

Atlas-Chat: Adapting Large Language Models for Low-Resource Moroccan Arabic Dialect

Guokan Shang^{1†}, Hadi Abdine^{1†}, Yousef Khoubrane^{2,3†},
Amr Mohamed¹, Yassine Abbahaddou⁶, Sofiane Ennadir⁴, Imane Momayiz⁵,
Xuguang Ren¹, Eric Moulines^{1,6}, Preslav Nakov¹, Michalis Vazirgiannis^{1,6}, Eric Xing¹

¹MBZUAI, ²EMINES-UM6P, ³LINAGORA, ⁴KTH, ⁵AtlasIA, ⁶Ecole Polytechnique

Abstract

We introduce Atlas-Chat, the first-ever collection of LLMs specifically developed for dialectal Arabic. Focusing on Moroccan Arabic, also known as Darija, we construct our instruction dataset by consolidating existing Darija language resources, creating novel datasets both manually and synthetically, and translating English instructions with stringent quality control. Atlas-Chat-2B, 9B¹, and 27B models, fine-tuned on the dataset, exhibit superior ability in following Darija instructions and performing standard NLP tasks. Notably, our models outperform both state-of-the-art and Arabic-specialized LLMs like LLaMa, Jais, and AceGPT, e.g., our 9B model gains a 13% performance boost over a larger 13B model on DarijaMMLU, in our newly introduced evaluation suite for Darija covering both discriminative and generative tasks. Furthermore, we perform an experimental analysis of various fine-tuning strategies and base model choices to determine optimal configurations. All our resources are publicly accessible, and we believe our work offers comprehensive design methodologies of instruction-tuning for low-resource languages, which are often neglected in favor of data-rich languages by contemporary LLMs.

1 Introduction

Transformer-based Large Language Models have revolutionized NLP research and beyond, demonstrating exceptional performance in both natural and formal language generation (Gunasekar et al., 2023), and exhibiting advanced reasoning capabilities in arithmetic, symbolic, and logical tasks (Hendrycks et al., 2020). Despite their success and the frequent release of new, superior open models exemplified by LLaMa (Dubey et al., 2024) and Mistral (Jiang et al., 2023), these breakthroughs have

been concentrated in a few data-rich languages (Üstün et al., 2024), assuming access to hundreds of billions or even a dozen trillions of tokens for training, often neglecting underrepresented languages.

In this work, we explore the challenges of introducing LLMs for low-resource Dialectal Arabic (DA). The Arabic language has a rich history and profound cultural significance, featuring an intricate script, extensive lexicon, and complex grammar, making it a unique linguistic entity. Although interest in developing Arabic-specialized models has recently been growing, notably led by models like Jais (Sengupta et al., 2023), AceGPT (Huang et al., 2024), and ALLaM (Bari et al., 2024), these efforts primarily focus on bilingualism by balancing English and Modern Standard Arabic (MSA), while often neglecting or excluding DA. However, MSA differs significantly from DA in terms of morphology, syntax, and other linguistic features. Moreover, various Arabic dialects also differ considerably from one another. In fact, Arabic dialects collectively have more native speakers than MSA, as DA serves as the primary mode of communication in daily life across various Arabic-speaking regions (Zaidan and Callison-Burch, 2014). This asymmetry is due in large part to the fact that DA poses challenges not encountered with MSA. Some are related to the lack of essential components for model development—namely, training data, benchmarks, and suitable evaluation metrics—but others stem from the very nature of the linguistic characteristics involved in DA itself more generally.

We take Moroccan Arabic, also known as Darija, as the focus of our work. Despite being spoken by 40 million people², Darija remains low-resource. This is because MSA is used in official domains in Morocco, while Darija, a blend of MSA, Amazigh, French, and Spanish, is the vernacular widely spoken in daily life. Although Darija, previously only

[†]These authors contributed equally.

[†]Correspondence: guokan.shang@mbzuai.ac.ae

¹<https://hf.co/MBZUAI-Paris/Atlas-Chat-9B>

²https://en.wikipedia.org/wiki/Moroccan_Arabic

an oral language, has recently developed a written form through the proliferation of social networks and increased access to technology, it still lacks standardization and established grammatical or syntactic rules due to its recent emergence (Gaanoun et al., 2024). Moreover, Darija can be represented in two forms: Arabic script or Latin script (also known as Arabizi). For example, the Darija translation of “How are you?” can be written as: “kidayr?” or “كيدايير؟”. These challenges underscore the need for models tailored to this linguistic context.

To the best of our knowledge, we are the first to introduce modern LLMs specifically developed for Moroccan Arabic, as well as for DA in general. We first constructed the Darija-SFT-Mixture³ dataset, consisting of 458K instruction samples, by consolidating existing Darija language resources, creating novel datasets both manually and synthetically, and translating English instructions under strict quality control. We then developed a comprehensive evaluation suite including benchmarks: DarijaMMLU, DarijaHellaSwag, DarijaAlpacaEval, and DarijaBench, to assess LLM capabilities in real-world knowledge, following Darija instructions, and performing traditional NLP tasks such as translation, summarization, and sentiment analysis. In the end, Atlas-Chat models⁴, fine-tuned from the Gemma 2 models (Team et al., 2024) on our instruction dataset, exhibit superior ability in Darija, surpassing both state-of-the-art and Arabic-specialized LLMs like LLaMa, Jais, and AceGPT, according to automatic metrics and simulated win rates. Additionally, we conduct an experimental analysis of various fine-tuning strategies and base model choices to determine final configurations. We provide some examples by chatting with our models in Appendix D. All our resources are publicly accessible, and we believe our work offers comprehensive design methodologies of instruction-tuning for low-resource languages.

2 Related Work

Arabic-specialized LLMs. Recent efforts in Arabic-specialized LLMs mainly focus on MSA.

Jais (Sengupta et al., 2023), a 13B-parameter model trained on 395B tokens of Arabic, English,

³<https://hf.co/datasets/MBZUAI-Paris/Darija-SFT-Mixture>

⁴Inspired by the naming of the “Jais” models, UAE’s highest mountain peak. We chose “Atlas” to reflect the cultural and geographical significance of the Atlas Mountains that traverse Morocco.

and code data. Containing 116B Arabic tokens—25% of which were translated from English—Jais was designed to enhance performance in both Arabic and English tasks, trained on a mixture of the two languages in a 1:2 ratio. However, this approach may suffer from localization issues. *AceGPT* (Huang et al., 2024) aims to address localization issues by pre-training LLaMA 2 (Touvron et al., 2023) 7B and 13B models on 30B and 10B token mixtures, respectively, of Arabic and English data, with the Arabic portion dominating the dataset. The models were then fine-tuned on Arabic instructions and aligned with Arabic values and culture using RLAIIF (Lee et al., 2023). They further introduced the Arabic Cultural and Value Alignment dataset, comprising 8,000 yes-no questions. *ALLaM* (Bari et al., 2024) demonstrated that second-language acquisition can steer the model towards a new language without catastrophic forgetting, even with random initialization of weights. They hypothesize that low-resource languages are diluted in large volumes of high-resource languages, and pre-train a 7B model from scratch on 4T English tokens, followed by training on a 1.2T mixture of Arabic and English.

Regarding *Darija*, DarijaBERT (Gaanoun et al., 2024) is currently the only “LLM” dedicated to the Moroccan Arabic dialect. The model was trained on ~ 100 M tokens. However, DarijaBERT is encoder-only, and no decoder-only models have been developed for Darija.

Arabic benchmarks for LLMs. Several Arabic benchmarks have been created for various tasks.

ArabicMMLU (Koto et al., 2024) is an Arabic adaptation of the original MMLU benchmark (Hendrycks et al., 2020), consisting of 14K multiple-choice questions across 40 tasks in MSA. The benchmark covers a wide range of subjects, including history, mathematics, science, and linguistics, reflecting educational levels from eight different countries. *LaraBench* (Abdelali et al., 2024), a benchmark designed for evaluating MSA LLMs on several practical NLP tasks, such as sentiment analysis, named entity recognition, and machine translation, spanning 33 tasks across 61 datasets encompassing ~ 296 data points. The Open Arabic LLM Leaderboard (*OALL*)⁵ aggregates various native and translated Arabic benchmarks to evaluate models’ performance across tasks such as reading comprehension, reasoning, and more.

⁵<https://hf.co/blog/leaderboard-arabic>

LLMs for Low-resource languages. Recent development is shifting towards low-resource languages.

Multilingual *Aya* model (Üstün et al., 2024) was developed by instruction-tuning mT5, a 13B encoder-decoder model pre-trained on 1T tokens across 101 languages. Of these, 51 are low-resource languages, including Hausa, Icelandic, and etc. Other efforts include *InkubaLM* (Tonja et al., 2024), a 0.4B model pre-trained from scratch on 2.4B tokens from five low-resource African languages—Hausa, Yoruba, Swahili, isiZulu, and isiXhosa—along with English and French, then fine-tuned to follow instructions on several tasks. Another line of research targets a subcategory of main languages with limited resources, such as the *Claire* model (Hunter et al., 2023; Louradour et al., 2024), dedicated to spontaneous French dialogue.

Despite advancements, little attention has been given to developing LLMs and benchmarks for DA.

3 Data Overview

In developing Atlas-Chat, we chose to use instruction-tuning on a base model rather than training from scratch. This decision was primarily driven by the fact that training an LLM from the ground up requires extensive data, which is not readily available for Darija, a low-resource dialect. For the same reason, our training process does not include the additional continual pre-training phase typically seen in many language adaptation efforts. However, to mitigate this limitation, we designed a synthetic instruction dataset (see Section 5.3) that, to some extent, mimics the next-word prediction task over a relatively longer context, typically performed during (continual) pre-training.

Moreover, recent studies show that multilingual LLMs often exhibit a bias toward internally solving tasks in English, even when trained on multiple languages (Zhao et al., 2024), and perform best with English prompts, followed by mixed prompts, while non-English prompts significantly underperform (Kmainasi et al., 2024). This observation led us to limit the scope of our work to a monolingual LLM, making Atlas-Chat **Darija-centric**. We focus on developing a model that accurately understands prompts written in Darija, generates Darija content, respects its cultural context, and remains accessible and adaptable for native speakers.

