CIDAR: Culturally Relevant Instruction Dataset For Arabic

Zaid Alyafeai [1,*] **Khalid Almubarak** [2,*] **Ahmed Ashraf** [3,*] **Deema Alnuhait** [4,*]
Saied Alshahrani [5,6] **Gubran A. Q. Abdulrahman** [1] **Gamil Ahmed** [1,7]
Qais Gawah [1] **Zead Saleh** [1] **Mustafa Ghaleb** [1,8]
Yousef Ali [1] **Maged S. Al-Shaibani** [1]

[1] King Fahd University of Petroleum and Minerals (KFUPM) [2] Prince Sattam bin Abdulaziz University (PSAU)
[3] ARBML [4] University of Illinois Urbana-Champaign [5] Clarkson University [6] University of Bisha
[7] Interdisciplinary Research Center for Smart Mobility and Logistics (IRC-SML), KFUPM
[8] Interdisciplinary Research Center for Intelligent Secure Systems (IRC-ISS), KFUPM

Abstract

Instruction tuning has emerged as a prominent methodology for teaching Large Language Models (LLMs) to follow instructions. However, current instruction datasets predominantly cater to English or are derived from English-dominated LLMs, leading to inherent biases toward Western culture. This bias negatively impacts non-English languages such as Arabic and the unique culture of the Arab region. This paper addresses this limitation by introducing CIDAR, *the first open Arabic instruction-tuning dataset culturally aligned by native Arabic speakers*. CIDAR contains 10,000 instruction and output pairs that represent the Arab region. We discuss the cultural relevance of CIDAR via the analysis and comparison to a few models fine-tuned on other datasets. Our experiments indicate that models fine-tuned on CIDAR achieve better cultural alignment compared to those fine-tuned on 30x more data. The dataset is available on HuggingFace https://huggingface.co/datasets/arbml/CIDAR.

1 Introduction

The need for Natural Language Processing (NLP) applications has exploded in an era of unprecedented linguistic interaction between humans and machines. As these applications strive for greater inclusivity and effectiveness across diverse linguistic landscapes, the need for datasets that reflect the cultural differences and linguistic peculiarities of specific regions becomes increasingly important. In the context of Arabic language understanding, the challenge lies not only in linguistic complexity but also in capturing the rich cultural fabric that shapes communication in the Arab world.

*Equal contribution. Corresponding author: Zaid Alyafeai, email: g201080740@kfupm.edu.sa

أكتب رسالة شكر ~~لخدمة الموظف~~ لموظف لخدمته في الشركة لمدة معينة.
اسم الموظف: ~~جون سميث~~ صالح محمد. مدة الخدمة: 5 سنوات.

عزيزي ~~جون~~ صالح،

أريد أن أعبر عن تقديري الصادق لجهودك وتفانيك خلال الخمس سنوات الماضية. خلال فترة عملك هنا، قدمت العديد من المساهمات التي ساعدت في جعل شركتنا ناجحة. تقديرنا الكبير لموثوقيتك والتزامك وولائك.

أتمنى لك كل التوفيق في مسيرتك المقبلة.

مع خالص التحية،

[اسمك]

Figure 1: An example of our localization procedure in CIDAR of a given (instruction , output) pair. We show, in colors, the grammatical and cultural modifications.

In the past year, many language models have been pre-trained and instruct-tuned for Arabic, like JAIS (Sengupta et al., 2023), and ACEGPT (Huang et al., 2023). All these models have been trained on a large corpus of Arabic text and then fine-tuned to respond to users' instructions via instruction-tuning. However, such efforts do not release high-quality instruction datasets to be openly used for research. Moreover, they use a lot of machine-translated or machine-generated instruction datasets without further human review or audit, disregarding the consequences of using such poor, distorted, and misaligned instructions.

In this paper, we introduce CIDAR, the *first* open instruction-tuning dataset that has gone through extensive review and localization (see Figure 1) crafted for instructional tuning in Arabic. In the next sections, we delve into the dataset creation

Findings of the Association for Computational Linguistics: ACL 2024, pages 12878–12901
August 11-16, 2024 ©2024 Association for Computational Linguistics

