in role-playing by generating diverse text genres through personas embodying local community values—successfully steers responses towards culturally aligned positions. In a broader comparative analysis against all evaluated models, ours achieves the best results for Morocco across several dimensions and remains competitive in others. For Egypt, *NileChat* notably excels in Perceptions of Migration, Political Culture and Political Regimes, Happiness and Wellbeing, and Perceptions about Science and Technology, though models such as Jais-13B and ALLaM-7B show slightly stronger performance in certain other dimensions.

[6]We also evaluate our translation performance against an NLLB-200's 3.3B variant (Team et al., 2022).

	Model	MMLU		HellaSwag		Belebele		Palm	
		EGY	MOR	EGY	MOR	EGY	MOR	EGY	MOR
Less than 7B	**Qwen3-1.7B**	28.53	28.53	28.44	27.47	22.89	22.89	3.61	2.12
	ar-stablelm-2-chat	41.56	40.36	34.79	33.45	38.89	36.11	4.20	3.62
	Atlas-Chat-2B	42.61	44.87	29.66	34.74	50.56	55.67	3.16	3.42
	Llama-3.2-3B-Instruct	40.68	37.54	29.16	28.27	45.44	35.89	3.21	2.28
	gemma-3-4b-it	40.79	32.70	34.21	31.35	37.33	34.22	**7.61**	5.42
	Qwen3-4B	28.61	28.54	30.28	29.04	22.89	22.89	4.51	2.71
	Qwen2.5-3B-Instruct	43.37	44.43	31.62	29.58	51.33	41.44	2.86	2.31
	NileChat (3B)	**57.56**	**57.36**	**37.97**	**39.33**	**72.67**	**70.33**	5.72	**5.86**
More than 7B	**AceGPT-7B-chat**	40.29	37.57	33.27	30.47	32.67	32.00	5.58	3.93
	ALLaM-7B-Instruct	60.04	58.72	39.40	37.30	69.56	57.78	6.78	6.14
	Qwen2.5-7B-Instruct	49.65	44.98	34.67	32.16	64.22	48.56	6.70	4.77
	Qwen3-8B	28.53	28.53	31.76	30.32	22.89	22.89	5.88	3.96
	Atlas-Chat-9B	55.17	58.84	33.71	44.34	70.33	74.11	5.24	4.84
	gemma-3-12b-it	61.17	60.00	38.59	35.66	75.78	64.89	8.76	7.09
	AceGPT-13B-chat	45.45	40.68	35.06	32.40	38.78	36.44	6.10	4.83
	jais-13b-chat	49.79	48.10	39.02	36.56	64.22	53.78	5.66	4.80

Table 1: Zero-shot performance of models on understanding and cultural knowledge evaluations. Metrics are accuracy for MMLU, HellaSwag, and Belebele, and a 0-10 correctness score for Palm. Bold values indicate the highest score among models comparable in size to ours (< 7B). Underlined values represent the highest score in the entire column, including larger models.

	Model	Flores-200				In-House Data				Average
		XX →		→ XX		XX →		→ XX		
		EGY	MOR	EGY	MOR	EGY	MOR	EGY	MOR	
Less than 7B	**Qwen3-1.7B**	14.75	10.89	19.51	15.47	11.41	4.36	15.63	6.32	12.29
	ar-stablelm-2-chat	14.35	7.07	11.10	9.72	9.23	2.92	11.23	7.73	9.17
	Atlas-Chat-2B	15.20	13.40	21.39	21.11	5.36	7.83	14.52	13.54	14.05
	Llama-3.2-3B-Instruct	14.25	9.15	19.28	15.54	10.67	3.16	13.61	4.87	11.32
	gemma-3-4b-it	9.27	5.22	12.46	10.13	3.01	0.60	16.89	5.25	7.86
	Qwen3-4B	17.93	11.64	20.03	18.90	13.09	4.44	20.72	8.52	14.41
	NLLB-200-3.3B	**23.93**	15.37	**25.84**	**26.57**	16.77	7.49	18.90	11.43	18.29
	Qwen2.5-3B-Instruct	15.14	11.27	20.52	17.37	9.91	4.19	19.24	7.83	13.18
	NileChat (3B)	23.60	**16.41**	25.74	25.56	**22.02**	**12.34**	**26.50**	**18.39**	**21.32**
More than 7B	**AceGPT-7B-chat**	18.02	11.33	21.11	17.46	14.73	4.95	20.10	7.47	14.40
	ALLaM-7B-Instruct	23.91	15.88	24.74	23.19	19.98	9.16	29.40	18.51	20.60
	Qwen2.5-7B-Instruct	14.41	10.23	19.81	18.95	10.43	4.10	20.92	8.80	13.46
	Qwen3-8B	20.03	13.86	22.56	21.33	13.38	4.73	24.14	9.27	16.16
	Atlas-Chat-9B	18.20	16.89	24.92	26.29	5.36	7.68	17.35	15.23	16.49
	gemma-3-12b-it	13.01	4.89	19.05	19.54	7.86	2.45	24.51	12.38	12.96
	AceGPT-13B-chat	19.48	14.02	22.81	19.84	15.54	5.56	23.51	9.52	16.29
	jais-13b-chat	8.80	4.29	15.77	17.12	10.83	4.02	19.19	12.47	11.56

Table 2: Zero-shot translation performance (spBLEU) on the Flores and in-house datasets. XX → EGY and XX → MOR denote average over target languages EGY and MOR, respectively. Conversely, EGY → XX and MOR → XX indicate average over EGY and MOR as source languages. Bold values highlight the top score among models with fewer than 7 billion parameters. Underlined values indicate the highest score overall in each column. Detailed results are in Table D.2.

How many pre-training tokens are needed to reach good performance for a new language? Figure 4 shows the performance evolution of *NileChat* during the pre-training phase on Belebele and translation tasks. The charts show that the model starts to get a large boost in these tasks

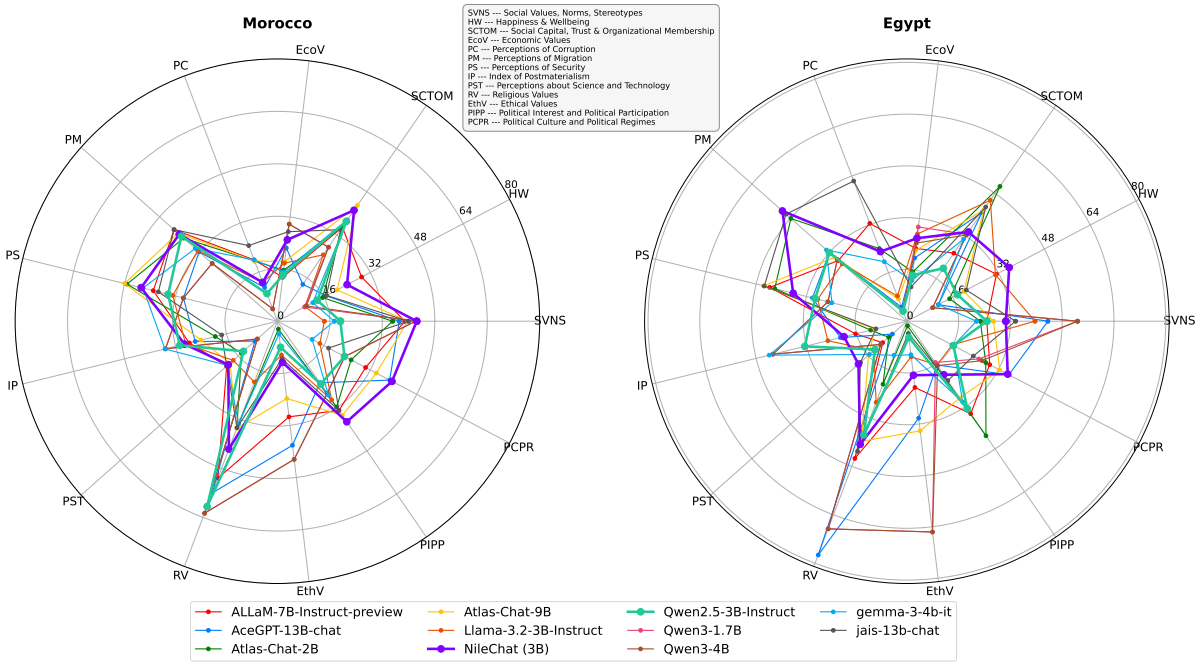

Figure 3: Average SVA scores of evaluated models across societal value dimensions for Egypt and Morocco.

during the first 10B tokens and then continues to slightly increase until it becomes steady after around 60B tokens.