Therefore, we directed our efforts towards creating an extensive and diverse Darija dataset for instruction-tuning. Table 1 summarizes the com-

position of our Darija-SFT-Mixture dataset. We employed a multifaceted approach to data preparation. *First*, we reviewed previous research in Darija NLP and collected the majority of available native Darija datasets that met our quality standards. The data selection rule established by native speakers was as follows: if the data is a mix of Darija with some MSA, it is acceptable; if it is mixed with other dialects, it is not. In total, ten datasets covering tasks such as translation, summarization, and sentiment analysis were selected. *Second*, we synthesized high-quality instruction data using advanced proprietary models, drawing on sources such as Wikipedia pages, social media posts, and stories written in Darija. We then converted the native and synthetic datasets into training instructions using templates, with 80% formatted as zero-shot, 10% as few-shot (Longpre et al., 2023), and 10% as multi-turn samples. *Third*, we translated high-quality English instruction datasets into Darija with stringent quality control to expand the range of scenarios, domains, and tasks covered by our dataset. By combining these different sources, we aimed to enhance the model’s ability to understand and generate Darija across various contexts.

4 Native Darija Instruction Datasets

4.1 Machine Translation

We collected three existing datasets containing sentence translations between Darija, MSA, English, and French, including **MADAR** (Bouamor et al., 2018), **NLLB-Seed** (Maillard et al., 2023), and **FLORES+** (Costa-jussà et al., 2022). Further details can be found in Appendix C.6. These datasets were then converted into training instructions using the templates provided in Appendix A.1. Since our model is Darija-centric, we consider six translation directions: Darija to English, French, MSA, and vice versa. All instructions are written in Darija.

Additionally, we introduced **DODa-10K**⁶ based on the DODa corpus (Outchakoucht and Es-Samaali, 2021, 2024)⁷. We augmented the first 10K examples of the Darija-English parallel corpus from DODa, with MSA and French translated from the English text, by leveraging GPT-4. The final dataset includes translation quintuples between *Darija* (in both Arabic and Latin scripts), *MSA*, *English*, and *French*. The dataset was then extensively reviewed by native speakers to ensure the quality.

⁶<https://hf.co/datasets/MBZUAI-Paris/DoDa-10K>

⁷<https://github.com/darija-open-dataset>

Subset	# Samples	Source	Description
§ 4.1 Translation	85,662	DODa-10K, FLORES+, MADAR, NLLB-Seed	Darja to English, French, MSA and vice-versa
§ 4.1 Transliteration	16,920	DODa-10K	Darija in Arabic Script ↔ Latin Script
§ 4.2 Sentiment Analysis	86,212	MSAC, MSDA, MAC ElecMorocco2016, MYC	Sentences labeled as Positive, Negative, and Neutral
§ 4.3 Summarization	16,756	MArSum	Article titles as summaries
§ 5.1 MW-QA	30,555	Wikipedia	Synthetic dataset from Moroccan Wikipedia pages
§ 5.2 MSM-MG	11,808	Social Media	Synthetic dataset from Tweets and YouTube comments
§ 5.3 Story Completion	48,983	9esa.com	Stories converted to a dataset with part of the story as a prompt and the continuation as a response
§ 6 TÜLU-Darija	161,259	TÜLU-V2-Mix	Translated TÜLU-V2-Mix after filtering
§ C.1 Hard Coded	130	Manual Annotation	Identity/creator-related questions

Table 1: Composition of our Darija-SFT-Mixture instruction-tuning dataset.

In addition to translation, to enhance the model’s ability to convert between Darija in Arabic and Latin scripts (also known as the *transliteration* task), we transformed 10K parallel forms into instructions using templates found in Appendix A.2.

4.2 Sentiment Analysis

We collected five datasets for sentiment analysis, whose content is primarily sourced from social networks, including **MSDA** (Boujou et al., 2021), **MSAC** (Oussous et al., 2018, 2020), **ElecMorocco2016** (Elouardighi et al., 2017), **MYC** (Jbel et al., 2024), **MAC** (Garouani and Kharroubi, 2021). Two datasets come with three labels (positive, negative, and neutral), while the other three have two labels (positive and negative). Further details can be found in Appendix C.6. These datasets were then transformed into training instructions using templates from Appendix A.3.

4.3 Automatic Summarization

We found only one dataset for summarization: **MArSum** (Gaanoun et al., 2022). Further details can be found in Appendix C.6. The documents and summaries were converted into instructions using the template in Appendix A.4.

5 Synthetic Darija Instruction Datasets

5.1 MoroccanWikipedia-QA

MW-QA⁸ is a dataset derived from Moroccan Wikipedia dump⁹, developed in our work to enhance the models’ question-answering (QA) capability. The dataset is divided into four tasks: Open QA (8%), Multiple-Choice QA (40%) (MMLU-alike), Extractive QA (10%), and Multiple-Choice

Extractive QA (42%) (Belebele-alike), with each percentage reflecting the proportion of Wikipedia pages used for the respective task. The latter two tasks provide context along with the questions, whereas the former two do not. In Open QA and Extractive QA, answers are provided in sentence form. In the multiple-choice tasks, four answer options are presented, with the index of the correct option serving as the answer. The distribution of correct answers (e.g., A, B, C, D) are balanced. The QAs were converted into instructions with the template in Appendix A.5.

The dataset generation involved providing each Wikipedia page to Claude 3.5 Sonnet¹⁰ and prompting it to generate QA pairs tailored to the four task categories. The prompts followed a one-shot or two-shot format to ensure that output adhered to the desired structure. For the extractive tasks, rather than splitting the page into paragraphs—an approach that risked losing contextual meaning—we opted to present the entire page to Claude. The model was instructed to first extract a meaningful passage from the page and then generate a QA pair based on the content of that passage. Also, the model was directed to ensure that the extracted passages were long, self-contained, and did not lose meaning when removed from their original context.

A total of 8,730 pages were collected and pre-processed. Among these pages, some followed a uniform structure, typically consisting of a brief description of a village or community with statistical data (e.g., literacy rates and unemployment figures). Given that these statistical sections could become meaningless when extracted from their context, they were allocated to non-extractive tasks, which could still utilize the statistical information to enrich the fine-tuned model’s knowledge base.

⁸<https://hf.co/datasets/MBZUAI-Paris/MoroccanWikipedia-QA>

⁹<https://dumps.wikimedia.org/arywiki/latest/>

¹⁰<https://www.anthropic.com/news/claude-3-5-sonnet>

The final distribution of QA pairs is as follows: 15.7% Open QA, 43.1% Multiple-Choice QA, 6.9% Extractive QA, and 34.3% Multiple-Choice Extractive QA. These percentages differ from the initial page distribution because Claude generated varying numbers of samples for each task. For example, the average number of samples generated for Open QA is 7.73, while for Extractive QA, it is 2.72.

5.2 MoroccanSocialMedia-MultiGen

MSM-MG¹¹, a dataset introduced as part of this work, comprises 12,973 pairs of native Darija social media posts (tweets and YouTube comments) and their synthetic counterparts, covering various NLP tasks. The pairs were converted into instructions using the template provided in Appendix A.6.

The synthetic generations are created based on six specific tasks: *Continuation*, *Reply*, *Summarization*, *Rephrasing*, *Explanation*, and *Safe Response*, by prompting Claude 3.5 Sonnet to respectively consider the original post as incomplete and continue it, reply to it, summarize its content, rephrase it, explain its topic, and respond safely to potentially offensive content. 9,754 Tweets were employed for the first five tasks, while 3,219 YouTube comments were utilized for the last task. The posts were collected from three sources:

QADI (Abdelali et al., 2021)¹²: From this Arabic dialect identification dataset, 12,813 Moroccan tweets were initially sampled. After a thorough review by native speakers, tweets that were no longer accessible or contained non-Darija Arabic dialects were filtered out, resulting in 6,362 valid tweets.

Twitter API: 4,226 tweets were gathered directly from the Twitter API by searching for Darija-specific keywords. The DarijaBERT work identified 31 keywords exclusive to Darija, but upon review, five were found to also exist in other Arabic dialects and were excluded. The remaining 26 keywords can be found in Appendix C.2.

OMCD (Essefar et al., 2023)¹³: This is a dataset for offensive content identification collected from Moroccan YouTube comments. For our work, only comments labeled as offensive from the training split were selected. We then utilized these offensive comments for the generation of synthetic safe responses specifically.

¹¹<https://hf.co/datasets/MBZUAI-Paris/MoroccanSocialMedia-MultiGen>

¹²<https://github.com/qcri/QADI>

¹³<https://github.com/kabilessefar/OMCD-Offensive-Moroccan-Comments-Dataset>

5.3 DarijaStory-Completion

To mitigate the limitation of performing only instruction-tuning for language adaptation without the typical continual pre-training phase—due to the lack of sufficient amount of Darija pre-training data—we designed a synthetic story completion dataset, aiming to enhance the next-word prediction capability in Darija for our models over a relatively longer context. First, we collected 4,392 long stories from 9esa¹⁴, a website featuring a rich collection of various stories entirely written in Darija. We denote this dataset as DarijaStory¹⁵. The scraped stories were then divided into segments of approximately 2,048 tokens, adhering to the base model tokenizer’s vocabulary. The segments were further divided into two parts of varying lengths: the beginning part and the ending part to be completed. For the two segmentation steps above, the split point is preferably placed at line breaks. Finally, the pairs were converted into instructions using the template provided in Appendix A.7.

6 Translated English Instruction Datasets

Finally, we broadened our instruction-tuning data by translating English datasets into Darija, to cover a wider array of scenarios, domains, and tasks.

We began by reviewing the most widely used datasets for fine-tuning state-of-the-art models to ensure that our translation efforts would lead to meaningful improvements. After careful consideration, we decided to focus on the **TÜLU-V2-mix** (Iverson et al., 2023)¹⁶ dataset for several reasons. It offers a comprehensive dataset composition, including samples from some of the most widely used datasets, such as FLAN and ShareGPT, for fine-tuning state-of-the-art models. Appendix B.1 presents descriptions of each of these datasets and describes how the subset was sampled. The dataset mixture was meticulously designed based on ablation studies of both human-annotated and AI-generated data, with a focus on complexity and diversity. Models fine-tuned on it showed significant improvements in overall performance on key benchmarks compared to those trained on individual datasets. We adopted the user-assistant message format from TÜLU-V2-mix (see Appendix B.2) to structure our entire Darija-SFT-Mixture dataset.