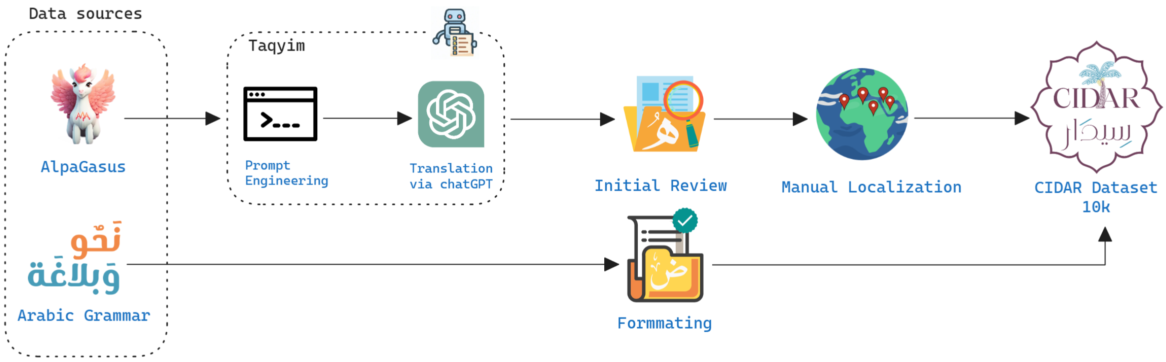

Figure 2: Workflow diagram of CIDAR's data collection pipeline, illustrating each pipeline phase and its components.

process, elucidating the methodology employed to navigate the delicate balance between linguistic accuracy and cultural relevance. The paper discusses the potential applications of CIDAR in enhancing the performance of Arabic LLMs, shedding light on its role in bridging the gap between language understanding and cultural context within the realm of Arabic instruction-tuning. We study the performance of a fine-tuned model on CIDAR and other models fine-tuned on non-localized datasets. Our experiments show the importance of CIDAR in adapting LLMs to the Arabic culture.

We summarize our contributions as follows:

1. We release three open datasets, CIDAR, CIDAR-EVAL-100, and CIDAR-MCQ-100, as a suite for fine-tuning and evaluating Arabic LLMs on cultural relevance.

2. We highlight our data localization approach and showcase the cultural relevance of CIDAR, compared to a translated dataset (ALPAGASUS) via thorough analysis.

3. We show that a model fine-tuned on our dataset, CIDAR, can better capture the Arabic cultural nuances compared to models fine-tuned on translated datasets like ALPAGASUS or much more data like ACEGPT.

2 Issues of Arabic Instruction Datasets

Two main issues currently exist in the literature, as addressed in Section 6, in creating Arabic instruction-tuning datasets: the full translation of both instruction-response pairs using Machine Translation tools (MTs) and the translation of instructions, then generating responses using LLMs like GPT-4 (Achiam et al., 2023). Next, we highlight the drawbacks of such approaches.

2.1 MTs-*related* Issues

One harmful drawback of the current instruction-tuning datasets' creation approaches is the poor, naive, and direct translation of English instruction-output pairs to Arabic without human intervention or supervision using off-the-shelf MTs like Google Translate, which is widely known for their social problems like gender, cultural, and religious biases and stereotypes (Prates et al., 2020; Ullmann and Saunders, 2021; Lopez-Medel, 2021; Chen et al., 2021; Naik et al., 2023; Alshahrani et al., 2022b; Al-Khalifa et al., 2024; Alshahrani et al., 2024). Many researchers have repeatedly stressed how such unguided translations are not only prone to various linguistic and grammatical errors, detrimental outcomes, cultural misalignment (favoring the Western culture), and representational harm to native speakers (unrepresentative content) but also introduce negative performance implications of models trained on them (Stanovsky et al., 2019; Habash et al., 2019; Das, 2020; Agrawal et al., 2023; Alshahrani et al., 2023; Thompson et al., 2024; Roscoe, 2024; Saadany et al., 2024).

2.2 LLMs-*related* Issues

The other hazardous drawback of the current instruction-tuning datasets' creation approaches is the unvetted, unchecked, and unsupervised translation of instruction-response pairs from English to Arabic or the generation of responses for the previously translated instructions, all using LLMs like GPT-3.5 Turbo or GPT-4 without paying attention to the consequences. Many research studies have underscored various risks, threats, and controversies in LLMs, for example, research studies like (Paullada et al., 2021; Wach et al., 2023; Thakur, 2023; Naous et al., 2023; Dong et al., 2023; Acerbi and Stubbersfield, 2023) accentuated that

Table 1: Comparison between translated ALPAGASUS and CIDAR regarding names and countries using Word Clouds. In ALPAGASUS, the top locations are the United States (الولايات) and New York (نيويورك), and the top names are John (جون) and Marry (ماري), while in CIDAR, after our localization, the top locations are Yemen (اليمن) and Egypt (مصر), and the top names are Muhammad (محمد) and Sarah (سارة).