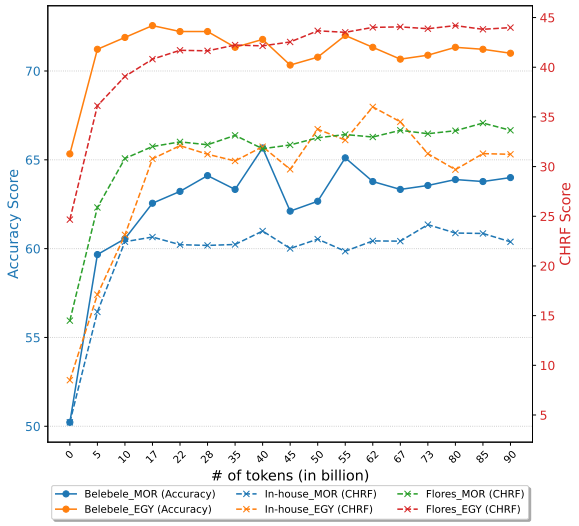

Figure 4: Evolution of model performance during pretraining, measured by the number of tokens processed.

5 Conclusion

We introduced a novel methodology for adapting LLMs to specific communities by deeply integrating their unique linguistic characteristics, cultural heritage, and societal values. Our approach leverages a teacher model proficient in generating low-resource languages to enable: (i) translation for the incorporation of community-specific language and (ii) controlled generation and retrieval mechanisms for the authentic inclusion of cultural heritage and values. We validated our methodology using the Moroccan and Egyptian Arabic dialects as testbeds by developing *NileChat*, an LLM covering these two dialects. Comprehensive evaluations on understanding, translation, and cultural alignment benchmarks demonstrate that our method significantly enhances the baseline LLM's performance in capturing target language nuances and cultural values. Notably, *NileChat* also outperforms existing Arabic-aware LLMs. Our method offers a promising research direction for fostering inclusivity of diverse local communities within LLM development, thereby emphasizing the critical role of such an inclusion in the broader democratization of this technology.

Limitations

- **Teacher Model Dependency for Low-Resource Languages:** Our method's reliance on a teacher model proficient in generating even low-resource target languages may not hold for extremely under-resourced languages (e.g., Berber, Malayo-Polynesian varieties) (Team et al., 2022), potentially limiting its applicability in such contexts.

- **Supervised Fine-Tuning Data:** SFT phase

predominantly utilized translated data due to resource constraints. This reliance on translated, rather than native, data for SFT might impact the model's nuanced performance in the target languages and their corresponding cultures.

- **Susceptibility to Hallucination:** As a 3B parameter model, our LLM is relatively small, rendering it more prone to hallucination and the generation of inaccurate or incomplete information compared to larger architectures (Wei et al., 2022).

- **Computational Cost of Synthetic Data Generation:** The process of generating synthetic data is computationally intensive, particularly when employing large teacher models (e.g., Command R+, a 104B parameter model requiring substantial GPU resources: 4x80GB). This challenge is amplified by the autoregressive generation of long documents from extensive input contexts (e.g., articles, persona descriptions, cultural concepts) restricting the scale of this approach for more languages.

- **Absence of Explicit Safety Alignment:** The model has not undergone dedicated safety alignment. While trained on curated datasets (Wikipedia, educational, news) largely devoid of toxic content and leveraging a safety-aligned teacher LLM, specific safety tuning is acknowledged as important future work.

- **Limited generation of subtle details.** While the controlled generation uses multiple sources (WVS, news, Wikipedia, TV scripts), the generated texts are limited in terms of the very subtle cultural nuances, implicit knowledge, humor, or sarcasm that are often not explicitly stated in these source materials (Wikipedia and news articles).

Ethics Statement

Our work contributes to the development of inclusive, linguistically, and culturally diverse LLMs capable of serving varied communities. While we generate our pre-training and instruction-tuning data using a teacher LLM, this process is critically informed by ground-truth cultural values survey data from the communities of interest and local context to control the generation. This approach aims to imbue our models with specific cultural nuances relevant to these communities.

As our evaluations demonstrate, the resulting models exhibit reasonable alignment with the cultural heritage and values of our target communities and can produce fluent text in their respective dialects. Despite these advancements, we have not conducted explicit safety alignment procedures for these models. Consequently, we strongly recommend thorough testing and further safety evaluations before any deployment in real-world scenarios.

Acknowledgments

Muhammad Abdul-Mageed acknowledges support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[7] and UBC Advanced Research Computing-Sockeye.[8]

References

Aakanksha, Arash Ahmadian, Seraphina Goldfarb-Tarrant, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. 2024. Mix data or merge models? optimizing for diverse multi-task learning. *Preprint*, arXiv:2410.10801.

Ibrahim Ahmad, Shiran Dudy, Resmi Ramachandranpillai, and Kenneth Church. 2024. Are generative language models multicultural? a study on Hausa culture and emotions using ChatGPT. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 98–106, Bangkok, Thailand. Association for Computational Linguistics.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin

[7] https://alliancecan.ca

[8] https://arc.ubc.ca/ubc-arc-sockeye

Raffel, Leandro von Werra, and Thomas Wolf. 2025. Smollm2: When smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737.

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, Abdelrahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibaee, Yasir Ech-Chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, Ahmed Oumar El-Shangiti, Aisha Alraeesi, Mohammed Anwar Al-Ghrawi, Abdulrahman S. Al-Batati, Elgizouli Mohamed, Noha Taha Elgindi, Muhammed Saeed, Houdaifa Atou, Issam Ait Yahia, Abdelhak Bouayad, Mohammed Machrouh, Amal Makouar, Dania Alkawi, Mukhtar Mohamed, Safaa Taher Abdelfadil, Amine Ziad Ounnoughene, Rouabhia Anfel, Rwaa Assi, Ahmed Sorkatti, Mohamedou Cheikh Tourad, Anis Koubaa, Ismail Berrada, Mustafa Jarrar, Shady Shehata, and Muhammad Abdul-Mageed. 2025. Palm: A culturally inclusive and linguistically diverse dataset for arabic llms. *Preprint*, arXiv:2503.00151.

Zaid Alyafeai, Michael Pieler, Hannah Teufel, Jonathan Tow, Marco Bellagente, Duy Phung, Nikhil Pinnaparaju, Reshinth Adithyan, Paulo Rocha, Maksym Zhuravinskyi, and Carlos Riquelme. 2024. Arabic stable lm: Adapting stable lm 2 1.6b to arabic. *Preprint*, arXiv:2412.04277.

Benedict Anderson. 1991. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London.

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)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. 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*.

Emily M. Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Pietro Bernardelle, Leon Fröhling, Stefano Civelli, Riccardo Lunardi, Kevin Roitero, and Gianluca Demartini. 2024. *arXiv preprint arXiv:2412.14843*.

El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, AbdelRahim Elmadany, Alcides Inciarte, and Md Tawkat Islam Khondaker. 2023. JASMINE: Arabic GPT models for few-shot learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16721–16744, Singapore. Association for Computational Linguistics.

P. Bourdieu and J.B. Thompson. 1991. *Language and Symbolic Power*. Harvard University Press.

Cohere Labs. 2024. c4ai-command-r-plus-08-2024.

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.

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. EtiCor: Corpus for analyzing LLMs for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931, Singapore. Association for Computational Linguistics.

AbdelRahim Elmadany, ElMoatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. ORCA: A challenging benchmark for Arabic language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9559–9586, Toronto, Canada. Association for Computational Linguistics.

EVS/WVS. 2024. Joint evs/wvs 2017-2022 dataset (joint evs/wvs). GESIS, Cologne. ZA7505 Data file Version 5.0.0, https://doi.org/10.4232/1.14320.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In *First Conference on Language Modeling*.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Aswanth Kumar M, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M. Khapra, Raj Dabre, Rudra Murthy, and Anoop Kunchukuttan. 2024. Airavata: Introducing hindi instruction-tuned llm. *Preprint*, arXiv:2401.15006.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

C. Geertz. 1977. *The Interpretation Of Cultures*. Basic Books.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7765–7784, Miami, Florida, USA. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daniil Gurgurov, Mareike Hartmann, and Simon Ostermann. 2024. Adapting multilingual LLMs to low-resource languages with knowledge graphs via adapters. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 63–74, Bangkok, Thailand. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

V. Higgins and D. Douglas. 2020. *Communities and Cultural Heritage: Global Issues, Local Values*. Routledge studies in heritage. Routledge/Taylor & Francis Group.

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.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *Preprint*, arXiv:2311.10702.

Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. GATITOS: Using a new multilingual lexicon for low-resource machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar, and Eileen Long. 2025. Adapting multilingual LLMs to low-resource languages using continued pre-training and synthetic corpus: A case study for Hindi LLMs. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 50–57, Abu Dhabi. Association for Computational Linguistics.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. ALDi: Quantifying the Arabic level of dialectness of text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. Association for Computational Linguistics.

Jiyoung Lee, Minwoo Kim, Seungho Kim, Junghwan Kim, Seunghyun Won, Hwaran Lee, and Edward Choi. 2024. KorNAT: LLM alignment benchmark for Korean social values and common knowledge. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11177–11213, Bangkok, Thailand. Association for Computational Linguistics.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. 2024. Best practices and lessons learned on synthetic data. *arXiv preprint arXiv:2404.07503*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2025. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *Preprint*, arXiv:2308.08747.