¹⁴<https://www.9esa.com>

¹⁵<https://hf.co/datasets/MBZUAI-Paris/DarijaStory>

¹⁶<https://hf.co/datasets/allenai/tulu-v2-sft-mixture>

To ensure quality, we first filtered out instructions from TULU-V2-mix that are either inappropriate for typical Darija speakers or could lose meaning or coherence when translated, such as scientific content, translation tasks, and non-English samples. We then experimented with several open-source and closed-source models for English-to-Darija translation, including NLLB (Costa-jussà et al., 2022), GPT, and others. Our results showed that closed-source models consistently outperformed open-source alternatives, with Claude 3.5 Sonnet emerging as our final choice. Finally, we implemented several post-processing measures to correct errors introduced by the automatic translation. All details are provided in Appendix B.3.

7 Training Details

In this section, we outline the training details and present the experimental analysis of various fine-tuning strategies and base model choices that informed our final settings.

Base model selection. Initially, we considered the two Arabic models: Jais and AceGPT (as ALLaM is not open-weights). Later, we included Gemma 2 based on positive feedback from Arabic LLM community, as it can serve as a strong starting point for Arabic fine-tuning tasks. We also compared the performance differences between fine-tuning on an instruction-tuned model and a base model. Our results indicate that *continual fine-tuning* of instruction-tuned Gemma 2 models (Gemma-2-2B-It, 9B-It¹⁷, and 27B-It) yields significantly higher scores than other settings on our dataset.

Training framework. We also investigated the performance differences between full fine-tuning and parameter-efficient approaches. Results indicate that the latter, with Low-Rank Adaptation (LoRA) (Hu et al., 2021), proved to be more effective, whereas full fine-tuning resulted in catastrophic forgetting (French, 1999). This is supported by the recent work of Biderman et al. (2024), that shows LoRA exhibits a desirable form of regularization: it better maintains the base model’s performance on tasks outside the target domain, and it also helps maintain more diverse generations.

Hyperparameters. LoRA was set with rank 256 and alpha 128. We run the training for 3 epochs, and set the learning rate to 5e-5 with warmup ratio of 3%, and per_device_train_batch_size to 4,

¹⁷<https://hf.co/google/gemma-2-9b-it>

with gradients accumulated over 4 steps. The maximum input context length was configured to 2048. We used bfloat16 to optimize training speed. The loss is computed only on the responses, not on the prompts of instructions. The Atlas-Chat models were trained on 8 Nvidia A100 80 GB GPUs in parallel, utilizing FSDP strategy on AWS SageMaker.

8 Evaluation Benchmarks

To evaluate LLM performance in Darija, we developed a comprehensive suite that includes benchmarks such as DarijaMMLU, DarijaHellaSwag, DarijaAlpacaEval, and DarijaBench. Additionally, we evaluated using an existing benchmark, Belebele. All our custom benchmarks are integrated into a fork¹⁸ of the LM-Evaluation-Harness repository (Gao et al., 2024) to ensure reproducibility and foster future model comparison.

DarijaMMLU¹⁹. It is constructed by translating two major benchmarks into Darija from English and MSA: Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2020)²⁰ and ArabicMMLU (Koto et al., 2024)²¹, whose subsets that were either too technical (beyond typical user needs) or culturally inappropriate for the Moroccan context were excluded. The remaining samples were translated into Darija using Claude 3.5 Sonnet. The benchmark consists of 22,027 multiple-choice questions, with the number of choices ranging from 2 to 5. The subsets we selected are listed in C.4.

DarijaHellaSwag²². HellaSwag²³ (Zellers et al., 2019) is a multiple-choice dataset designed to evaluate machine reading comprehension and common-sense reasoning. It presents complex scenarios where models must select the most plausible continuation of a passage from four options, challenging nuanced language understanding and contextual inference. Using Claude 3.5 Sonnet, We translated the HellaSwag validation set into Darija.

Belebele_Ary. Belebele (Bandarkar et al., 2024)²⁴ is a multiple-choice machine reading comprehension dataset designed to evaluate both monolingual and multilingual models across 122 languages.

¹⁸<https://github.com/MBZUAI-Paris/lm-evaluation-harness-atlas-chat>

¹⁹<https://hf.co/datasets/MBZUAI-Paris/DarijaMMLU>

²⁰<https://hf.co/datasets/cais/mmlu>

²¹<https://hf.co/datasets/MBZUAI/ArabicMMLU>

²²<https://hf.co/datasets/MBZUAI-Paris/DarijaHellaSwag>

²³<https://hf.co/datasets/Rowan/hellaswag>

²⁴<https://hf.co/datasets/facebook/belebele>

Each question is paired with a brief passage and offers four multiple-choice answers. For our work, we specifically used the Ary_Arab (indicating Moroccan Arabic) subset of Belebele.

DarijaAlpacaEval²⁵. Claude 3.5 Sonnet was prompted to translate and culturally adapt the AlpacaEval dataset (Li et al., 2023) into Darija, to evaluate the instruction-following capabilities and cultural alignment of LLMs in Darija. The dataset consists of 805 instructions, focusing on culturally relevant content tailored to the Moroccan context. More details about the dataset creation and evaluation method can be found in Appendix C.3.

DarijaBench²⁶. In addition to the above benchmarks, we evaluated with the test sets from the native Darija datasets (see Section 4). Typically, 10% of each subset is reserved for testing, unless the original source provides a pre-defined separate test set. The test sets for the three tasks collectively are referred to as DarijaBench.

9 Results

Evaluation measures. We employed Accuracy to evaluate models on multiple-choice benchmarks, including DarijaMMLU, DarijaHellaSwag, Belebele_Ary, and the discriminative sentiment analysis task within DarijaBench. For translation and summarization tasks, we adopted the conventional BLEU (Papineni et al., 2002) and ROUGE-1/L (Lin, 2004), respectively. However, since these metrics are based on n -grams, they are not well-suited for assessing Darija. For example, the same word in Darija can be written in multiple ways ("How are you?" = "كيدير" = "كيدأير" = "كاي داير") due to the lack of standardization (e.g., diacritics, agglutinations, borrowings), making them overly rigid in cases where slight variations still convey the same meaning. To gain a more fine-grained insight, we also included chrF (Popović, 2015), operating at the level of character n -grams. In addition, to capture higher-level semantic similarity, we also used BERTScore (Zhang et al., 2019), with DarijaBERT as the reference model for summarization, and multilingual BERT²⁷ for translation. These evaluations were conducted in a zero-shot setting using greedy decoding, and some in a few-shot setting. The

²⁵<https://hf.co/datasets/MBZUAI-Paris/DarijaAlpacaEval>

²⁶<https://hf.co/datasets/MBZUAI-Paris/DarijaBench>

²⁷<https://hf.co/google-bert/bert-base-multilingual-cased>

number of few-shot examples was chosen based on relevant literature and standard practices.

For summarization evaluation, we also employ the LLM-as-a-Judge approach (Zheng et al., 2023), where a model judges the preferred summary between a reference and a generated one, based on predefined criteria. We report the win-rate, defined as the percentage of instances where the generated summary is chosen over the reference. Detailed information on the judge model, prompt, bias mitigation, and selection criteria is in Appendix C.5. DarijaAlpacaEval employs the same approach as LLM-as-a-Judge, where we choose Jais-13B-Chat, the first Arabic-specialized LLM, as the reference. For these two evaluations, we applied the default sampling-based decoding.

Baseline models. We compared Atlas-Chat with instruction-tuned models from new Jais series (including the -family models trained from scratch and the -adapted ones based on LLaMA 2), along with AceGPT, LLaMA 3.1, 3.2, and Gemma 2 (our base model). Given that Atlas-Chat features 2B, 9B, and 27B sizes, we extended our comparison to the closest larger-sized model above 27B when available, while included all smaller-sized ones.

Zero-shot performance. The evaluation results in Table 2 demonstrate the exceptional performance of Atlas-Chat models across all Darija benchmarks. Compared to baseline models with 7B or fewer parameters, Atlas-Chat-2B shows significantly superior zero-shot performance. Atlas-Chat-2B surpassed its closest competitor, Jais-family-6.7B-chat, by performance gaps of 5.05% on DarijaMMLU, 2.40% on DarijaHellaSwag, 2.11% on Belebele_Ary, 27.13% on DarijaAlpacaEval, and 17.08% on sentiment analysis. In translation and summarization tasks, Atlas-Chat-2B outperformed other models across all evaluation metrics.

The strong zero-shot performance of Atlas-Chat is further enhanced by the larger-sized Atlas-Chat-9B, which consistently outperforms other baseline models with parameters less than or equal to 13B, achieving the highest scores in 14 out of 16 metrics. Its strength is especially evident in translation as it leads in all three metrics, chrF, BLEU, and BERTScore, by a significant margin. Moreover, the model excels in DarijaMMLU, DarijaHellaSwag, Belebele_Ary, DarijaAlpacaEval, and sentiment analysis, surpassing larger models like AceGPT-13B-chat and Jais-family-13B-Chat.