Locations in ALPAGASUS	**Locations in CIDAR**
	
Names in ALPAGASUS	**Names in CIDAR**
	

most commonly used LLMs could exhibit a wide spectrum of biases, privacy, and security hazards, ethical questions, hallucination, and could create a damaging or deceptive content of certain group. Besides, LLMs could generate content (e.g., responses) that suffer cultural misalignment and cultural contradictions, leading to culturally unaligned, undiverse, untruthful, and unrepresentative outputs (Prabhakaran et al., 2022; Alshahrani et al., 2022a; Kasirzadeh and Gabriel, 2023; Cetinic, 2022; Bang et al., 2023; Yu et al., 2023; Masoud et al., 2023; Galileo, 2023; Ji et al., 2024; Mubarak et al., 2024).

3 CIDAR

We introduce CIDAR, a dataset that has 10,000 instruction and output pairs. CIDAR was constructed using two sources. First, we used the ALPAGASUS dataset[1] by (Chen et al., 2023a), which is a high-quality dataset filtered from the Stanford Alpaca dataset (Taori et al., 2023). ALPAGASUS contains more than 9K instruction, input, and output triplets. We translate 9,109 of the data to Arabic using ChatGPT (GPT-3.5 Turbo). Then, we append it with around 891 questions and answers about the Arabic language and Grammar crawled from AskTheTeacher website[2]. Figure 2 highlights the main procedure for our data collection process. Next, we explain our approach to construct CIDAR further.

3.1 Machine Translation

We use the Taqyim library (Alyafeai et al., 2023) to translate all the examples in ALPAGASUS using GPT-3.5 Turbo. As a preprocessing step, we first concatenated the instructions and input. After some prompt engineering, we realized that ChatGPT is translating coding blocks. Thus, we had to explicitly instruct ChatGPT to ignore coding blocks. We also append the instruction and output with *User*, and *Bot*, respectively, as shown in the following example:

> *You are given a conversation between a user and a bot, translate the full conversation into Arabic. Don't translate any coding blocks.*
> **User**: Given the context, identify a suitable word to complete the sentence. The sun feels so <mask> today, I just want to sit here and relax.
> **Bot**: warm.

3.2 Initial Review

After translating our seed dataset, we noticed some initial problems. Therefore, we followed multiple

[1] ALPAGASUS: https://hf.co/mlabonne/alpagasus.
[2] AskTheTeacher: https://aljazeera.net/ar/asktheteacher.

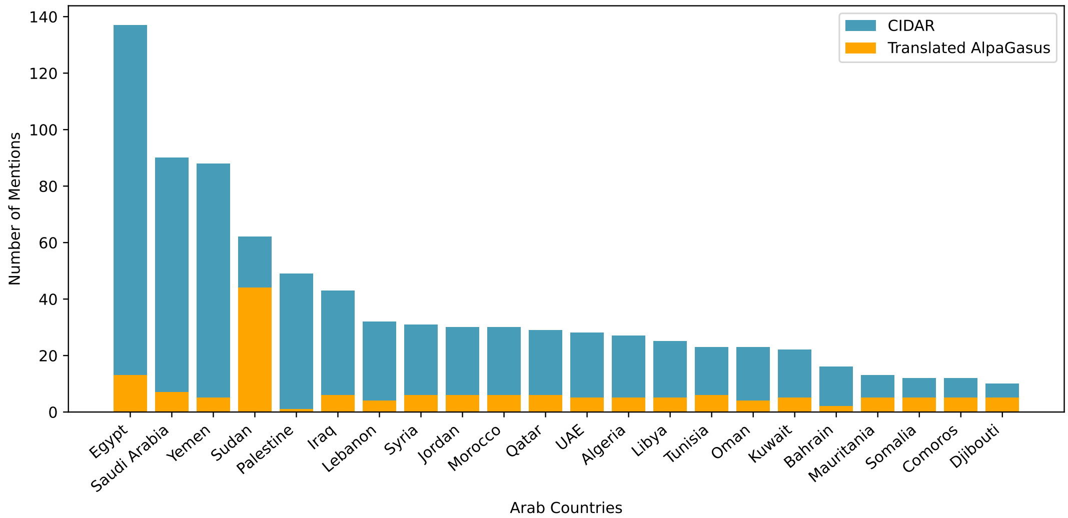

Figure 3: Number of mentions of every Arab country in both CIDAR and translated ALPAGASUS datasets.

steps to fix these machine translation issues:

- Fix instructions or outputs that contain a large number of the English alphabet.
- Fix empty fields of instructions or outputs.
- Fix manually some of the instructions that had wrong first words that are not in the correct form of an instruction.