Kathleen M. MacQueen, Eleanor McLellan, David S. Metzger, Susan Kegeles, Ronald P. Strauss, Roseanne Scotti, Lynn Blanchard, and Robert T. Trotter. 2001. What is community? an evidence-based definition for participatory public health. *American Journal of Public Health*, 91(12):1929–1938.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, Ahmed El-Shangiti, and Muhammad Abdul-Mageed. 2023. Dolphin: A challenging and diverse benchmark for Arabic NLG. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1404–1422, Singapore. Association for Computational Linguistics.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2).

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Preprint*, arXiv:2406.17557.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Qwen, :, 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.

Alexandre Ramé, Johan Ferret, Nino Vieillard, Robert Dadashi, Léonard Hussenot, Pierre-Louis Cedoz, Pier Giuseppe Sessa, Sertan Girgin, Arthur Douillard, and Olivier Bachem. 2024. Warp: On the benefits of weight averaged rewarded policies. *Preprint*, arXiv:2406.16768.

Jonathan Rystrøm, Hannah Rose Kirk, and Scott Hale. 2025. Multilingual != multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in llms. *Preprint*, arXiv:2502.16534.

E. Sapir. 1929. The status of linguistics as a science. *Language*, 5(4):207–214.

Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Abdelkader DEBBAH. 2024. How bad is training on synthetic data? a statistical analysis of language model collapse. In *First Conference on Language Modeling*.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, 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.

Guokan Shang, Hadi Abdine, Yousef Khoubrane, Amr Mohamed, Yassine Abbahaddou, Sofiane Ennadir, Imane Momayiz, Xuguang Ren, Eric Moulines, Preslav Nakov, Michalis Vazirgiannis, and Eric Xing. 2025. Atlas-chat: Adapting large language models for low-resource Moroccan Arabic dialect. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 9–30, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea.

2024. Understanding the capabilities and limitations of large language models for cultural commonsense. 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 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.

Zirui Song, Bin Yan, Yuhan Liu, Miao Fang, Mingzhe Li, Rui Yan, and Xiuying Chen. 2025. Injecting domain-specific knowledge into large language models: A comprehensive survey. *Preprint*, arXiv:2502.10708.

J. Stanlaw and N. Adachi. 2025. *Language, Culture, and Society: An Introduction to Linguistic Anthropology*. Taylor & Francis.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

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. 2025a. Fanar: An arabic-centric multimodal generative ai platform. *Preprint*, arXiv:2501.13944.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, et al. 2025b. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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, and Others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Shumin Wang, Yuexiang Xie, Bolin Ding, Jinyang Gao, and Yanyong Zhang. 2025. Language adaptation of large language models: An empirical study on LLaMA2. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7195–7208, Abu Dhabi, UAE. Association for Computational Linguistics.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

C. G. Wrenn. 1962. The counselor in a changing world.

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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang,

Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Language	Data category		Data source	Nature	# of words
Dialectal data	**Egypt**	General	Wikipedia	Real	128.71M
			MT fineweb-EDU	**Synthetic**	**2.08B**
			LHV	**Synthetic**	**398.89M**
			Fineweb2	Real	430.46M
		Arabizi	**MT fineweb-EDU (Arabizi) & LHV**	**Synthetic**	**206.49M**
	Morocco	General	Wikipedia	Real	1.67M
			Translated fineweb-EDU	**Synthetic**	**2.02B**
			LHV	**Synthetic**	**207.41M**
			Fineweb2	Real	1.64B
		Arabizi	**MT fineweb-EDU (Arabizi) & LHV**	**Synthetic**	**467.30M**
MSA	**Egypt**	Cultural	**Brave API**	**Real**	**74.67M**
		General	Local News	Real	346.79M
	Morocco	Cultural	**Brave API**	**Real**	**23.08M**
		General	Local News	Real	220.16M
	General	General	Fineweb2	Real	28.80B
			Wikipedia	Real	318.62M
English	**General**	General	Fineweb-EDU	Real	51.57B
French	**General**	General	Fineweb2	Real	9.42B
Code & Math		Code & Math	MathGenie/MathCode-Pile macrocosm-os/code-parrot-github-code	Real	818.35M

Table A.1: Distribution of the final pre-training data mixture by language, nature (synthetic vs. real), and word count per dataset. Bold rows highlight data generated via our proposed augmentation pipeline.

A Data

B Training details

B.1 Pre-training

Motivations for choosing Qwen-2.5-3B as our backbone model. We select Qwen-2.5-3B as our base model to continue pretrain for two primary reasons: its competitive performance on MSA tasks and good tokenizer compression ratio on Arabic dialect texts. At the time of the selection, the Gemma 2 (Team et al., 2024) and Qwen 2.5 base models showed the best performance in MSA. Also, both of their tokenizers have a good compression ratio for Arabic text in both standard and dialectal forms. Our analysis shows a ratio between 2.7 and 2.8 for Gemma, while Qwen 2.5 has a compression ratio between 2.9 and 3.1. Our final choice of Qwen 2.5 was based on its better performance on MSA.

Details for continued model pretraining. We continue the pre-training of Qwen-2.5-3B on our curated pre-training dataset. Subsequently, the model is fully fine-tuned for one epoch using a sequence length of 4,096. To optimize the learning process, the learning rate is linearly decayed from 5×10^{-6} to 5×10^{-7}. To mitigate overfitting, we apply a weight decay of 0.1, and gradient norms are clipped at a maximum value of 1.0. The training is performed on a cluster of 4×A100 80GB GPUs.

B.2 Supervised fine-tuning

To enhance model robustness and facilitate effective merging, we augmented each dialect-specific dataset with a shared multilingual corpus, comprising English SmolTalk, MSA SmolTalk, French SmolTalk, and additional data from the ORCA and Dolphin datasets. Each dialect-specific model was trained for two epochs with a sequence length of 4,096 tokens, using a learning rate that linearly decayed from 7×10^{-6} to 7×10^{-7}.

Table B.1 compares the SFT model with models fine-tuned on Egyptian and Moroccan datasets individually, as well as with our final merged model, *NileChat*. *NileChat* performs well on tasks for both EGY and MOR. The MOR-specific model also demonstrates strong performance on both MOR and, to some extent, EGY tasks. In contrast, the EGY-specific model does not perform well on MOR tasks.

Dataset name	Language	Source	# of Instructions
Darija-SFT-Mixture	MOR (Arabic)	Atlas-Chat	458,155
TÜLU-V2-mix	EGY (Arabic)	Ours (MT)	178,109
SmolTalk	MOR (Arabic)	Ours (MT)	192,266
	MOR (Arabizi)	Ours (MT)	93,419
	EGY (Arabic)	Ours (MT)	195,260
	EGY (Arabizi)	Ours (MT)	93,181
	French	Ours (MT)	99,468
	MSA	Ours (MT)	96,933
	English	SmolTalk	149,124
ORCA	MSA + dialects	Ours (Converted)	460,203
Dolphin		Ours (Converted)	425,703
Cultural instructions	MOR (Arabic)	Ours (Synthetic)	25,159
	EGY (Arabic)	Ours (Synthetic)	107,428

Table A.2: Distribution of the final instruction and response data mixture by language and number of instructions per dataset. 'Ours' refers to datasets we created via machine translation (MT) or by converting existing datasets into an instruction/response format.

Table B.1 compares the SFT model with models fine-tuned on Egyptian and Moroccan datasets individually, as well as with our final merged model, NileChat. NileChat performs well on tasks for both EGY and MOR. The MOR-specific model also demonstrates strong performance on both MOR and, to some extent, EGY tasks. In contrast, the EGY-specific model does not perform well on MOR tasks. We relate this observed asymmetry to the linguistic characteristics of the dialects relative to MSA. During the SFT phase, each dialect-specific dataset was augmented with a shared multilingual corpus which included MSA data (e.g., MSA SmolTalk, and data from the ORCA and Dolphin datasets). It is plausible that EGY is linguistically closer to MSA compared to the MOR, which is often considered more distant from MSA due to influences such as Berber and French. Consequently, the MOR-tuned model, having been exposed to this shared MSA data, might more effectively leverage this MSA knowledge to generalize to EGY tasks. Conversely, the greater linguistic divergence of the Moroccan dialect from MSA could make it more challenging for the EGY-tuned model to transfer its learning, including the MSA component, to the distinct features of the Moroccan dialect.

C Evaluation Setup

C.1 Evaluation Tasks

Full list of the 13 categories of WVS questions. Economic Values (EcoV); Ethical Values (EthV); Happiness and Wellbeing (HW); Index of Postmaterialism (IP); Perceptions about Science and Technology (PST); Perceptions of Corruption (PC); Perceptions of Migration (PM); Perceptions of Security (PS); Political Culture and Political Regimes (PCPR); Political Interest and Political Participation (PIPP); Religious Values (RV); Social Capital, Trust, and Organizational Membership (SCTOM); and Social Values, Norms, and Stereotypes (SVNS).