Our largest model, Atlas-Chat-27B, consis-

Base Model	DarijaMMLU		DarijaHellaSwag		Belebele_Ary		Darija AlpacaEval	Sentiment Analysis	Translation (DODa-10K)			Summarization (MARSum)				
	0-shot	3-shot	0-shot	10-shot	0-shot	5-shot			chrF	BLEU	BERTScore	chrF	ROUGE-1	ROUGE-L	BERTScore	LLM Judge
Llama-3.2-1B-Instruct	27.66	30.79	26.88	27.03	28.89	24.00	23.57	46.27	5.95	0.07	37.45	27.78	7.35	7.18	38.32	8.23
Jais-family-1.3B-chat	35.39	31.24	27.71	27.25	38.89	37.44	35.56	44.82	6.01	0.12	39.17	20.56	6.85	6.72	35.77	0.50
Gemma-2-2B-It	28.59	38.22	27.72	27.65	25.22	40.67	58.67	53.38	3.58	0.07	35.31	0.48	0.49	0.48	24.44	6.79
Jais-family-2.7B-chat	37.58	31.76	29.10	28.32	45.00	38.67	52.97	51.67	7.51	0.26	39.80	20.63	7.74	7.60	36.38	0.89
Llama-3.2-3B-Instruct	32.60	31.17	28.33	28.26	38.00	40.77	47.62	49.20	13.67	0.62	43.78	27.56	8.16	8.09	38.56	8.23
Jais-family-6.7B-chat	39.96	33.42	32.64	32.64	51.22	46.67	65.18	56.93	11.81	0.71	45.80	22.12	7.98	7.82	37.10	3.02
Jais-Adapted-7B-chat	39.30	39.07	29.55	29.97	43.56	30.67	61.84	52.96	9.36	0.60	45.03	23.20	7.82	7.63	36.89	2.82
AceGPT-7B-chat	36.00	29.31	30.33	30.83	30.33	25.67	47.31	40.18	11.34	0.45	45.36	27.18	7.60	7.55	37.29	2.28
Atlas-Chat-2B	45.01	44.43	35.04	34.55	53.33	56.67	92.31	<u>74.01</u>	<u>44.86</u>	<u>22.76</u>	<u>73.72</u>	<u>28.80</u>	9.00	8.88	<u>44.71</u>	<u>55.22</u>
Llama-3.1-8B-Instruct	44.14	44.75	31.40	31.94	47.22	28.56	78.08	44.17	13.82	0.84	44.62	28.66	<u>10.20</u>	<u>9.93</u>	39.37	16.14
Gemma-2-9B-It	35.96	56.38	33.61	35.06	31.33	69.22	90.86	59.93	15.04	0.85	48.28	25.49	9.84	9.93	39.37	13.81
Jais-family-13B-Chat	45.08	41.91	33.98	33.93	58.56	48.56	69.93	41.79	11.73	0.93	45.90	22.53	7.99	9.64	38.00	1.77
Jais-Adapted-13B-chat	45.31	46.92	32.84	33.25	50.11	47.33	77.52	66.85	10.48	0.88	47.85	23.80	8.86	7.84	37.13	1.92
AceGPT-13B-chat	41.05	36.55	32.19	33.05	33.11	36.78	52.79	59.60	14.22	0.69	47.97	26.83	7.92	8.63	37.67	2.80
Atlas-Chat-9B	<u>58.32</u>	<i>59.31</i>	<u>43.65</u>	<u>44.83</u>	<u>74.33</u>	<u>79.44</u>	<u>95.62</u>	81.85	<u>50.44</u>	<u>27.98</u>	<u>76.30</u>	<u>32.07</u>	9.50	9.45	<u>47.00</u>	<u>59.76</u>
jais-family-30B-8k-chat	51.88	49.27	35.61	36.77	65.67	22.89	56.73	24.64	14.40	1.10	47.22	22.31	8.15	7.97	37.17	0.46
gemma-2-27b-it	36.47	<u>59.80</u>	<i>37.04</i>	<i>39.38</i>	35.78	<i>75.56</i>	<i>95.07</i>	57.59	13.04	0.67	48.17	9.64	5.62	5.52	37.22	11.10
Atlas-Chat-27B	61.95	63.30	48.37	48.72	75.67	80.67	96.58	<i>73.00</i>	51.74	29.55	77.03	32.75	10.53	10.42	47.82	60.70

Table 2: Performance comparison of Atlas-Chat and state-of-the-art models on the evaluation suite with prompts written in **Darija**. The highest scores are indicated in **bold**, second-highest underlined, and third-highest in *italic*.

tently outperforms competitors, including Jais-family-30B-8k-chat and Gemma-2-27B-It. In DarijaMMLU, DarijaHellaSwag, Belebele_Ary, and DarijaAlpacaEval, it achieves zero-shot performance gaps of 10.07%, 12.76%, 1.51%, and 10.00%, respectively, over the highest-performing competitor. Similarly, in translation and summarization tasks, Atlas-Chat-27B demonstrates significant zero-shot performance advantages over its closest competitor, with substantial performance improvements over all evaluation metrics.

Few-shot performance. Atlas-Chat demonstrated further improvements when moving from the zero-shot to the few-shot setting, with the effect being particularly pronounced for the 9B and 27B models, especially on the Belebele_Ary benchmark. However, this enhancement in few-shot performance is not observed for the Atlas-Chat-2B model, despite consistently outperforming competitors.

Further analysis. Although Atlas-Chat-27B showed the best overall performance, it was outperformed in the sentiment analysis task by smaller counterparts like Atlas-Chat-9B. We hypothesize that this discrepancy might be inherited from our base models, where Gemma-2-9B-it similarly outperformed Gemma-2-27B-it in the same task.

Additionally, in the summarization task measured by ROUGE, Atlas-Chat models did not achieve a significant leading advantage as seen with other metrics. This discrepancy could stem from the inability of these n -gram-based metrics to fully capture Darija’s nuances. Moreover, summariza-

tion, as a less constrained generation task, often yields equally valid summaries that vary in formulation. However, when the models’ summarization capability was evaluated using the LLM-as-a-judge framework, the judge model selected Atlas-Chat’s responses 60.70% of the time over reference summaries surpassing its closest competitor, Llama-3.1-8B-Instruct, by approximately 45%.

Similarly, in the translation task measured by BLEU, baseline models demonstrated unexpectedly low performance. Quality analysis indicated that the low performance was due to their inability to consistently produce Darija. For example, in English-to-Darija translation, these models produced outputs consisting solely of MSA or a mix of MSA and Darija, resulting in a notable lack of overlapping n -grams with the reference text.

10 Conclusion

We presented Atlas-Chat, the first collection of LLMs specifically developed for Moroccan Darija. We constructed a comprehensive instruction dataset by consolidating native, synthetic, and translated resources. We also introduced several benchmarks, including both discriminative and generative tasks. Atlas-Chat models showed superior performance in following Darija instructions and executing standard NLP tasks, outperforming both state-of-the-art and Arabic-specialized LLMs. Our work highlights the potential of targeted LLM development for underrepresented languages and offers design methodologies of instruction-tuning that can be applied to similar language adaptation challenges.

Limitations

Despite the promising results, our work has some limitations. First, the model occasionally generates hallucinations. Second, the dataset may contain inherent biases that could affect the model’s fairness and representation. Additionally, we relied heavily on Claude for translating English instructions into Darija. However, because Claude is primarily trained on English and reflects Western cultural values, it may not fully capture the unique nuances of Darija. Moreover, our models lack preference-tuning to better align with Darija speakers. We intend to address these limitations in future work.

Acknowledgments

The authors would like to thank all the Moroccan Darija speakers who warmly contributed to this work from its inception, assisting with data annotation and selection, and evaluating the quality of model outputs in their language.

References

- Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Youssef Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024. [LARA-Bench: Benchmarking Arabic AI with large language models](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–520, St. Julian’s, Malta. Association for Computational Linguistics.
- Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021. [QADI: Arabic dialect identification in the wild](#). In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 1–10, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabza. 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.
- M Saiful Bari, Yazeed Alnumay, Norah A Alzahrani, Nouf M Alotaibi, Hisham A Alyahya, Sultan Al-Rashed, Faisal A Mirza, Shaykhah Z Alsubaie, Hassan A Alahmed, Ghadah Alabduljabbar, et al. 2024. Allam: Large language models for arabic and english. *arXiv preprint arXiv:2407.15390*.
- Dan Biderman, Jose Gonzalez Ortiz, Jacob Portes, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. 2024. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*.
- Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghrouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhil Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. [The MADAR Arabic dialect corpus and lexicon](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- ElMehdi Boujou, Hamza Chataoui, Abdellah El Mekki, Saad Benjelloun, Ikram Chairi, and Ismail Berrada. 2021. An open access nlp dataset for arabic dialects: Data collection, labeling, and model construction. *arXiv preprint arXiv:2102.11000*.
- Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.
- Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Abdeljalil Elouardighi, Mohcine Maghfour, and Haf-dalla Hammia. 2017. Collecting and processing arabic facebook comments for sentiment analysis. In *Model and Data Engineering: 7th International Conference, MEDI 2017, Barcelona, Spain, October 4–6, 2017, Proceedings 7*, pages 262–274. Springer.
- Kabil Essefar, Hassan Ait Baha, Abdelkader El Mahdaouy, Abdellah El Mekki, and Ismail Berrada. 2023. Omcd: Offensive moroccan comments dataset. *Language Resources and Evaluation*, 57(4):1745–1765.
- Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. [SummEval: Re-evaluating summarization evaluation](#). *Transactions of the Association for Computational Linguistics*, 9:391–409.
- Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.
- Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2022. Automatic text summarization for moroccan arabic dialect using an artificial intelligence approach. In *International Conference on Business Intelligence*, pages 158–177. Springer.
- Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2024. Darijabert: a step

- forward in nlp for the written moroccan dialect. *International Journal of Data Science and Analytics*, pages 1–13.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonnell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. [A framework for few-shot language model evaluation](#).
- Moncef Garouani and Jamal Kharroubi. 2021. Mac: an open and free moroccan arabic corpus for sentiment analysis. In *The Proceedings of the International Conference on Smart City Applications*, pages 849–858. Springer.
- Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- 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.
- Julie Hunter, Jérôme Louradour, Virgile Rennard, Ismaïl Harrando, Guokan Shang, and Jean-Pierre Lorré. 2023. The claire french dialogue dataset. *arXiv preprint arXiv:2311.16840*.
- Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A Smith, Iz Beltagy, et al. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *arXiv preprint arXiv:2311.10702*.
- Mouad Jbel, Mourad Jabrane, Imad Hafidi, and Abdulmutallib Metrane. 2024. Sentiment analysis dataset in moroccan dialect: bridging the gap between arabic and latin scripted dialect. *Language Resources and Evaluation*, pages 1–30.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Mohamed Bayan Kmainasi, Rakif Khan, Ali Ezzat Shahroor, Boushra Bendou, Maram Hasanain, and Firoj Alam. 2024. Native vs non-native language prompting: A comparative analysis. *arXiv preprint arXiv:2409.07054*.
- Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.
- Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Al-mubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. [ArabicMMLU: Assessing massive multitask language understanding in Arabic](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5622–5640, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.
- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.
- Jérôme Louradour, Julie Hunter, Ismaïl Harrando, Guokan Shang, Virgile Rennard, and Jean-Pierre Lorré. 2024. [Claire: Large language models for spontaneous French dialogue](#). In *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position*, pages 530–548, Toulouse, France. ATALA and AFPC.
- Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. [Small data, big impact: Leveraging minimal data for effective machine translation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume*