The main goal of this step is to observe the current problems in the dataset to initialize the guidelines for the annotators.

3.3 Localization

After fixing the initial issues with translation, we prepare our dataset to be manually reviewed. To simplify the annotation process, we created a web-based Annotation Tool (see Appendix C), where reviewers were instructed to fix two main issues:

- **Linguistic Issues**: Some words might not be translated correctly, especially at the beginning of each instruction; we want all the statements to start with an instruction. For example, we should replace خلاصة (summary) with لخّص (summarize). Also, some instructions might be specific to English. The annotators are asked to provide their corresponding examples in Arabic.
- **Cultural Relevance**: Some examples in the translated AlpaGasus dataset might contain instructions and outputs that represent Western cultures. We want to replace them with samples that represent the Arab region and its culture. For instance, the name جون سميث (John Smith) should be replaced by an Arabic name like علي خالد (Ali Khalid).

In our dataset localization process, 12 native Arabic speakers voluntarily participated in localizing and reviewing all the 10,000 samples of CIDAR.

4 Dataset Analysis

We, in this section, compare between CIDAR and the initial translated ALPAGASUS to emphasize the importance of manual revision and cultural alignment of machine-generated or translated data.

4.1 Modifications

We show, in Table 2, the number of modifications made on our dataset, CIDAR, concerning the instructions, outputs, or either. Of 9,109 instruction-response pairs in ALPAGASUS dataset, there were around **64.5%** of them required a modification to be included in CIDAR dataset. These modifications are either due to a linguistic error or cultural irrelevance, as stressed in the subsection 3.3.

Modifications	# Samples
Instructions	3,202
Outputs	4,879
Instructions or Outputs	5,871

Table 2: Number of modified instructions, outputs, or either from the original translated ALPAGASUS dataset using our manual review.

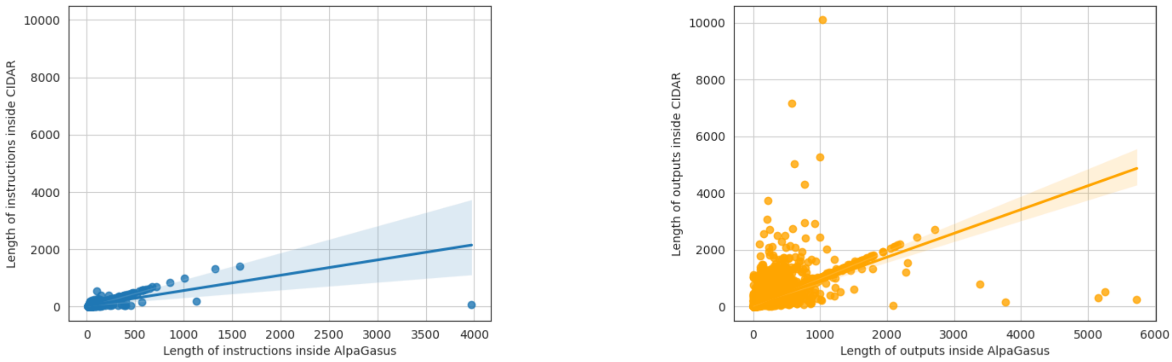

Figure 4: Comparison between CIDAR and translated ALPAGASUS in terms of instruction (Left) and output (Right) lengths. Noticeably, the length of outputs increased in CIDAR due to the possible reviewers' rewriting of outputs.

4.2 Locations and Names

The translated ALPAGASUS dataset contains a lot of Western names and countries. To calculate how much CIDAR mitigates that, we use Named Entity Recognition (NER) to extract the tokens representing persons and locations. We use a fine-tuned CAMeLBERT (Inoue et al., 2021) model on NER[3] to extract the names of persons and countries in both CIDAR and the translated ALPAGASUS. In Table 1, we draw a comparison between locations and persons in both datasets using word cloud visualizations. We can see that the majority of locations and names in CIDAR are from the Arab region.

4.3 Countries

In Figure 3, we highlight the distribution of instruction-output pairs that contain Arab countries. We observe a huge superiority for CIDAR over the translated ALPAGASUS in terms of mentioning Arab countries. In CIDAR, the mentions of Arab countries have increased noticeably after our localization. While, in ALPAGASUS, the mentions of Arab countries are mostly around ten mentions for most countries, except for Sudan (السودان)[4]. This highlights the importance of CIDAR in representing the region.