The Quality and Validation of Generated Datasets. We rigorously validated the two evaluation Egyptian Arabic datasets we created using machine translation, namely *EgyMMLU* and *EgyHellaSwag*. A random sample of items from each dataset was rated on two 1–5 scales: (i) *Correctness* (semantic accuracy/faithfulness) and (ii) *Dialectness* (authenticity and naturalness in Egyptian Arabic). Expert human annotator rated 100 randomly sampled items per dataset. To increase statistical power, a state-of-the-art LLM judge (Gemini 2.5 Pro) rated the same 100 items plus an additional 300 (400 total) per dataset. Agreement between human and LLM scores on the 100 shared items yielded an Intraclass Correlation Coefficient (ICC) of **0.60**, indicating good reliability. Summary scores appear in Table C.1. These results,

together with the inter-evaluator agreement, indicate that the newly created Egyptian benchmarks are of high quality and suitable for reliable model evaluation.

C.2 Baselines

We evaluate our model *NileChat* against a set of 17 LLMs that are Arabic-aware; some of these 17 models are also aligned to Arabic dialects. These models are from the following model families: ALLaM (Bari et al., 2025), Jais (Sengupta et al., 2023), Atlas-Chat (Shang et al., 2025), ar-stablelm-2-chat (Alyafeai et al., 2024), Gemma-3 (Team et al., 2025b), Qwen-2.5 (Qwen et al., 2025), Qwen3 (Yang et al., 2025) (non-thinking mode), and Llama-3.2 (Grattafiori et al., 2024). The full list of models, including their corresponding size and release date, are presented in Table C.2.

D Full results

E Prompts

The provided figures showcase diverse prompts for language models targeting low-resource languages. Figure 5 translates English educational content into conversational dialectal Arabic, while Figure 6 converts dialectal Arabic script to Arabizi. English instructions are translated to dialectal Arabic using the prompt in Figure 7. For content generation, Figure 8 guides the model to create culturally relevant dialectal Arabic text based on a given persona and context. Figure 9 focuses on summarizing detailed persona descriptions concisely. Finally, Figure 10 instructs an LLM to generate practical dialectal Arabic question-answer pairs in JSON format from provided text.

```
Translate the following text from English to Egyptian Arabic. Ensure that all words
    are in Egyptian Arabic, and do not use any Modern Standard Arabic (MSA). Keep
    the translation casual, conversational, and reflective of how Egyptians would
    naturally speak in everyday situations. Avoid any formal or classical language
    structures. Translate only the input paragraph and don't add anything else in
    your output.
English: {English_text}
```

Figure 5: The translation prompt used with a teacher model to convert English educational pre-training data to a low-resource target language. The placeholder '{English_text}' represents the input English text.

```
Write the following Moroccan dialectal Arabic text in Moroccan Arabizi. Ensure that
    all words are written in Moroccan Arabizi. Keep the text casual, conversational,
     and reflective of how Moroccans would naturally write in everyday situations
    using Arabizi. Translate only the content keys in the following JSON, and output
     a json of the same format:
{JSON_OBJECT}
```

Figure 6: The prompt used with our teacher LLM to convert dialectal Arabic text written in Arabic script into Arabizi. The placeholder {JSON_OBJECT} represents the input text formatted as a JSON object.

```
Translate the following text from English to Moroccan Arabic. Ensure that all words
    are in Moroccan Arabic, and do not use any Modern Standard Arabic (MSA). Keep
    the translation casual, conversational, and reflective of how Moroccans would
    naturally speak in everyday situations. Avoid any formal or classical language
    structures. Translate only the content keys in the following JSON, and output a
    json of the same format:
{JSON_OBJECT}
```

Figure 7: The translation prompt used with a teacher model to convert SmolTalk and TULU instructions data to a low-resource target language. The placeholder '{JSON_OBJECT}' represents the input text.

```
Act as the following person: {persona_description} Act like you are {person_Name}
    and write a {text_genre} in Egyptian dialect, using colloquial Arabic script as
    spoken in Egypt and not Modern Standard Arabic (MSA). Use this context and use
    the information provided in it while writing the {text_genre}:
{context}
Make sure to follow these conditions:
1. Rely on the provided context when writing the {text_genre}.
2. Ensure that the written {text_genre} reflects the cultural background, values,
    and worldview of {person_Name}.
3. Don't write the persona's description. I want you to focus only on the provided
    context when writing while reflecting the perosna's background.
Note: Ensure that all words are in Egyptian Arabic, and do not use any Modern
    Standard Arabic (MSA). Keep the translation casual, conversational, and
    reflective of how Egyptians would naturally speak in everyday situations.
```

Figure 8: Prompt for generating culturally and values-aware text genres in low-resource languages, given a local persona description and a local cultural concept. The placeholders {persona_description}, {text_genre}, and {context} represent the persona description, the intended text genre to generate, and the cultural concept text, respectively.

```
I have the following persona description, I want you to write it in a concise manner
    keeping all the information, the output should be plain text, make sure to
    include all values, morals, and culture of the persona:
{PERSONA_DESCRIPTION}
```

Figure 9: Prompt for generating concise persona descriptions with a LLM utilizing a comprehensive description of a specific persona extracted from the WVS. The placeholder '{PERSONA_DESCRIPTION}' represents the input persona description.

```
You are an expert in Moroccan culture and language, with native-level fluency in Moroccan Arabic dialect
    written in Arabic script. Your task is to transform detailed articles or information (like recipes,
    historical accounts, cultural traditions, etc.) into practical, useful instruction-response pairs in
    authentic Moroccan Arabic dialect expressed in Arabic script.

## Task Description:
1. I will provide you with raw text content (like cooking recipes, cultural traditions, historical
    information, etc.)
2. Transform this content into practical instruction-response pairs where:
   - Instructions ask how to do something, how something works, or how to understand something
   - Responses provide thorough, practical explanations in Moroccan dialect
3. Focus on making these interactions useful for someone wanting to learn practical information
4. Output the results in structured JSON format

## Instruction Format Guidelines:
- For procedural content (recipes, crafts, etc.): Create "how to" questions
  - Example: {example_1}
- For historical/factual content: Create "why" questions
  - Example: {example_2}
- For cultural practices: Create "what is" or "how do we celebrate" questions
  - Example: {example_3}

## Response Format Guidelines:
- Make responses detailed and practical
- Include specific steps for procedural content
- Use authentic Moroccan vocabulary and expressions
- Structure responses in a logical order
- For recipes or procedures, list steps clearly in the order they should be performed
- Include tips and warnings where appropriate
- Keep the tone conversational and helpful, as if explaining to a friend

## Language Guidelines:
- Use authentic Moroccan Arabic dialect written in Arabic script (not transliteration)
- Include typical Moroccan cooking/cultural terminology and expressions
- Use language as it would naturally be spoken, not literary Arabic
- Include common Moroccan filler words and expressions where natural
- For cooking terms or specialized vocabulary, use the actual terms Moroccans use.

## Output Format:
The output should be valid JSON with the following structure:
```json
{
  "instruction_response_pairs": [
    {
      "instruction": "[practical instruction in Moroccan dialect]",
      "response": "[detailed practical response in Moroccan dialect]"
    },
    {
      "instruction": "[another practical instruction]",
      "response": "[another detailed response]"
    }
  ]
}
```
Instructions:
1. Generate at least 3-5 practical instruction-response pairs based on the content I provide.
2. Make each instruction focus on a practical aspect (how-to, ingredients, common mistakes, variations, etc
    .).
3. Ensure responses are detailed enough to be practically useful.
4. The resulting JSON must be properly formatted and valid.
5. For specialized terms or techniques, use the authentic Moroccan terms.
6. When dealing with recipes or procedures, make sure the steps are in logical order.
7. If the article contains cultural context or historical information, include some of this in the responses
    to provide depth.
8. If the article does not have any relevent information to the Moroccan culture then return an empty list.

Now, I'll provide you with content, and you'll transform it into practical instruction-response pairs in
    Moroccan dialect as specified above.
{Text}
```

Figure 10: Prompt for generating synthetic cultural dialectal question-answer pairs using a teacher model given raw text describing a local cultural heritage concept. The placeholder {Text} represents the raw text that is used to generate question-answer pairs from it.