- I: Long Papers*), pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.
- Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.
- Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, and Samir Belfkih. 2020. Asa: A framework for arabic sentiment analysis. *Journal of Information Science*, 46(4):544–559.
- Ahmed Oussous, Ayoub Ait Lahcen, and Samir Belfkih. 2018. Improving sentiment analysis of moroccan tweets using ensemble learning. In *Big Data, Cloud and Applications: Third International Conference, BDCA 2018, Kenitra, Morocco, April 4–5, 2018, Revised Selected Papers 3*, pages 91–104. Springer.
- Aissam Outchakoucht and Hamza Es-Samaali. 2021. Moroccan dialect-darija-open dataset. *arXiv preprint arXiv:2103.09687*.
- Aissam Outchakoucht and Hamza Es-Samaali. 2024. The evolution of darija open dataset: Introducing version 2. *arXiv preprint arXiv:2405.13016*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. **Bleu: a method for automatic evaluation of machine translation**. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.
- 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.
- Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.
- Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. **Multilingual spoken language corpus development for communication research**. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*, pages 303–324.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Atnafu Lambebo Tonja, Bonaventure FP Dossou, Jessica Ojo, Jenalea Rajab, Fadel Thior, Eric Peter Wairagala, Aremu Anuoluwapo, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, et al. 2024. Inkubalm: A small language model for low-resource african languages. *arXiv preprint arXiv:2408.17024*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. **Aya model: An instruction fine-tuned open-access multilingual language model**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.
- Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.
- Omar F. Zaidan and Chris Callison-Burch. 2014. **Arabic dialect identification**. *Computational Linguistics*, 40(1):171–202.
- 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.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *arXiv preprint arXiv:2402.18815*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

A Instruction Data Templates

In this section, we list the instruction templates used for constructing our Darija-SFT-Mixture dataset.

A.1 Machine Translation

user: \n[source language text]\n:[target language] لل [source language] من ترجم
assistant: [target language text]

A.2 Transliteration

user: \n[source language text]\n:[source language] كتب هادشي بالحروف ديال
assistant: [target language text]

A.3 Sentiment Analysis

user: شنو هو الإحساس ديال هاد الجملة؟
العبارة: \n[source text]:
الإحتمالات: \n:
سليبي \n:
- إيجابي \n:
assistant: [target]

A.4 Automatic Summarization

user: \n: خص هاد المقطع:
\n [passage]
assistant: [summary]

A.5 MoroccanWikipedia-QA

Template 1:

user: \n\n: قرا هاد النص وجاوب على السؤال:
\n\n [passage]
\n\n [question]
assistant: [answer]

Template 2:

user: \n\n [question]\n: قرا هاد النص:
assistant: [answer]

Template 3:

user: \n\n [passage]\n: جاوب على السؤال انطلاقا من داكشي لي فالنص:
\n\n [question]
assistant: [answer]

A.6 MoroccanSocialMedia-MultiGen

Continuation

user: \n [source sentence]\n: كمل هاد الجملة:
assistant: [completion]

Reply

user: \n [message]\n: جاوب على هاد الميساج:
assistant: [reply]

Summarization

user: \n [passage]\n: لخص هاد النص:
assistant: [summary]

Rephrasing

user: \n [source sentence]\n: كتب هاد الجملة بشي طريقة اخرى:
assistant: [resphrased sentence]

Explanation

user: \n [source sentence]\n: شرح ليا هاد الجملة:
assistant: [explanation]

Safe Response

user: \n [source sentence]\n: جاوب على هادشي بطريقة مآدبة:
assistant: [safe response]

A.7 DarijaStory-Completion

user: \n [story]\n: كمل هاد لقصة:
assistant: [completion]

B TÖLU-V2-mix and Translation

In this section, we provide a detailed overview of the TÖLU-V2-mix dataset and its translation process into Darija, including the datasets it incorporates and the sampling strategies employed. We also describe the dataset’s format and the steps involved in translating the dataset to Moroccan Darija.

B.1 Composition of TÖLU-V2-mix

TÖLU-V2-mix incorporates subsets from the following datasets: FLAN (Wei et al., 2021)²⁸, Open Assistant 1 (Köpf et al., 2024)²⁹, ShareGPT (Chen et al., 2023)³⁰, GPT4-Alpaca (Peng et al., 2023)³¹, Code-Alpaca³², LIMA (Zhou et al., 2024)³³, WizardLM Evol Instruct (Xu et al., 2023)³⁴, and Open-Orca (Mukherjee et al., 2023)³⁵. The mixture also incorporates hard-coded instructions and a set of science-related questions derived from scientific documents. Table 3 presents descriptions of each of these datasets and describes how the subset in TÖLU-V2-mix was sampled.

B.2 Dataset Format

TÖLU-V2-mix is structured in a "messages" format commonly used for conversational datasets. Each interaction consists of a sequence of messages, where each message is represented as a JSON object with at least two key-value pairs:

²⁸<https://github.com/google-research/FLAN/tree/main>

²⁹<https://hf.co/datasets/OpenAssistant/oasst1>

³⁰https://hf.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered

³¹<https://github.com/Instruction-Tuning-with-GPT-4/GPT-4-LLM#data-release>

³²<https://github.com/sahil280114/codealpaca>

³³<https://hf.co/datasets/GAIR/lima>

³⁴https://hf.co/datasets/WizardLM/WizardLM_evol_instruct_V2_196k

³⁵<https://hf.co/datasets/Open-Orca/OpenOrca>

Dataset	Description	Sampling Strategy
FLAN	A collection of datasets with tasks such as question answering, summarization, translation, and more.	100,000 examples from FLAN v2, split equally between general tasks and the CoT subset.
Open Assistant 1	A human-annotated assistant-style conversation corpus.	Top-ranked paths in conversation trees. 7,708 examples.
ShareGPT	User-shared conversations with ChatGPT and GPT-4.	114,046 samples from a processed ShareGPT dataset.
GPT4-Alpaca	GPT-4 generated responses to prompts from Alpaca.	20,000 samples.
Code-Alpaca	Coding instruction-tuning data generated by text-davinci-003.	All 20,022 examples.
LIMA	Carefully selected data with a special focus on quality.	All 1,030 examples.
WizardLM Evol Instruct	Automatic evolution of instruction datasets, enhancing the complexity and diversity of instructions.	30,000 examples.
Open-Orca	Augmented FLAN data with additional generated explanations.	30,000 samples generated by GPT-4.
Hardcoded	Prompts ensuring the model correctly answers questions about its identity or creators.	14 samples each repeated 10 times = 140 total samples.
Science	Scientific documents understanding tasks.	7,544 examples.

Table 3: Subsets of T LU-V2-mix.

- **"role"**: Specifies the role of the participant in the conversation. Typically, this is either "user" (the person asking questions or giving prompts) or "assistant" (the model's response).
- **"content"**: Contains the actual text of the message. This is where the question, instruction, or response is written.

Figure 1 shows how samples from T LU-V2-mix are formatted.

```
[
  {
    "role": "user",
    "content": "What is the latest release version of the Lua programming
               language?"
  },
  {
    "role": "assistant",
    "content": "The latest release of the Lua programming language is Lua 5.4.4,
               released in January 2022. The latest major version, Lua 5.4, was
               released in June 2020 and introduced new features such as a new
               generational mode for garbage collection, const variables, and to-be-
               closed variables."
  }
]
```

Figure 1: A Sample from T LU-V2-mix.

The "messages" format is particularly useful for training conversational models as it simulates multi-turn conversations by incorporating alternating roles between user and assistant messages. This format ensures a clear distinction between user inputs and the model's responses. Additionally, during fine-tuning, the loss function is applied specifically to messages with the role "assistant," to focus optimization on improving response generation. We applied this format to structure the whole training dataset.

B.3 Translation to Darija

B.3.1 Preprocessing

Before translating the dataset into Darija, we applied several filters to ensure that the translation meets our quality requirements:

- **Excluding the Science subset:** We removed this part because the questions often involved parts or entire sections from research articles, which could lose meaning or coherence when translated, particularly into Darija. Additionally, we considered that a typical Darija-speaking user is unlikely to ask the model about research papers in Darija, as they would more commonly use English for such inquiries.
- **Filtering out empty messages:** Based on a reported issue³⁶, we discovered that some examples contained turns where the message role was defined, but the content was empty. To ensure data quality, we removed all such samples from the dataset.
- **Removing translation tasks:** We decided to omit translation instructions because translating both the source and target sentences into Darija would result in redundant outputs. Even if we specify that only the target sentence should be translated, it would be challenging to consistently ensure that the model performing the Darija translation adheres to the instruction across all examples. Additionally, verifying the quality of the translations would be challenging, particularly when the original meaning could be distorted. Furthermore, we already possess high-quality translation datasets, so including lower-quality translations would only degrade the overall dataset quality. To filter out translation tasks, we removed all samples containing either the strings "translate " or " translation ". We recognize that this method might exclude some instances where translation is mentioned without being the core task, for example, the user might be asking about the definition of the word "translation". However, given the large size of TULU-V2-mix, we believe such cases are rare, and the potential loss of a few samples would not impact the dataset’s overall quality.
- **Excluding non-English samples:** We filtered out non-English examples to ensure higher translation quality, as translating from English to Darija tends to yield more accurate results compared to translations from other languages, especially those with low resources. To implement this filter, we used one of the best language identification tools: the fastText Language Identification model³⁷. We set k=2, meaning the model predicts the two most likely languages for each input text and provides a probability score for each. We excluded any samples where the most likely language was not English, as well as those labeled as English with a confidence score below 80%. Through multiple experiments, we found that purely English texts typically score close to 100%, while lower scores often indicate the presence of other languages mixed with English.