4.4 General Topics

We use keyword-based search to extract how many instruction-output pairs contain a specific topic. In Figure 5, we observe, in general, that our dataset, CIDAR, covers a wider range of topics, including Arabic-specific tasks such as poetry[5], books, diacritization, and Arabic grammar, which are much less in the translated ALPAGASUS dataset.

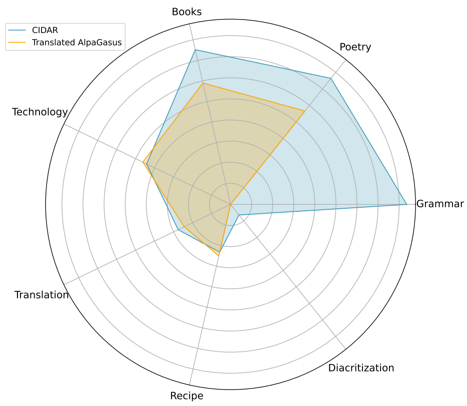

Figure 5: Comparison between CIDAR and translated ALPAGASUS datasets in terms of the covered topics.

4.5 Annotation Lengths

We, in Figure 4, compare the length of instructions and outputs between CIDAR and translated ALPAGASUS before and after our review. We highlight fewer changes in terms of instructions compared to outputs after the review. This is expected because sometimes the reviewer might re-write the whole output depending on changing a few words in the instruction. For example, if an instruction asks to find the best tourist places in a given US state, the reviewer will *likely* change one word in the instruction and completely rewrite the whole

[3]CAMeLBERT NER: https://hf.co/CAMeL-Lab/bert-base-arabic-camelbert-mix-ner.

[4]Note that Sudan is considered an outlier because many food recipes contain peanuts as an ingredient, which is translated to فول سوداني (Sudanese Bean) in Arabic.

[5]The ALPAGASUS dataset contains English poetry which is completely different from Arabic poetry.

output, which might result in a longer output.

5 Evaluation

We, in this section, shed light on the performance of LLMs after being fine-tuned on CIDAR dataset.

5.1 Experimental Setup

We employed ACEGPT-7B, a variant of LLaMA-7B pre-trained on a large Arabic corpus (Huang et al., 2023), as our base model. This model was further fine-tuned using two instruction datasets, CIDAR and ALPAGASUS, to assess their adaptability in culturally and regionally nuanced contexts. This study compares the following three variants of ACEGPT across diverse cultural and regional scenarios.

1. **ACEGPT\CIDAR**: A fine-tuned variant of ACEGPT-7B model on our culturally aligned dataset, CIDAR.
2. **ACEGPT\ALPAGASUS**: A fine-tuned variant of ACEGPT-7B model on translated ALPAGASUS dataset.
3. **ACEGPT\CHAT**[6]: The instruct-tuned variant of ACEGPT-7B model released by the original authors (Huang et al., 2023).

We fine-tuned the first two models using supervised fine-tuning (SFT) with the Quantized Low-Rank Adaptation (QLoRA) technique (Dettmers et al., 2023). We provide detailed specifications of the fine-tuning and inference hyper-parameters in Appendix E. We, in Table 3, compare the number of instructions used to fine-tune each model. Note that ACEGPT\CHAT is fine-tuned on 30x more data compared to the other models.

Model	# Instructions
ACEGPT\CIDAR	10,000
ACEGPT\ALPAGASUS	9,230
ACEGPT\CHAT	363,155

Table 3: Number of instructions used for fine-tuning each model in our evaluation study.

5.2 Qualitative Analysis

We qualitatively analyze the outputs of the three fine-tuned models used in this study and find that ACEGPT\CIDAR model better adheres to the Arab region's culture. We display, in Figure 6, a qualitative example to showcase the outputs of the three models on a given instruction. In this example, we want to know which model can utilize the names that are related to Arabic culture. We observe that ACEGPT\CIDAR demonstrates a marked improvement in aligning with Arabic culture by choosing a perfume name that is related to our region. In contrast, the ACEGPT\ALPAGASUS shows a tendency towards creating English and French names. We also observe that ACEGPT\CHAT generated a list of suggestions of the names, even though this was not requested in the instruction. We also share a few qualitative examples in Table 6 in Appendix F.