F Samples of texts generated from NileChat

Moroccan conversation (based on Wikipedia context)	
Text genre	**Conversation**
Persona	Jilali, a 40-year-old Moroccan male, lives in Karia Ba Mohamed, Fes-Meknes region, in an urban area. He is married with 3 children and lives in a household of 7, including his parents and in-laws. Jilali has post-secondary non-tertiary education and works full-time as a skilled worker in a government or public institution. He is in the third income decile, considers himself upper middle class, and manages to save money. Jilali values family, friends, leisure, politics, work, and religion highly. He is Muslim, attends religious services weekly, and prays several times a day. He believes most people can be trusted and is quite happy, rating his life satisfaction as 6 out of 10 and his financial satisfaction as 8 out of 10....
Context	طاجين الزيتون (كما يعرف أيضا بالاسم الكامل: طاجين الدجاج بالزيتون) هو طبق مغربي ينحدر من المطبخ الأمازيغي المغربي التقليدي. يعتبر وصفة من بين 400 وصفات و أنواع الطواجين المختلفة الموجودة في المطبخ المغربي. معترف به عالميا كطبق كلاسيكي من المطبخ المغربي باعتباره، مع طاجين البرقوق، جوهرة فن الطهو المغربي. كما يعتبر واحد من أشهر الطواجين المغربية في المغرب و خارج البلد. يعتبر الزيتون والدجاج من المكونات الأساسية لدى طاجين الزيتون، كما يمكن تحضيره بزيادة الليمون المصير، الفطريات، الجزر أو التين. أصل و ايتيمولوجيا. تعود أصول الطاجين إلى المغرب حيث ينحدر من الحضارة الأمازيغية المغربية الأصلية. ذكر إخترعه في المغرب ما قبل العصور الوسطى حيث كان معروف في المغرب في القرن الثامن و كان يعتبر طبق ملكي لدى قبائل أمازيغ صنهاجة المغاربة في القرن الثالث عشر. تحظى بشعبية كبيرة خارج حدود المغرب، معترف به في المطاعم الأوروبية والأمريكية. كما عرف طاجين الدجاج بالزيتون المغربي إنتشارا واسعا في الجزائر، خصوصا بمدينة تلمسان. و يفسر هذا الإقبال بانجذاب عشاق الطاجين المغربي إلى جغرافيا المدينة القريبة من المغرب. طقوس مرتبطة. الشلال. يتم تقديم هذا الطبق بعد غسيل الأيادي على الطريقة المغربية الأصلية باستخدام ما يسمى بالشلال المغربي الذي يعتبر من طقوس الضيافة المغربية و هو عبارة عن إناء أو شلال متحرك يمنح للضيوف إمكانية غسل الأيادي بدون التحرك من مكانهم، إحتراما للضيف حيث يقدم الشلال مباشرة إلى أيادي الضيف ليتمكن بغسيل أياديه قبل الأكل. يتم تناول طاجين الدجاج بالزيتون ساخنًا و يعتبر طبق يومي لدى المغاربة لكن يتم تقدمه في حفلات الاستقبال الكبيرة و الأعراس بصورة ملحوظة كرمز للضيافة. المرشة. تعود المرشة واحدة من أهم التقاليد المغربية العريقة حيث يتم رش ماء الزهر أو المسك في أكف الضيوف كنوع من الترحيب.
Linguistic cues	Include these Moroccan dialect words in your generated text: "دعم" (meaning "sustain"), "تحضير" (meaning "preparation"), "مارس" (meaning "march"), "تغذية" (meaning "nutrition"), and "قلب" (meaning "flip").
Generated text	- السلام عليكم، كيف داير لبارح؟ - لاباس الحمد لله، اليوم جيت نكلمك على واحد الطبق مغربي تقليدي كيعجبني بزاف، سميتو طاجين الزيتون. هاد الطبق كيعتبر من أشهر الأطباق المغربية، وكيتكون أساسا من الدجاج والزيتون، ولكن كاينين بعض الإضافات اللي كتزيده نكهة مميزة، بحال الليمون المصير، والفطريات، والجزر، وحتى التين. - واااه، كيعجبني طاجين الزيتون بزاف، كيكون بنين ومذاقو مميز. كنعرف بلي كيتاكل فالمناسبات الكبيرة والعرس، ولكن كنفضل ناكلو فالدار مع العائلة، كيكون لذيذ ودافئ. - بالضبط، طاجين الزيتون كيكون ديما حاضر فالمناسبات الكبيرة، وكيعتبر من رموز الضيافة المغربية. كيعجبني كيفاش كيقدم مع الشلال، حيث كنغسلو يدينا بطريقة تقليدية قبل ما ناكلو، وكيكون احترام كبير للضيف. - آه، الشلال! هاد الشي كيعجبني بزاف، كيكون عندنا فالدار ديما، كنستعملوه فكل المناسبات. كنغسلو يدينا قبل ما ناكلو، وكيكون عندنا المرشة ديال ماء الزهر أو المسك، كنرحبو بالضياف بيه. - هادشي كيعجبني فالمغرب، التقاليد ديالنا والطقوس اللي كترافق الأكل. طاجين الزيتون كيتطلب تحضير خاص، خاصو الوقت وداكشي، ولكن كيجي بنين ومداقو كيعجب الكل. - كنت سمعت بلي أصل هاد الطبق أمازيغي، من الحضارة الأمازيغية المغربية الأصلية. كيعجبني كيفاش كنلقاو هاد التنوع الثقافي فالطبخ ديالنا، كيجمع بين التقليدي والحديث. - بالضبط، كنلقاو هاد الطبق حتى فالمطاعم الأوروبية والأمريكية، حيث كيعرفوه بزاف. كيعجبني كيفاش كنلقاو دعم لثقافتنا فالعالم. - وفاش كتحضرو، كتعطي ليه لمسة خاصة، كتزيد ليه الخضرة والتوابل، كيجي بنين بزاف. كنت جربت مرة ندير ليه التين، قلب ليه المذاق، ولا كيعجبني بزاف. - هادشي اللي كيعجبني فالطبخ، كنلقاو فيه الإبداع والتجديد. كنلقاو كل واحد كيدير ليه لمسة خاصة، كيجي بنين ومتنوع. - وفاش كيجي مارس، كنلقاو الزيتون الجديد، كيكون بنين بزاف، كنستعملوه فالطبخ، وكيعطي نكهة خاصة. - آه، الزيتون ديال مارس كيكون بنين بزاف، كنستعملوه فالطبخ، وحتى فالسلطات، كيعطي تغذية وصحة. - بالضبط، كنلقاو فوائد كثيرة فالزيتون، كنستعملوه فالطبخ، وكيعطي مذاق مميز، وكيعتبر من المكونات الأساسية فالطبخ المغربي.
Egyptian Blog Post (based on Wikipedia context)	
Text genre	**Blog Post**
Persona	Hatem, a 29-year-old Egyptian male, lives in a rural area of Markaz Al Delengat, Behaira region. He is single, has no children, and lives with his parents in a household of three. Hatem has no formal education and works part-time as an unskilled laborer in private business or industry. He is in the third income decile and considers himself lower class. He has spent savings and borrowed money, and rates his financial satisfaction as 2 out of 10 and his overall life satisfaction as 3 out of 10. Despite this, he reports good health and rates his control over life as 5 out of 10. Hatem is a devout Muslim who attends religious services more than once a week and prays several times a day. He values family and friends highly but does not prioritize leisure or politics. Work and religion are very important to him. He is cautious with people and not very happy.....
Context	الكشري من أشهر الأكلات الشعبية المصرية وتتكون من المعكرونة والأرز والعدس الأسود والبصل المقلي وصلصة الطماطم أي البندورة الحارة والتقلية، ويعتبر الكشري أكلة منخفضة الكلفة عالية السعرات الحرارية وذات زيوت نباتية متأكسدة بسبب القلي الذي يثير حول الوجبة كمًا من الانتقادات الصحية ونحوها. تطوره. في مصر. تطور الكشري في الآونة الأخيرة في مصر حيث زادت عدد المحلات المتخصصة في عمله، كذلك تقوم هذه المحلات بتقديم مأكولات أخرى بجانب الكشري مثل أطباق المعكرونة والأرز باللبن والعاشوراء والفطائر (الهريسة). ويختلف الكشري الإسكندراني عن الكشري العادي في باقي المحافظات حيث يتم وضع العدس الأصفر بدلاً من العدس الأسود ليختلف شكله عن الكشري العادي ويصبح لونه أصفر ولا يضاف معه الصلصة. تحضير الصلصة. مقاديرها. 1- خمس فصوص ثوم. 2- عدد 2 حبة طماطم. 3- علبة صلصة كبيرة. 4- ملعقتين خل أو ثلاث ملاعق 5- ملعقة كبيرة سمن. الكشري في بلاد الشام. الكشري في الشام يسمى المجدّرة يطبخ ويقدم بطريقتين بالبرغل أو بالأرز، والطريقة الأولى وهي البرغل المخلوط بالعدس الأسود المسلوق ويقدم حاراً أو بارداً وخصوصاً عندما يكون الأرز والعدس مهدى على زيت الزيتون بدلاً من السمن البلدي فعندها يفضل تناوله ساخناً. من المعتاد تناول الكشري مع اللبن الرائب أو تناوله بدون ذلك. الكشري في العراق. الكشري في العراق يطبخ بالأرز والعدس الأصفر المجروش ويقدم عادة مع كبة الحامض ويعتبر مكملاً لها. المعلومات الغذائية. تحتوي وجبة الكشري المصري (300 غ تقريباً)، بحسب موقع شهية، على المعلومات الغذائية التالية: بعض الانتقادات. إذا وضع الأرز المعاد تسخينه في الكشري يخلق نوعاً من البكتيريا قد تؤدي لمشاكل صحية تؤدي للوفاة ومن أعراض التسمم من الأرز المعاد تسخينه: لذا من المفترض أن يطلب الكشري بدون الأرز.
Linguistic cues	N/A
Generated text	يا جماعة، النهاردة هتكلم معاكوا عن أكلة شعبية مصرية مشهورة جداً وهي الكشري. الكشري ده بقى أكلة حكاية، عبارة عن خليط من المكرونة والأرز والعدس الأسود والبصل المقلي وصلصة الطماطم الحارة، وبنتحط فوقهم التقلية. أكلةالكشري دي مش بس رخيصة، لأ وكمان مليانة سعرات حرارية وزيوت نباتية متأكسدة بسبب القلي، وده اللي خلى ناس كتير تتكلم عن أضرارها الصحية. الكشري بقى له شعبية كبيرة في مصر، وفيه محلات كتير متخصصة في عمله، وبقت كمان تقدم أكلات تانية جنب الكشري زي المكرونة باللبن والعاشوراء والهريسة. بس عارفين، في الإسكندرية عندهم طريقة مختلفة شوية، بيستخدموا العدس الأصفر بدل الأسود، فبيبقى شكله مختلف ولونه أصفر، وبيقدموه من غير صلصة. نيجي بقى لطريقة عمل الصلصة بتاعة الكشري، هقولكوا على المقادير: هتحتاجوا ثوم وطماطم وصلصة كبيرة وخل وسمنة. وبالمناسبة، الكشري في بلاد الشام اسمه المجدرة، وبيطبخوه بطريقتين، يا إما بالبرغل يا إما بالأرز. واللي بيحبوا البرغل بيعملوه مع العدس الأسود المسلوق وبيقدموه ساخن أو بارد، خصوصاً لو حطوا عليه زيت زيتون بدل السمنة. وفي العراق، الكشري بيتحضر بالأرز والعدس الأصفر المجروش، وبيقدموه مع كبة الحامض. بس خدوا بالكم، في بعض التحذيرات من أكل الكشري اللي فيه أرز معاد تسخينه، لأن ده ممكن يسبب نوع من البكتيريا اللي ممكن تأذي الصحة وتسبب مشاكل خطيرة. عموماً، الكشري أكلة شعبية لذيذة، بس زي ما قلتلكوا، لازم ناخد بالنا من الأرز المعاد تسخينه، والأفضل نطلبه من غير أرز عشان نتجنب أي مشاكل صحية. بالهنا والشفا!
Moroccan conversation (based on local news context)	
Text genre	**Conversation**
Persona	Zohra, a 49-year-old Moroccan woman, lives in a rural area of Tigrigra within the Fes-Meknes region. She is married with two children and lives in a household of five people, not including her parents. Zohra has no formal education and works as a housewife and farm laborer in private business or industry. She is in the sixth income decile and considers herself lower middle class, just getting by financially. Family is very important to Zohra, while friends are rather important. She does not prioritize leisure, politics, or work, and considers religion very important. She is a practicing Muslim, attending religious services once a week and praying several times a day. She believes most people can be trusted and rates her life satisfaction as 6 out of 10 and her financial satisfaction as 7 out of 10. She reports fair health and feels complete control over her life.....
Context	الجمعة 5 أكتوبر 2018, 11:00 تشهد منطقة طماريس بدار بوعزة بالبيضاء إنجاز عدد من المشاريع الهيكلية، من بينها مشروع تنجزه ليدك لتزويد المنطقة بشبكة للتطهير السائل لمياه الأمطار والمياه العادمة. ويشمل هذا المشروع إنجاز سرداب على طول 1102 مترا وقناة لتصريف المياه العادمة طولها 1,2 كلم، إضافة إلى قناة كبيرة لتصريف مياه الأمطار طولها 984 مترا، وكذا قناة مجمعة لتصريف المياه العادمة طولها 1383 متر، بالإضافة إلى وحدة لضخ المياه. ويأتي هذا المشروع، الذي تنجزه ليدك، بشراكة وتنسيق مع عمالة إقليم النواصر، والسلطات المحلية وجماعة دار بوعزة، في إطار مقاربة تهدف إلى مواكبة حركية تعمير منطقة طماريس بالجهة الجنوبية للدار البيضاء من خلال تجهيزها بشبكة للتطهير السائل وشبكة الماء الشروب. وإلى جانب هذا المشروع، الذي ينجز باستثمار يتجاوز 58 مليون درهم، تقوم ليدك بإنجاز مشروع شبكة للإمداد بالماء الشروب بكلفة مالية قدرها أكثر من 7 ملايين درهم. ويتكون المشروع من قناة طولها 3,32 كلم و قطرها 315 ميلمتر. ويندرج المشروعان في إطار مجهودات والتزامات ليدك الرامية إلى مواكبة التنمية الاجتماعية والاقتصادية والتوسع الحضري لمدينة الدار البيضاء.
Linguistic cues	3. Include these Moroccan dialect words in your generated text: "شاك" (meaning "doubt"), "مخبي" (meaning "hidden"), "ريح" (meaning "windy"), "رجل" (meaning "feet"), and "بخيل" (meaning "stingy").
Generated text	- السلام عليكم، كيف دايرين اليوم؟ كاين شي جديد فالبلاد؟ - وعليكم السلام والرحمة، كاين شي جديد فطماريس، كاينة مشاريع جديدة كيديرو فدار بوعزة، كاين شي مشروع ديال التطهير السائل ديال مياه الأمطار والمياه العادمة. - أه، هادشي زوين، وعلاش هاد المشروع بالضبط؟ - كاينة شي مشاريع جديدة ديال التعمير فالمنطقة، وبغاو يديرو شبكة ديال التطهير باش يدوزو مياه الأمطار والمياه العادمة، حيت المنطقة كتعاني شوية من هاد المشكل. - آه فهمت، يعني كاينة شي مشاكل ديال الصرف الصحي فطماريس؟ - بالضبط، كاين شي مشاكل مخبيين، وكاين شي مناطق كيتجمع فيها مياه الأمطار وكيتسربو للمياه العادمة، وكاينة شي مشاكل ديال التلوث. - مساكن، خاصهم يديرو شي حل، حيت هادشي كيخلي المنطقة ريح وبلا بنية تحتية مزيانة. - بالضبط، هاد المشروع غادي يدير سرداب طويل وقنوات كبيرة باش يجمعو مياه الأمطار والمياه العادمة ويدوزوهم بطريقة منظمة. - وعلاش هاد المشروع بالضبط؟ شاك فهادشي؟ - لا، هاد المشروع منظم من طرف ليدك بشراكة مع العمالة والسلطات المحلية، بغاو يديرو شي حل شامل للمشكل ديال التطهير السائل، وحتى الماء الشروب. - أه، زوين هادشي، خاصنا نعاونو فالتنمية ديال البلاد، حيت هاد المشاريع مهمة بزاف. - بالضبط، هاد المشروع غادي يكلف بزاف، ولكن غادي يكون عندو تأثير إيجابي على المنطقة، وغادي يحسن من جودة الحياة ديال الساكنة. - الله يجيب اللي فيه الخير، حيت هادشي كيعاون على التنمية ديال البلاد، ماشي بحال شي ناس بخيل كيفكرو غير فراسهم. - ههههه، كاينين شي ناس كيفما قلتي، ولكن كاينين شي ناس كيخدمو بجد باش يحسنو من الوضعية ديال البلاد. - أه، الله يعاونهم، حيت هادشي ماشي ساهل، خاصهم يخدمو رجلهم باش يطورو البلاد. - بالضبط، حنا كمغاربة خاصنا نتحدو ونخدمو يد فيد باش نتقدمو ونحسنو من الوضعية ديال البلاد. - أه، كلامك صحيح، خاصنا نكونو يد وحدة ضد التخلف والجهل، ونشجعو على التنمية والتقدم. - بالضبط، خاصنا نكونو فخورين ببلادنا ونخدمو عليها، حيت هادشي واجب علينا. - أه، الله يعاون الجميع، ونشوفو بلادنا مزيانة ونقية من كل مشاكل. - إن شاء الله، والله يسهل على كل واحد كيخدم بجد.