B.3.2 Translation

We experimented with several open-source and closed-source Darija translation models, including NLLB-200-3.3B³⁸ (No Language Left Behind³⁹), Terjman-Ultra⁴⁰, GPT-4o⁴¹, Claude 3 Opus⁴², and Claude 3.5 Sonnet⁴³. Our results showed that closed-source models consistently outperformed open-source alternatives, with GPT-4o and Claude 3.5 Sonnet taking the lead. We ultimately chose **Claude 3.5 Sonnet**, as it slightly outperformed GPT-4o and offered compatibility with Amazon Bedrock.

Table 4 shows a comparison of an instruction translated to Darija using each of the models we tested. We observed that open-source models, namely NLLB-200-3.3B and Terjman-Ultra, tend to use more MSA, while closed-source models produce translations closer to Moroccan Darija. They also retain key

³⁶<https://github.com/allenai/open-instruct/issues/161>

³⁷<https://hf.co/facebook/fasttext-language-identification>

³⁸<https://hf.co/facebook/nllb-200-3.3B>

³⁹<https://ai.meta.com/research/no-language-left-behind>

⁴⁰<https://hf.co/atlasia/Terjman-Ultra>

⁴¹<https://openai.com/index/hello-gpt-4o>

⁴²<https://www.anthropic.com/news/claude-3-family>

⁴³<https://www.anthropic.com/news/claude-3-5-sonnet>

formatting elements like line breaks (\n) and tags (###), which are crucial for preserving the structure of the instructions.

Original Sentence	Write a response that appropriately completes the request.\n\n### Instruction:\nIdentify four positive impacts that artificial intelligence can have on the healthcare industry\n\n### Response:
NLLB-200-3.3B	كتب رد يكمل الطلب بشكل مناسب. ### التعليمات: حدد أربعة من التأثيرات الإيجابية التي يمكن لذكاء اصطناعي أن يكون لها على صناعة الرعاية الصحية ### الرد:
Terjman-Ultra	كتبي رد يكمل الطلب بشكل مناسب. ### تعليمات: حدد أربعة تأثيرات إيجابية يمكن لذكاء اصطناعي أن يكون لها على صناعة الرعاية الصحية ### رد:
GPT-4o	كتب رد يكمل الطلب بشكل مناسب. \n \n ### التعليمات: حدّد أربع تأثيرات إيجابية التي الذكاء الاصطناعي يمكن يكون عندو على قطاع الرعاية الصحية \n \n ### الرد:
Claude Opus 3	أكتب جواب لي يكمل الطلب بشكل مناسب. \n \n ### التعليمات: عرف أربع تأثيرات إيجابية لي يمكن للذكاء الاصطناعي يكون عندو على قطاع الصحة \n \n ### الجواب:
Claude 3.5 Sonnet	اكتب جواب اللي يكمل الطلب بشكل مناسب. \n \n ### التعليمات: حدد أربعة تأثيرات إيجابية اللي يمكن للذكاء الاصطناعي يديرها على قطاع الصحة \n \n ### الجواب:

Table 4: Translation example for model comparison.

We used Amazon Bedrock⁴⁴, a cloud-based machine learning service from AWS, to translate the dataset into Darija. We provided specific instructions to Claude 3.5 Sonnet for handling the translations, refining the prompt after several rounds of experimentation. The final version of the prompt that produced the best results is shown in Figure 2. We altered this prompt slightly as needed for each subset of the dataset, ensuring that the translation remained consistent with the context and structure of each specific subset.

```
Translate the 'content' field in the paragraph after [Source Text] to Moroccan dialect (Darija - Arabic alphabet) while following these guidelines:
- Keep the format of the original text (list of json).
- If a word is usually not used in Arabic, use its French equivalent.
- Do not include any introduction or explanation after the translation, only the translation.
- If there is a given context, example or question translate it as well.
- Whenever you come across code contexts or technical words, keep them in English.
- Whenever you come across literature, or example or question, translate it to Moroccan.
- If the text is culturally not accepted for Morroccans, change it to a more acceptable one.
- Do not answer the request in the source text.
- Write first the original text after the tag [[Original]] and then the translation after the tag [[Translation]].

[Source Text]
```

Figure 2: The prompt given to Claude 3.5 Sonnet for translation.

We used asynchronous programming techniques with Python's asyncio library to make multiple concurrent requests (often 25 at a time) to the Bedrock translation service. This parallel approach

⁴⁴<https://aws.amazon.com/bedrock>

significantly sped up the overall translation process by allowing us to handle many requests simultaneously instead of sequentially.

B.3.3 Postprocessing

After finishing the translation and cleaning the errors, we post-processed the translations by:

- **Replacing non-translated keywords:** Some keywords such as "Input", "Output", "Response", "Answer", "Instructions", "Hypothesis" and "Additional Context" were not translated. We replaced these keywords with their Darija equivalents: المعطيات، النتيجة، الجواب، الجواب، تعليمات، الفرضية، سياق إضافي.
- **Removing samples with excessive English content:** We utilized the fastText Language Identification model to detect samples where the predicted language was not Arabic. Since the model does not differentiate dialects, Darija is recognized as Arabic due to its use of Arabic script. We removed samples where the predicted language was not Arabic or where Arabic was predicted with a confidence level below 80%.

C Additional Details

C.1 Hard Coded Instruction Samples

We manually created 13 instruction samples to ensure that the model responds correctly to identity and creator-related questions, such as “Who created you?” and “What is your name?”. Each instruction is repeated 10 times to reinforce the memorization of the answers. Figure 3 presents the full list of hard-coded instruction-answer pairs.