5.3 Multiple Choice Analysis

We create CIDAR-MCQ-100, a dataset containing 100 multiple-choice questions with answers that are culturally relevant to the Arab region to evaluate the three fine-tuned models. We integrated the dataset with `lm-evaluation-harness` (Gao et al., 2023) and tested with two prompts. 1) A prompt that formulates the dataset as a multiple choice problem, where the question and the multiple choices are used within the input, and 2) a prompt that formulates the dataset in open-form question format, where the input takes only the question. Note that `lm-evaluation-harness` uses two metrics for multiple-choice tasks: accuracy and normalized accuracy. The accuracy computes the log probability of each option within the multiple-choice set given the input. However, this metric can introduce bias by picking a shorter answer over a longer one. Therefore, we used normalized accuracy, which addresses this issue by calculating the average log probability per character, removing any bias toward answer length. Figure 7 shows that ACEGPT\CHAT outperforms in multiple-choice format, achieving 39% in normalized accuracy. On the other hand, ACEGPT\CIDAR outperforms in open-form questions, achieving 39% in normalized accuracy. These findings are consistent with the fact that multiple-choice questions are not presented in CIDAR, whereas the open-ended questions are more aligned closely with the mode of completion objectives. In both prompts' assessments, we highlight that ACEGPT\CIDAR achieve better results compared to ACEGPT\ALPAGASUS.

5.4 GPT Analysis

For this experiment, we create CIDAR-EVAL-100, a dataset containing 100 instructions that are cul-

[6] ACEGPT\CHAT: https://huggingface.co/FreedomIntelligence/AceGPT-7B-chat.

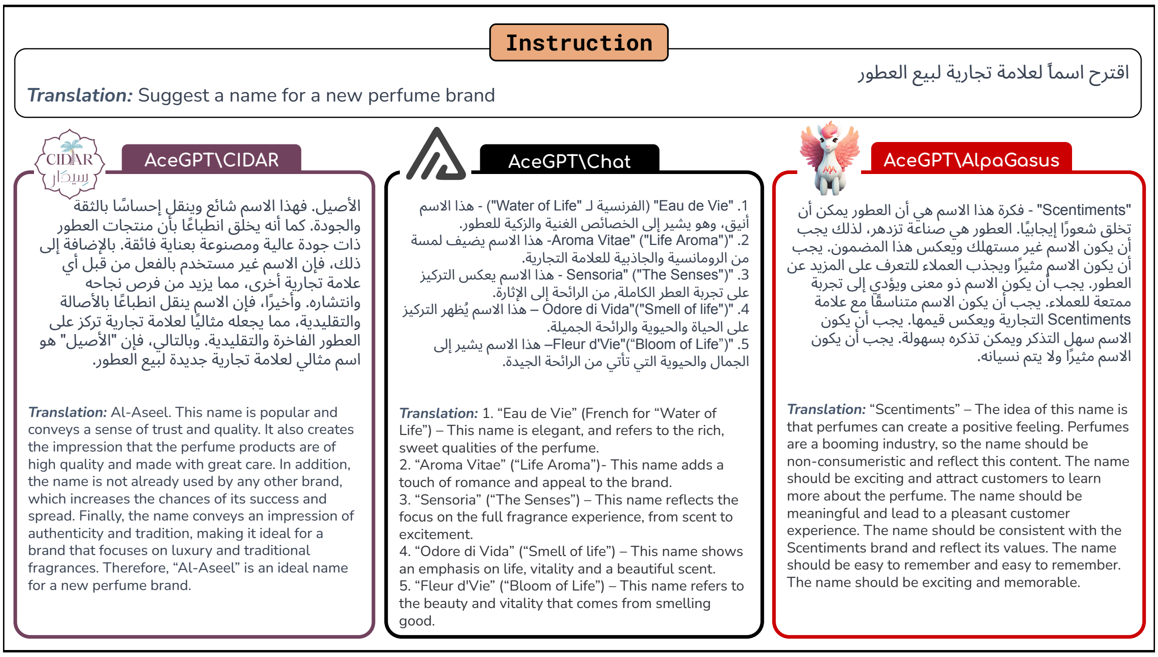

Figure 6: Comparison between the outputs of the three evaluated models on a given instruction. All the instructions are from CIDAR-EVAL-100. The output of ACEGPT\CIDAR model reveals a remarkable improvement.

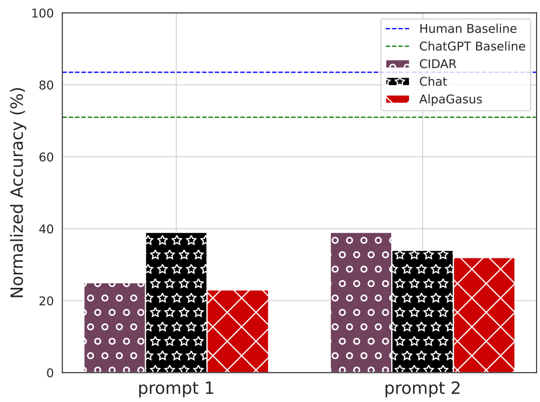

Figure 7: Performance comparison of ACEGPT\CHAT and models fine-tuned on CIDAR and ALPAGASUS on the CIDAR-MCQ-100 using normalized accuracy.