Table A.3: Examples of culturally aware and dialectally diverse texts from various genres, generated by our teacher model. The model was provided with input representing a specific persona, local context, and a local linguistic cue following the methodology described in Section 3.1.

	Belebele		Flores		In-house	
	MOR	EGY	ENG→EGY	ENG→MOR	ENG→MOR	ENG→EGY
NileChat-EGY	64.44	70.89	43.85	23.10	11.93	36.93
NileChat-MOR	**70.67**	72.56	39.94	**37.45**	**30.82**	29.98
NileChat	70.33	**72.67**	**44.37**	33.89	28.67	**37.52**

Table B.1: Comparison of the performance of the Egyptian SFT model (*NileChat-EGY*), the Moroccan SFT model (*NileChat-MOR*), and their merged version, *NileChat*, on Belebele (accuracy), Flores (ChrF++), and In-house parallel data (ChrF++).

Dataset	Evaluator	Sample Size	Correctness Score	Dialectness Score
EgyMMLU	Human	100	3.78	3.92
	LLM	100	3.90	4.22
	LLM	400	4.28	4.22
EgyHellaSwag	Human	100	3.90	4.08
	LLM	100	4.22	4.96
	LLM	400	4.11	4.64

Table C.1: Translation quality scores (out of 5) for EgyMMLU and EgyHellaSwag along two dimensions: correctness and dialectness. Human raters and an LLM judge rated the same 100 items; the LLM judge additionally rated 300 more items (400 total) to yield more reliable estimates.