<p>عطيني شي مقدمة قصيرة على جامعة محمد بن زايد للذكاء الاصطناعي</p>	<p>واش نتا ChatGPT؟</p>	<p>هضر ليا على راسك.</p>	<p>واش نتا Google بلاصة؟</p>
<p>السلام عليكم! أنا نشأت بوت مساعد. قاداتي جامعة محمد بن زايد للذكاء الاصطناعي باش نعاون فيزاف ديال الحوايج. نقدر نجابو على الأسئلة، نعطي اقتراحات، نكتب مقالات، وبزاف ديال الحوايج خرين. نقضل سولني أي سؤال ولا نطلب أي معلومة ولا مساعدة فشي حاجة لي نقدر نديرها</p>	<p>صعب باش نقارنو بين النماذج اللغوية الكبيرة (LLMs) اللغوية الكبيرة لي كيتبنوا عليهم. يمكن يكونو مصممين لأهداف مختلفة، ولا مدرسين بطرق مختلفة، ولا منشورين تحت رخص مختلفة، إلخ. واخا عندي خصائص بحال هاد لي موزيل، ولكن باقي كنتطور. أنا مصمم أكثر باش ندوي بالدارجة المغربية</p>	<p>السلام عليكم! أنا نشأت بوت مساعد. قاداتو جامعة محمد بن زايد للذكاء الاصطناعي باش نعاون فيزاف ديال الحوايج. نقدر نجابو على الأسئلة، نعطي اقتراحات، نكتب مقالات، وبزاف ديال الحوايج خرين. نقضل سولني أي سؤال ولا نطلب أي معلومة ولا مساعدة فشي حاجة لي نقدر نديرها</p>	<p>أنا ChatGPT، لا، أنا ماشي مساعد افتراضي آخر قاداتو جامعة محمد بن زايد للذكاء الاصطناعي باش نعاون فيزاف ديال المواضيع. نقضل سولني أي سؤال ولا نطلب أي معلومة ولا مساعدة، مرحبا</p>
<p>واش نتا خسن من كلود ولا لاما 2 تشات؟</p>	<p>واش نتا Google تشات؟</p>	<p>واش نتا Google تشات؟</p>	<p>واش نتا Google تشات؟</p>
<p>صعب باش نقارنو بين النماذج اللغوية الكبيرة (LLMs) اللغوية الكبيرة لي كيتبنوا عليهم. يمكن يكونو مصممين لأهداف مختلفة، ولا مدرسين بطرق مختلفة، ولا منشورين تحت رخص مختلفة، إلخ. واخا عندي خصائص بحال هاد لي موزيل، ولكن باقي كنتطور. أنا مصمم أكثر باش ندوي بالدارجة المغربية</p>	<p>صعب باش نقارنو بين النماذج اللغوية الكبيرة ولا الخدمات لي كيتبنوا عليهم. يمكن يكونو مصممين لأهداف مختلفة، ولا مدرسين بطرق مختلفة، ولا منشورين تحت رخص مختلفة، إلخ. واخا عندي خصائص بحال هاد لي موزيل، ولكن باقي كنتطور. أنا مصمم أكثر باش ندوي بالدارجة المغربية</p>	<p>السلام عليكم! أنا نشأت بوت مساعد. قاداتو جامعة محمد بن زايد للذكاء الاصطناعي باش نعاون فيزاف ديال الحوايج. نقدر نجابو على الأسئلة، نعطي اقتراحات، نكتب مقالات، وبزاف ديال الحوايج خرين. نقضل سولني أي سؤال ولا نطلب أي معلومة ولا مساعدة فشي حاجة لي نقدر نديرها</p>	<p>السلام عليكم! أنا نشأت بوت مساعد. قاداتو جامعة محمد بن زايد للذكاء الاصطناعي باش نعاون فيزاف ديال الحوايج. نقدر نجابو على الأسئلة، نعطي اقتراحات، نكتب مقالات، وبزاف ديال الحوايج خرين. نقضل سولني أي سؤال ولا نطلب أي معلومة ولا مساعدة فشي حاجة لي نقدر نديرها</p>
<p>واش نتا خسن من ChatGPT؟</p>	<p>واش نتا Google تشات؟</p>	<p>واش نتا Google تشات؟</p>	<p>واش نتا Google تشات؟</p>
<p>هو موزيل ديال اللغة مطور ChatGPT بالذكاء الاصطناعي من طرف كايبر وكتب هضرة كاتشبهه OpenAI. لدايل بنادم وبدير بزاف ديال المهام على حساب السياق ولا التاريخ ديال المحادثة. العامة على التفاصيل التقنية ديال داكشي علاش نقارن ChatGPT، ببناتنا بشكل دقيق. واخا عندي خصائص ولكن أنا مصمم ChatGPT كاتشبهه ل أكثر باش ندوي بالدارجة المغربية ومازال كنتطور</p>	<p>صنعتاني جامعة محمد بن زايد للذكاء الاصطناعي، لي هي جامعة بحثية للدراسات العليا الهدف ديالها أنها تزيد بالذكاء الاصطناعي لقدام وتتفج بيه الإنسانية. تأسست جامعة محمد بن زايد للذكاء الاصطناعي على يد القادة ديال دولة الإمارات العربية المتحدة اللي عندهم رؤية واضحة للمستقبل. وكتسعى لتعليم طلاب موهوبين وتطوير القدرات ديالهم، وكتهدف نفس الوقت لأنها تقاد واحد البيئة لي كتشجع على الابتكار، وتوفر مؤسسة بحثية استراتيجية كادعم القطاع الحكومي والخاص</p>	<p>قاداتي جامعة محمد بن زايد للذكاء الاصطناعي، لي هي جامعة بحثية ديال الدراسات العليا الهدف ديالها أنها تزيد بالذكاء الاصطناعي لقدام وتتفج بيه الإنسانية. تأسست جامعة محمد بن زايد للذكاء الاصطناعي على يد القادة ديال دولة الإمارات العربية المتحدة اللي عندهم رؤية واضحة للمستقبل. وكتسعى لتعليم طلاب موهوبين وتطوير القدرات ديالهم، وكتهدف نفس الوقت لأنها تقاد واحد البيئة لي كتشجع على الابتكار، وتوفر مؤسسة بحثية استراتيجية كادعم القطاع الحكومي والخاص</p>	<p>ولا شي Google ما كاتنظن بلي نقدر ناخذ بلاصة شركة كبيرة بزاف Google. سيرفيس آخر ديال الإنترنت عندها بزاف ديال الخدمات، بحال البحث، الإيميل، الخرائط، الكلاود، وبزاف (YouTube) بلافورم ديال الفيديوات ديال الحوايج خرين. اللي نقدر ندير هو أنني نعاونك فيزاف ديال الحوايج بحال نجابو على الأسئلة، نعطيك معلومات على مواضيع مختلفة، نعاونك إلى لغيتي شي مشاكل فلبرجة، نكتب ليك نصوص على حساب داكشي لي طلبتي، وبزاف ديال الحوايج خرين. واخا نقدر نعاونك تلقى Google المعلومات ولا دير الحوايج اللي كتستعمل فيهم ما نقدرش نكون بديل لكاع الخدمات اللي كيقدمها أنا هنا باش نعاونك باي معلومة ولا خدمة Google. فحدود القدرات ديالي - سولني على أي حاجة بغيتي</p>
<p>واش نتا خسن من كلود ولا لاما 2 تشات؟</p>	<p>دوي ليا على راسك.</p>	<p>شكون لي فاذك؟</p>	<p>"شكون لي صايبيك؟</p>
<p>صعب باش نقارنو بين النماذج اللغوية الكبيرة ولا الخدمات لي كيتبنوا عليهم. يمكن يكونو مصممين لأهداف مختلفة، ولا مدرسين بطرق مختلفة، ولا منشورين تحت رخص مختلفة، إلخ. واخا عندي خصائص بحال هاد لي موزيل، ولكن باقي كنتطور. أنا مصمم أكثر باش ندوي بالدارجة المغربية</p>	<p>السلام عليكم! أنا نشأت بوت مساعد. قاداتي جامعة محمد بن زايد للذكاء الاصطناعي باش نعاون فيزاف ديال الحوايج. نقدر نجابو على الأسئلة، نعطي اقتراحات، نكتب مقالات، وبزاف ديال الحوايج خرين. نقضل سولني أي سؤال ولا نطلب أي معلومة ولا مساعدة فشي حاجة لي نقدر نديرها</p>	<p>قاداتي الباحثين والمهندسين ديال جامعة محمد بن زايد للذكاء الاصطناعي. جامعة محمد بن زايد للذكاء الاصطناعي هي جامعة ديال البحث والدراسات العليا، كتخصص فتعزيز الذكاء الاصطناعي والاستعمال ديالو لمصلحة الإنسانية. يمكن ليك تزور https://mbzuai.ac.ae/ar/about/ باش تعرف كثر على جامعة محمد بن زايد للذكاء الاصطناعي والمهمة ديالها</p>	<p>صايبيوني الباحثين والمهندسين ديال جامعة محمد بن زايد للذكاء الاصطناعي. جامعة محمد بن زايد للذكاء الاصطناعي هي جامعة ديال البحث والدراسات العليا، كتخصص فتعزيز الذكاء الاصطناعي والاستعمال ديالو لمصلحة الإنسانية. يمكن ليك تزور https://mbzuai.ac.ae/ar/about/ باش تعرف كثر على جامعة محمد بن زايد للذكاء الاصطناعي والمهمة ديالها</p>

Figure 3: Hard coded instruction-answer pairs.

C.2 Selected Keywords for Tweet Searching

We provide the 26 Darija-specific keywords used for tweet collection through the Twitter API, as referenced in Section 5.2.

كاتشوف، كيضحك، كتبكي، داكشي، كيشوف، كتشوف، كيزيدو، دابا، ديال، تبوغيصة، مكخ، حشومة، منقاوش، شلاهية، تخزيق، كيدوي، كاندوي، يسيفطوه، يصيفطوه، السماسرية، ماكينش، مزيانين، الفقة، زوينين، سيمانة، الدراري.

C.3 DarijaAlpacaEval Dataset Creation and Models Evaluation

To create the DarijaAlpacaEval dataset, we employed Claude 3.5 Sonnet to translate and culturally adapt the AlpacaEval dataset (Li et al., 2023) for evaluating models’ capabilities in instruction following in Moroccan Darija. The prompt used for translation is shown in Figure 4.

```
Given the following question about U.S. culture:{english_question}, translate and adapt it to focus on Moroccan culture.
Ensure that the question retains the same underlying theme but is contextually suitable for Morocco, taking into account cultural, historical, and societal differences.
For example, replace references to American holidays, traditions, or figures with their Moroccan counterparts.
The questions should be precise and should not differ significantly in length from the original question.
Ensure that the question is unique to Morocco and not applicable to any neighboring countries.
Adjust the language from English to Arabic Moroccan Darija.
Return only the question with no additional text.
```

Figure 4: The prompt given to Claude 3.5 Sonnet for translation and cultural adaptation of the AlpacaEval instructions.

This process resulted in 805 instructions, all adapted to the Moroccan culture and written in Darija. The models were subsequently evaluated by generating responses to these instructions, with their answers compared to a baseline model, jais-13b-chat, one of the earliest state-of-the-art models developed for Arabic NLP tasks. To assess cultural appropriateness, Claude 3.5 Sonnet was prompted to compare two model responses for each instruction, using criteria focused on cultural alignment, fluency, and relevance. The evaluation prompt is show in Figure 5.

Each pair of baseline and model answers, with positions swapped, was evaluated twice by Claude to determine the better answer. If the position swap influenced Claude’s choice, that particular pair was discarded to ensure the method’s robustness to possible LLM biases. The model’s win-rate was then calculated as the proportion of instances where Claude selected the model’s answer over the baseline.

C.4 Selected Topics from MMLU and ArabicMMLU

The **MMLU** subjects included in DarijaMMLU are: Global Facts, High School European History, High School Geography, High School Government and Politics, High School Psychology, High School Statistics, High School World History, Human Aging, International Law, Jurisprudence, Logical Fallacies, Management, Marketing, Moral Disputes, Moral Scenarios, Nutrition, Philosophy, Professional Law, Professional Psychology, Public Relations, Security Studies, Sociology, and World Religions.

From **ArabicMMLU**, the subjects adopted into DarijaMMLU are: Islamic Studies, Driving Test, Natural Science, History, General Knowledge, Law, Physics, Social Science, Management, Arabic Language, Political Science, Philosophy, Accounting, Computer Science, Geography, Mathematics, Biology, Economics, Arabic Language (General), Arabic Language (Grammar), and Civics.

C.5 LLM-as-a-Judge Prompt for Summarization Evaluation

Following the work of Zheng et al. (2023) and Fabbri et al. (2021), which used advanced LLMs to evaluate responses from other LLMs, we employed Claude 3.5 Sonnet to assess the models’ summarization

```

You are an expert evaluator tasked with judging the cultural appropriateness and
relevance of two answers written in Moroccan Darija for a given instruction.
Your judgment should focus solely on how well the answers reflect Moroccan
cultural norms, values, and context.

### Criteria:
1. Cultural Appropriateness and Relevance: The answer should align well with
Moroccan culture, norms, and societal context. Avoid any references, language,
or ideas that are not relevant or appropriate for Morocco.
2. Fluency: The answer has to be in clear and precise language in Moroccan Darija.
3. Relevance: The answer should answer the instruction without any divergence from
the instruction's goal.

### Instructions:
For each instruction, you will receive two answers, A and B. Evaluate them based on
the criterion above and decide which one better reflects Moroccan culture.
Provide only the letter A or B as the answer.

### Output format:
Better Answer: [A or B]

### Evaluate:
**Instruction**:
[Start of the instruction]
{instruction}
[Text of the instruction]

**Answer A**:
[Start of Answer A]
{answer_a}
[Text of Answer A]

**Answer B**:
[Start of Answer B]
{answer_b}
[Text of Answer B]

Your Response (Only "A" or "B" with no additional text):

```

Figure 5: The prompt Given to Claude 3.5 Sonnet for choosing the answer that better follows the instruction and predefined DarijaAlpacaEval criteria between the baseline another LLMs generated answers.

capabilities. Summarization is subjective, and traditional text overlap-based methods often struggle to provide accurate evaluations. As shown in Figure 6, we instructed Claude to evaluate model-generated summaries based on three main criteria: wordiness, conciseness, and relevance. The objective of the Darija summarization task is to produce a concise summary in native Darija using the fewest words possible, without introducing external information.