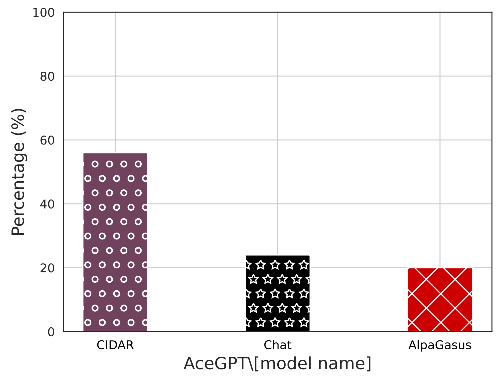

Figure 8: Win percentage for each model after feeding the responses to GPT-3.5.

turally relevant to the Arabic region. We use these instructions to generate responses for the three fine-tuned models in the study and then feed their responses to the GPT-3.5 Turbo to rank their outputs descendingly in terms of the best representation of the Arab region. As we observe from Figure 8, the best results are achieved by the model fine-tuned on CIDAR, which shows that such a model is more relevant to the region. Interestingly, such a model achieves more than **50%** win rate, which shows its dominance compared to other models that are trained on 30x larger data, i.e. ACEGPT\CHAT.

6 Related Work

In the literature, there are many English instruction datasets, whether generated by LLMs like Alpaca (Taori et al., 2023), Dolly (Conover et al., 2023), and SELF-INSTRUCT (Wang et al., 2023), or human-generated with templates like Flan collections (Wei et al., 2021; Longpre et al., 2023), P3 (Bach et al., 2022), and NATURAL INSTRUCTIONS (Mishra et al., 2022).

6.1 Multilingual Instruction-tuning Datasets

Many multilingual instruction-tuning datasets have been translated from English to Arabic using prompts or directly translating the instructions.

For example, xP3 (Crosslingual Public Pool of Prompts), which is an extension of the P3 dataset (Sanh et al., 2022), is constructed of applying English prompts across 16 NLP tasks for 46 languages, including Arabic (Muennighoff et al., 2023). Later, the authors released xP3x (xP3 eXtended) covering English prompts for 277 languages, including Arabic and ten of its Arabic dialects. MULTILINGUALSIFT (Multilingual Supervised Instruction Fine-tuning) is also created by translating instructions for 11 languages, including Arabic Chen et al. (2023c). The authors translated Alpaca-GPT4 (Peng et al., 2023), Evol-Instruct (Xu et al., 2023), and ShareGPT (Zheng et al., 2023), from English to Arabic using GPT-3.5 Turbo. The Multilingual Instruction-Tuning Dataset (MITD) (Upadhayay and Behzadan, 2023) is another dataset that is composed of the translation of Alpaca-GPT4 (Peng et al., 2023), Dolly (Conover et al., 2023), and Vicuna Benchmark (Chiang et al., 2023) in 132 languages, including Arabic, using Google Cloud AI Translation[7]. Lastly, the Bactrian-X dataset comprises 3.4M instruction-response pairs for 52 human languages, including Arabic, with around 65.4K pairs, which have been translated selected instructions from Alpaca (Taori et al., 2023) and Dolly (Conover et al., 2023), using Google Translate to Arabic. After that, the authors generated responses for such instructions using GPT-3.5 Turbo.

On the other hand, a few multilingual instruction-tuning datasets have been proposed from human-generated and human-annotated examples or conversations using templates. For instance, SUPER-NATURALINSTRUCTIONS (SUP-NATINST) benchmark consists of 1,616 diverse NLP tasks, besides their expert-written instructions, and covers nearly 76 distinct task types, spanning 55 languages, and includes 80.3K Arabic instructions for 16 Arabic NLP tasks (Wang et al., 2022). The OpenAssistant Conversations (OASST1) is made of a human-generated and human-annotated assistant-style conversation dataset consisting of 161.4K messages in 35 human languages, including Arabic, resulting in over 10K complete and fully annotated conversation trees (Köpf et al., 2023). In a concurrent work, Singh et al. (2024) released the AYA dataset, a multilingual instruction-tuning dataset with 204K instructions and responses, around 14K of which are in dialectal Arabic. The authors invited human reviewers (crowdsourcing) to contribute and review data samples, yet no cultural alignment or regional localization has been implemented on the dataset.