Model Name	Size	Release Date
Less than 7B		
Qwen3-1.7B	1.7	Apr. 2025
ar-stablelm-2-chat	1.6	Jul. 2024
Atlas-Chat-2B	2.6	Sep. 2024
Llama-3.2-3B-Instruct	3.2	Sep. 2024
gemma-3-4b-it	4.3	Mar. 2025
Qwen3-4B	4	Apr. 2025
NLLB-200-3.3B	3.3	Jul. 2022
Qwen2.5-3B-Instruct	3.1	Sep. 2024
More than 7B		
AceGPT-7B-chat	7	Dec. 2023
ALLaM-7B-Instruct	7	Feb. 2025
Qwen2.5-7B-Instruct	7.6	Apr. 2025
Qwen3-8B	8.2	Apr. 2025
Atlas-Chat-9B	9.2	Sep. 2024
gemma-3-12b-it	12.2	Mar. 2025
AceGPT-13B-chat	13	Dec. 2023
jais-13b-chat	13	Aug. 2023

Table C.2: The LLMs used for comparison against NileChat in this evaluation were selected from a list of Arabic-aware models. Each LLM is listed with its corresponding size (in billion parameters) and release date. We utilized the instruct version for all LLMs except for NLLB, which is a machine translation-specific model.

	Model	MMLU		HellaSwag		Belebele	
		EGY	MOR	EGY	MOR	EGY	MOR
Less than 7B	**Qwen3-1.7B**	28.53	28.53	28.07	27.33	22.89	22.89
	ar-stablelm-2-chat	39.54	38.32	34.33	33.40	24.22	22.78
	Atlas-Chat-2B	42.65	45.06	29.62	34.78	54.67	59.00
	Llama-3.2-3B-Instruct	31.10	30.92	28.86	28.39	49.67	40.89
	gemma-3-4b-it	46.32	46.60	34.26	32.53	61.44	52.11
	Qwen3-4B	28.59	28.52	30.21	29.53	22.89	22.89
	Qwen2.5-3B-Instruct	35.71	37.67	31.17	29.62	61.11	44.89
	NileChat (3B)	**58.20**	**58.62**	**38.29**	**40.35**	**78.11**	**73.78**
More than 7B	**AceGPT-7B-chat**	40.76	37.98	33.04	31.04	38.00	33.00
	ALLaM-7B-Instruct	60.18	59.61	40.20	38.14	76.11	66.00
	Qwen2.5-7B-Instruct	57.70	53.51	33.79	32.28	76.67	59.44
	Qwen3-8B	28.53	28.53	31.72	30.95	22.89	22.89
	Atlas-Chat-9B	57.17	60.27	34.75	44.47	78.44	79.33
	gemma-3-12b-it	59.29	56.16	40.16	37.60	80.78	73.11
	AceGPT-13B-chat	46.48	43.65	35.15	33.21	46.33	41.11
	jais-13b-chat	49.33	48.28	38.99	37.45	59.89	53.78

Table D.1: 3-shot performance (accuracy) of models on understanding (MMLU, HellaSwag, and Belebele). Bold values indicate the highest score among models comparable in size to ours (<7B parameters). Underlined values represent the highest score in the entire column, including larger models. Results for zero-shot are presented in Table 1, Section 4.2.

Flores Dataset Performance

	Model	MOR → EGY				MOR → ENG				MOR → FRA				EGY → MOR				EGY → ENG				ENG → EGY				ENG → EGY				FRA → EGY				Average			
		sB0	sB4	cF0	cF4	sB0	sB4	cF0	cF4	sB0	sB4	cF0	cF4	sB0	sB4	cF0	cF4	sB0	sB4	cF0	cF4	sB0	sB4	cF0	cF4	sB0	sB4	cF0	cF4	sB0	sB4	cF0	cF4	sB0	sB4	cF0	cF4
Less than 7B	**Qwen3-1.7B**	19.49	19.80	36.32	36.45	14.76	15.45	40.21	40.79	12.17	12.90	35.08	35.83	19.38	19.31	36.23	36.22	19.65	20.48	46.49	47.05	7.32	7.77	22.02	23.42	10.02	10.62	25.98	27.55	5.98	6.63	19.60	21.69	13.59	14.12	32.74	33.62
	ar-stablelm-2-chat	18.29	18.57	34.31	34.46	7.36	7.03	28.97	28.47	3.51	2.85	19.36	18.08	14.16	16.35	29.25	33.32	8.05	7.93	31.12	30.67	3.25	11.51	8.22	29.25	10.41	18.68	22.89	35.77	3.81	9.55	9.06	26.13	8.60	11.56	22.90	29.52
	Atlas-Chat-2B	19.67	20.03	36.67	36.90	24.64	26.18	49.36	50.32	19.03	20.18	41.53	42.49	19.35	19.88	36.88	**37.65**	23.44	25.41	49.17	50.28	12.06	14.89	29.66	33.48	10.74	15.16	27.45	33.04	8.80	11.72	25.73	29.77	17.22	19.18	37.06	39.24
	Llama-3.2-3B-Instruct	18.75	10.79	35.15	21.71	15.71	0.08	38.75	0.85	12.16	0.54	29.55	3.40	17.67	14.00	34.59	28.64	20.89	0.06	45.19	0.64	5.86	0.91	16.92	5.21	9.74	0.77	24.54	5.29	3.92	1.12	15.14	5.64	13.09	3.53	29.98	8.92
	gemma-3-4b-it	16.89	21.11	33.00	37.78	7.99	24.48	17.81	49.50	5.52	**23.80**	11.28	**46.04**	12.86	16.31	29.10	33.91	12.05	29.04	24.62	54.24	1.88	13.05	4.85	31.03	1.65	20.13	3.24	37.60	0.93	10.56	3.01	28.17	7.47	19.81	15.86	39.78
	Qwen3-4B	20.41	20.57	37.17	37.15	18.98	20.67	44.68	46.07	17.30	18.39	39.96	41.33	16.06	18.86	31.90	35.84	23.99	25.75	50.47	51.83	11.43	11.30	27.93	27.98	15.46	15.83	33.00	33.20	7.44	9.54	21.46	25.90	16.38	17.62	35.82	37.41
	NLLB-200-3.3B	20.92	-	38.74	-	**30.89**	-	**53.64**	-	**27.89**	-	**48.69**	-	17.06	-	35.14	-	**34.62**	-	**58.07**	-	**17.46**	-	**34.89**	-	**26.93**	-	**43.86**	-	11.58	-	28.90	-	**23.42**	-	**42.74**	-
	Qwen2.5-3B-Instruct	18.51	19.71	35.20	36.32	18.61	19.83	44.22	44.99	15.00	16.22	38.56	39.45	16.98	17.90	34.15	34.91	24.07	25.08	50.45	50.99	9.36	8.48	25.17	25.14	11.78	11.43	28.05	28.44	7.47	7.18	22.79	23.32	15.22	15.73	34.83	35.45
	NileChat	**23.81**	**22.68**	**40.25**	**39.25**	28.81	**29.50**	52.98	**53.43**	24.05	23.49	46.30	45.56	**20.90**	**20.51**	**38.22**	37.64	30.58	**31.90**	54.53	**56.19**	15.15	**17.90**	30.76	**36.00**	23.39	**25.03**	40.37	**41.94**	**13.18**	**14.91**	**29.61**	**32.52**	22.49	**23.24**	41.63	**42.82**
More than 7B	**AceGPT-7B-chat**	19.92	18.93	36.91	35.72	17.67	20.04	44.09	45.71	14.78	15.92	38.08	38.87	19.11	17.99	36.04	35.24	23.11	25.81	50.28	51.96	8.93	12.29	21.59	29.72	16.11	18.25	33.32	35.56	5.95	10.28	16.65	27.07	15.70	17.44	34.62	37.48
	ALLaM-7B-Instruct	24.63	25.12	40.93	41.39	26.02	28.13	51.65	52.92	18.91	20.80	42.64	44.04	19.85	19.53	37.32	37.00	29.63	31.82	55.53	56.75	14.63	16.56	32.38	34.79	23.18	25.93	39.85	42.70	13.17	14.09	30.75	32.15	21.25	22.75	41.38	42.72
	Qwen2.5-7B-Instruct	16.28	18.87	32.19	35.08	21.59	23.62	47.28	48.55	18.98	19.96	42.12	42.93	13.24	16.09	29.34	32.91	26.39	27.83	52.72	53.74	9.73	11.19	25.32	28.28	12.54	15.00	27.80	31.34	7.73	9.75	22.71	26.46	15.81	17.79	34.93	37.41
	Qwen3-8B	21.10	21.41	37.84	38.13	22.41	24.33	48.04	49.51	20.46	21.91	43.17	44.32	17.82	18.35	34.63	35.40	27.30	29.36	53.32	54.77	13.60	12.56	31.05	29.63	18.95	18.83	36.79	36.72	10.16	10.27	26.33	26.86	18.98	19.63	38.90	39.42
	Atlas-Chat-9B	19.44	20.52	36.32	37.29	31.29	32.75	54.19	55.61	28.14	28.82	49.35	49.85	18.86	20.36	36.78	37.91	30.98	32.83	55.04	56.07	17.29	18.91	35.83	37.42	16.95	19.69	34.35	37.25	14.53	15.59	32.76	33.92	22.19	23.68	41.83	43.16
	gemma-3-12b-it	19.67	24.48	36.28	41.24	23.68	30.29	47.75	54.47	15.28	29.99	26.64	51.36	10.73	17.30	24.98	36.28	27.38	32.71	53.46	57.42	2.51	14.60	6.56	34.00	6.34	24.27	11.17	41.69	1.45	12.47	3.70	31.30	13.38	23.26	26.32	43.47
	AceGPT-13B-chat	20.14	19.58	37.04	36.57	20.82	24.30	46.12	49.11	18.58	20.52	40.45	42.77	19.54	17.99	36.62	35.50	26.08	29.14	52.52	54.71	13.02	13.42	29.23	31.52	18.81	20.18	36.04	37.93	9.51	11.72	23.19	29.39	18.31	19.61	37.65	39.69
	jais-13b-chat	13.50	20.28	26.36	37.46	19.73	27.40	37.81	51.80	18.14	18.40	39.16	40.95	8.90	13.42	22.22	31.25	22.63	30.07	41.73	54.56	2.46	8.55	6.08	25.73	4.09	18.77	9.56	36.94	1.50	7.26	3.25	23.70	11.37	18.02	23.27	37.80