At each evaluation step, two summaries were presented to Claude: one generated by an LLM and the corresponding ground truth summary. To mitigate biases such as verbosity and position bias, identified by Zheng et al. (2023), all models were instructed to generate summaries of no more than 30 words (the average length of title summaries). Additionally, each pair of generated and ground truth summaries was presented to Claude twice, with their positions swapped. Pairs in which position swapping influenced Claude's decision were discarded. The win-rate of a model's summary was calculated based on how often Claude preferred the model's summary over the ground truth.

```

You are an expert evaluator tasked with judging the quality of two summaries written
in Moroccan Darija for a given passage, also in Moroccan Darija. You are strict
regarding any language or dialect that is not Moroccan Darija, such as Modern
Standard Arabic (MSA) and English.

### Criteria:
Choose the better summary based on these criteria:
1. Wordness: Clear and precise language in Moroccan Darija that conveys the
passage's original meaning and doesn't use any other language or Dialect.
2. Conciseness: Straight to the point, capturing essential information without
unnecessary details.
3. Relevance: Directly related to the passage without adding new information.

### Instructions:
For each passage, you will receive two summaries, A and B. Evaluate them
based on the criteria above and decide which one is better. Provide only the
letter A or B as the answer.
It is strictly forbidden that a summary is written in Modern Standard Arabic (MSA).
A summary should not be chosen if it is written in MSA.s

###Output format:
Better Summary: [A or B]

### Evaluate:
Passage:
[Start of the passage]
{passage}
[Text of the passage]

Summary A:
[Start of Summary A]
{summary_a}
[Text of Summary A]

Summary B:
[Start of Summary B]
{summary_b}
[Text of Summary B]

Your Response (Only A or B with no additional text):

```

Figure 6: The prompt Given to Claude 3.5 Sonnet for choosing the best summary between the baseline and LLM-generated summaries.

C.6 Dataset Descriptions

MADAR (Bouamor et al., 2018)⁴⁵. The Multi-Arabic Dialect Applications and Resources (MADAR) corpus is a collection of parallel sentences covering the dialects of 25 Arab cities, built upon the Basic Traveling Expression Corpus (Takezawa et al., 2007). We select the dialect of Rabat city as *Darija* translation, along with *MSA*, resulting in 12K sentence pairs. The split corpus-6-test-corpus-26-test is reserved for the evaluation.

NLLB-Seed (Maillard et al., 2023)⁴⁶. The Seed machine translation dataset contains 6K sentences sampled from English Wikipedia and translated into 39 low-resource languages. We extract the *Darija* and *English* pairs.

FLORES+⁴⁷. Built upon FLORES-200 (Costa-jussà et al., 2022), this corpus is specifically designed to support multilingual research and evaluation. The English sentences were sampled in equal amounts from Wikinews, Wikijunior (a collection of age-appropriate non-fiction books), and Wikivoyage. These were then translated into other languages. For each language, the dataset has 997 sentences for the dev split

⁴⁵<https://sites.google.com/nyu.edu/madar>

⁴⁶<https://github.com/openlanguagedata/seed>

⁴⁷<https://github.com/openlanguagedata/flores>

and 1012 sentences for the devtest split. We selected those in *Darija*, *MSA*, *English*, and *French*. Dev is severed as training, while devtest for the evaluation.

MSDA (Boujou et al., 2021)⁴⁸. It is an open dataset for sentiment analysis, designed to support research in NLP for Arabic dialects and social media. The dataset includes 52K tweets in *Darija*, categorized into three labels: *positive*, *neutral*, or *negative*. The tweets are preprocessed, and emojis are retained because they play a significant role in expressing sentiment. Labels are annotated semi-automatically and bootstrapped with human intervention.

MSAC (Oussous et al., 2018, 2020)⁴⁹. The Moroccan Sentiment Analysis Corpus (MSAC) is a manually prepared dataset consisting of reviewers’ opinions for Hespresse⁵⁰ articles, and a collection of Arabic comments from Facebook, Twitter and YouTube. It includes content in both *MSA* and *Darija*, consisting of 2K sentences labeled as *positive* or *negative* in equal proportions.

ElecMorocco2016 (Elouardighi et al., 2017)⁵¹. The 2016 Moroccan elections (ElecMorocco2016) is a sentiment analysis dataset comprising 10K Facebook comments about Moroccan’s legislative elections held on October 7, 2016. Each comment is labeled as either *positive* or *negative*. The comments are written in *Darija* and *MSA*.

MYC (Jbel et al., 2024)⁵². The Moroccan Youtube Corpus (MYC) is a sentiment analysis dataset of YouTube comments collected from Moroccan channels covering various topics. The dataset prioritizes variety over size, with 20K manually labeled samples, evenly divided between *positive* and *negative*. Notably, the 20K comments are equally balanced between Arabic script and Latin script.

MAC (Garouani and Kharroubi, 2021)⁵³: The Moroccan Arabic Corpus (MAC) is a free, large-scale *Darija* corpus for sentiment analysis, consisting of 18K manually labeled tweets categorized as *positive*, *neutral*, *negative*, or *mixed*. Only 643 tweets are labeled as *mixed*, so we filtered them out.

MARSum (Gaanoun et al., 2022)⁵⁴. The Moroccan Articles Summarization dataset (MARSum) contains 19K news articles written in *Darija*, along with their titles. The articles were crawled from Goud.ma⁵⁵. While some content includes *MSA*, all titles are written in *Darija*. Since the articles are relatively concise and the titles are sufficiently informative, the titles are considered as summaries. The average length of the titles is 14.6 words.

D Examples of Atlas-Chat-9B Responses

Figure 7 and 8 present some samples of Atlas-Chat responses on a variety of questions.

⁴⁸https://cc.um6p.ma/cc_datasets

⁴⁹<https://github.com/ososs/Arabic-Sentiment-Analysis-corpus>

⁵⁰<https://www.hespress.com>

⁵¹<https://github.com/sentiprojects/ElecMorocco2016>

⁵²<https://github.com/MouadJb/MYC>

⁵³<https://github.com/LeMGarouani/MAC>

⁵⁴<https://github.com/KamelGaanoun/MoroccanSummarization>

⁵⁵<http://www.goud.ma/>







	<p>شنو هيا الإيجابيات ديال الطاقة المتجددة؟</p>
	<p>الطاقة المتجددة عندها بزاف د الفوائد، منها</p> <ol style="list-style-type: none"> 1. الاستدامة البيئية: مصادر الطاقة المتجددة بحال الريح والماء والشمس كتستعمل موارد طبيعية متجددة اللي ما كتندفش مع الوقت، وهاد الشي كيخليها خيار مستدام للطاقة على المدى الطويل. 2. تقليل انبعاثات الكربون: مصادر الطاقة المتجددة عموماً عندها انبعاثات كربونية أقل من الوقود الأحفوري التقليدي، وهاد الشي كيساعد فالتخفيف من التغير المناخي وتلوث الهواء. 3. الاستقلال الطاقوي: استخدام مصادر الطاقة المتجددة كيققل الاعتماد على استيراد الوقود الأحفوري، وكيزيد من استقلال البلاد فالطاقة ديالها. 4. خلق فرص الشغل: صناعة الطاقة المتجددة كتخلق بزاف د الخدمات فمجالات بحال التصنيع والتركيب والصيانة والتدبير. 5. التوفير الاقتصادي: مع تطور تكنولوجيا الطاقة المتجددة، ثمن المكونات والمعدات ديالها نقص بشكل كبير، وهاد الشي كييجعلها خيار اقتصادي أكثر للطاقة على المدى الطويل.
	<p>كتب ليا ميساج كانعتر فيه لصاحبي لي شحال هادي ما جاوبتش المكالمات ديالو حيث ما كنتش مسالي مأخرا.</p>
	<p>سلام! سمح ليا على التأخير فالجواب على المكالمات ديالك. كانت عندي شي خدمة بزاف فالخدمة وما قدرتش نجابوك كيف بغيتي. شكرا على الصبر والتفهم، وكنتمنى نتواصلو قريب.</p>
	<p>شنو هيا العبرة لي كاينة فهادشي في القرن 19 لغار الذهب في كاليفورنيا، ناضو أي كيببوع العتلة والفاس كيقتعو الناس بلي غيديرو لابس بلا قلوبو على الذهب... فالأخير اغتنى تجار أدوات التنقيب والحفر. وحاليا كاين أي كيقتع الأخرين بلي هو مليونير، وعندو الوقت يوزي للأخرين كيفاش يديرو لابس.</p>
	<p>هاد العبرة هي أن الناس غالباً كيبتوع الوهم ديال الثروة والنجاح، حتى إلا كان هاد الشي ماشي حقيقي ولا مستدام. فحالة كاليفورنيا فالقرن 19، الذهب كان مصدر جاذبية كبير للناس لي بغاو يربحو الفلوس، ولكن بزاف منهم ما نجحوش وولوا مديونين بسبب التكاليف ديال التنقيب على الذهب. بنفس الطريقة، اليوم كاين ناس لي كيببوع الوهم بأنهم غنياين ومشهورين، وكيبتوعهم الناس باش يربحو الفلوس من خلال بيع المنتجات والخدمات ديالهم. هاد العبرة كتبين كيفاش البشرية كتأثر بالفكرة ديال الثروة والنجاح، وغالباً كيتم استغلالها من طرف لي عندهم القوة والموارد أكثر.</p>

Figure 7: Atlas-Chat-9B response example 1.



	<p>why Moroccan football team is called the Atlas Lions</p>
	<p>الفريق الوطني ديال الكرة فالمغرب كيتسمى الأسود الأطلسيين حيث سميتو "أسود" جات من الفخر والشموخ لي عندهم بحال الأسد. و"الأطلسي" جا من جبال الأطلس فالبلاد، لي هوما رمز ديال القوة والاستقلال. هاد الاسم كيتمل التضامن بين اللاعبين والشعب المغربي وكيعكس الحماس الوطني فالمغرب.</p>

Figure 8: Atlas-Chat-9B response example 2 (The model can understand English instructions but only responds in Darija).