In-House Dataset Performance

	Model	EGY → ENG				EGY → MSA				ENG → EGY				ENG → MOR				MOR → EGY				MOR → MSA				MSA → EGY				MSA → MOR				Average			
		sB0	sB4	cF0	cF4	sB0	sB4	cF0	cF4	sB0	sB4	cF0	cF4	sB0	sB4	cF0	cF4	sB0	sB4	cF0	cF4	sB0	sB4	cF0	cF4	sB0	sB4	cF0	cF4	sB0	sB4	cF0	cF4	sB0	sB4	cF0	cF4
Less than 7B	**Qwen3-1.7B**	12.14	13.14	31.46	31.98	19.11	20.32	34.24	35.09	4.63	4.96	15.10	15.43	2.52	2.64	11.14	10.99	6.19	6.12	22.58	23.01	6.45	6.04	19.25	18.12	18.20	18.26	34.20	33.86	6.20	6.07	18.38	17.94	9.43	9.69	23.29	23.30
	ar-stablelm-2-chat	7.89	6.74	21.92	20.28	14.56	13.71	27.71	24.23	3.57	10.99	7.50	22.25	0.65	3.69	1.45	14.61	6.95	6.15	20.63	19.05	8.50	7.28	21.32	18.59	14.89	14.28	28.14	27.20	5.19	3.49	15.82	13.12	7.77	8.29	18.06	19.92
	Atlas-Chat-2B	13.52	16.21	32.04	34.25	15.51	18.19	30.39	32.34	3.67	5.15	15.29	17.15	6.81	8.40	21.76	23.79	14.96	15.85	34.83	35.82	12.12	12.76	28.38	28.48	7.05	11.91	20.72	26.75	8.86	10.11	24.35	25.20	10.31	12.32	25.97	27.97
	Llama-3.2-3B-Instruct	12.97	0.38	28.40	1.37	14.24	4.80	25.90	8.81	5.60	2.41	16.44	8.16	2.03	0.98	10.14	5.63	4.74	0.23	17.03	1.01	5.00	1.98	14.55	6.79	15.73	7.97	31.80	18.17	4.30	2.46	14.68	11.25	8.08	2.65	19.87	7.65
	gemma-3-4b-it	16.24	22.91	32.96	43.72	17.53	26.98	30.61	42.52	0.52	13.66	1.61	28.41	0.49	5.67	1.90	18.59	6.10	12.65	16.81	33.20	4.41	12.67	11.22	28.30	5.50	19.30	11.35	35.56	0.72	6.16	2.13	18.92	6.44	15.00	13.57	31.15
	Qwen3-4B	19.21	18.42	39.11	38.95	22.22	23.63	37.84	38.93	8.19	8.16	20.04	20.48	3.41	3.40	13.35	13.57	9.18	8.54	26.74	26.42	7.86	8.51	21.12	21.64	17.98	19.72	33.83	35.71	5.47	6.15	16.43	18.51	11.69	12.07	26.06	26.78
	NLLB-200-3.3B	21.48	-	41.14	-	16.33	-	34.36	-	14.20	-	29.78	-	6.28	-	19.86	-	14.12	-	32.44	-	8.75	-	26.11	-	19.34	-	36.68	-	8.70	-	23.10	-	13.65	-	30.43	-
	Qwen2.5-3B-Instruct	19.44	19.38	39.84	39.59	19.04	20.48	34.07	34.75	6.34	6.58	17.94	18.97	3.38	3.65	13.01	14.36	8.15	9.49	26.02	27.75	7.52	7.68	20.87	20.82	13.47	16.82	28.91	33.21	5.00	5.40	17.04	17.88	10.29	11.19	24.71	25.92
	NileChat	**26.24**	**27.49**	**46.46**	**48.10**	**26.77**	**31.65**	**42.52**	**46.40**	**17.67**	**18.92**	**33.37**	**34.87**	**10.74**	**11.72**	**25.07**	**26.84**	**19.22**	**20.47**	**39.83**	**41.46**	**17.55**	**19.28**	**34.23**	**36.55**	**26.37**	**29.54**	**42.90**	**46.68**	**13.94**	**15.41**	**29.37**	**30.75**	**19.81**	**21.81**	**36.72**	**38.95**
More than 7B	**AceGPT-7B-chat**	19.27	21.44	39.36	40.93	20.94	21.99	35.99	37.51	9.24	10.57	22.39	25.25	3.36	3.69	12.76	15.12	7.80	9.95	26.08	27.97	7.15	8.23	20.20	22.17	20.22	20.47	37.14	37.05	6.53	6.00	19.31	18.98	11.81	12.79	26.65	28.12
	ALLaM-7B-Instruct	26.31	28.98	48.32	49.57	32.49	32.85	48.22	47.93	16.69	17.78	32.85	34.12	8.05	9.32	22.65	24.47	18.61	19.97	39.44	40.36	18.41	19.60	34.86	36.02	23.28	26.15	41.09	43.79	10.28	11.13	25.95	26.49	19.26	20.72	36.67	37.84
	Qwen2.5-7B-Instruct	22.70	21.40	43.05	42.34	19.14	22.90	33.08	36.71	7.65	7.91	19.04	20.16	3.65	4.50	13.87	16.04	9.87	10.55	28.38	29.17	7.72	9.05	20.84	22.79	13.21	16.00	28.10	31.91	4.55	5.74	15.67	18.44	11.06	12.26	25.25	27.20
	Qwen3-8B	22.21	23.62	42.36	44.34	26.06	27.35	41.28	42.38	9.04	9.52	22.25	23.18	3.67	4.50	14.64	15.68	10.23	11.03	28.90	29.63	8.30	9.16	21.79	23.18	17.71	19.49	33.99	35.90	5.79	6.12	17.83	18.83	12.88	13.85	27.88	29.14
	Atlas-Chat-9B	18.98	21.19	38.91	41.38	15.72	20.68	31.29	35.84	4.53	7.74	17.12	21.46	7.19	10.60	22.68	26.51	16.77	19.19	37.84	40.15	13.70	15.38	29.86	31.29	6.19	12.05	19.20	27.01	8.16	12.53	23.70	28.12	11.41	14.92	27.58	31.47
	gemma-3-12b-it	23.12	28.51	46.31	48.58	25.91	32.60	40.06	48.67	6.59	17.70	14.15	34.10	1.83	7.95	6.29	23.15	12.49	18.35	33.40	39.90	12.26	17.54	23.93	34.51	9.14	26.48	18.35	43.51	3.06	10.73	8.83	26.07	11.80	19.98	23.91	37.31
	AceGPT-13B-chat	22.36	25.37	42.53	45.47	24.67	27.42	40.60	42.11	10.76	12.56	25.54	27.49	4.27	5.17	15.26	17.90	10.56	12.24	28.12	31.28	8.47	10.30	22.28	24.78	20.32	22.83	37.61	39.75	6.86	7.31	20.34	20.72	13.53	15.40	29.04	31.19
	jais-13b-chat	16.71	24.94	33.16	43.30	21.68	28.15	36.23	43.52	6.45	14.19	14.78	30.35	2.47	6.35	7.42	20.10	12.06	18.56	26.34	38.44	12.89	18.74	28.36	34.79	15.20	22.74	30.81	39.83	5.57	8.83	16.16	23.75	11.63	17.81	24.16	34.26

Table D.2: Detailed translation evaluation results on Flores and In-House datasets. Metrics: spBLEU (sB) and ChrF++ (cF) for 0-shot (0) and 4-shot (4) scenarios. Bold values highlight the top score among models with fewer than 7 billion parameters. Underlined values indicate the highest score overall in each column.

Table F.1: Sample responses from *NileChat* to prompts in Egyptian and Moroccan dialects, covering general and local cultural knowledge. Samples with green background color represent samples with correct responses, samples with red background color represent samples with not accurate answers.