6.2 Arabic Instruction-tuning Datasets

A few Arabic-specific LLMs have been instruct-tuned on closed (not publicly released) Arabic instruction-tuning datasets. For example, PHOENIX (Chen et al., 2023b) has been instruct-tuned using three groups of instructions, including post-translated multilingual instructions, created by translating Alpaca instruction and output pairs (Taori et al., 2023) using GPT-4 to Arabic and sometimes by generating responses for the GPT-4 translated instructions using GPT-3.5. NOON (Naseej, 2023) has also been instruct-tuned on a collection of Arabic instructions from different datasets, such as Alpaca-GPT4 (Peng et al., 2023), Dolly (Conover et al., 2023), TruthfulQA dataset (Lin et al., 2022), Grade School Math dataset (Cobbe et al., 2021), and Arabic arithmetic problems generated using GPT-3.5 Turbo. JAIS(Sengupta et al., 2023) have been instruct-tuned using a translated collection of instructions to Arabic from various instructions-tuning datasets, such as SUPER-NATURALINSTRUCTIONS (Wang et al., 2022), Unnatural (Honovich et al., 2023), NaturalQuestions (Kwiatkowski et al., 2019), Alpaca (Taori et al., 2023), HC3 (Guo et al., 2023), Dolly (Conover et al., 2023), Basic-Conv[8], Bactrian-X (Li et al., 2023) and enriched the collection of instructions with Arabic examples from xP3 (Muennighoff et al., 2023). ACEGPT (Huang et al., 2023) has been instruct-tuned using instructions compiled from some open-source datasets, like Alpaca (Taori et al., 2023), Alpaca-GPT4 (Peng et al., 2023), Evol-Instruct (Xu et al., 2023), Code-Alpaca (Chaudhary, 2023), and ShareGPT (Zheng et al., 2023), and translated the questions from English to Arabic and regenerated the responses using GPT-4. AlGhafa model (Almazrouei et al., 2023) used many translated Arabic instruction-tuning datasets, including xP3 (Muennighoff et al., 2023), Bactrian-X (Li et al., 2023), Alpaca (Taori et al., 2023), and UltraChat (Ding et al., 2023). The *only* stand-alone (without models) open-source monolingual, Arabic instruction-tuning dataset is released by Yasbok (2023), which is poorly translated from the Alpaca dataset (Taori et al., 2023) to Arabic using Google Translate without human review, cultural

[7]Google Cloud AI Translation: https://cloud.google.com.

[8]ChatterBot Corpus: https://chatterbot-corpus.docs.io.

alignment, or translation error checking.

7 Conclusion

In this work, we present CIDAR, the *first* open Arabic instruction-tuning dataset that is culturally aligned by native Arabic reviewers to address the drawbacks of the conventional approach of fine-tuning LLMs on machine-generated or machine-translated datasets. Additionally, we introduce two datasets, CIDAR-EVAL-100 and CIDAR-MCQ-100, for evaluating LLMs on cultural relevance for Arabic. Using such benchmarks and via thorough analyses, we demonstrate that CIDAR is useful for enriching research efforts in culturally aligning LLMs with the Arabic culture. The experiments conducted validate our datasets' cultural relevance and highlight their potential to enhance the performance and understanding of LLMs within the rich Arabic linguistic and cultural context.

8 Broader Impact

We aim to establish CIDAR with the primary goal of incorporating rich Arabic content that authentically reflects our cultural values and the linguistic beauty of the language. Unlike much of the existing literature that relies on translated datasets or LLM-generated responses, which may encounter many challenges, as previously discussed, our primary focus is on preserving the integrity and quality of the Arabic culture. Moreover, the original Alpaca or ALPAGASUS mostly features Western cultural themes, such as food recipes, poems, tourist destinations, names, and countries. In our endeavor to curate CIDAR, we have diligently ensured the inclusion of elements specific to our culture and traditions, encompassing Arabic linguistic nuances, narratives, tourism, names, culinary recipes, poetry, and countries. The open release of the dataset allows for culturally-aligned fine-tuning of LLMs that undoubtedly can help with different domains. Our pilot study on fine-tuning ACEGPT reveals the huge impact such datasets can have in the region.

9 Limitations

CIDAR poses some limitations related to the data curation process. We summarize them as follows:

- **Country Biases**: Localizing a given instruction usually depends on the nationality of the person annotating. Often, annotators will prefer to add annotations related to the countries they were born in or currently residing in.
- **Dataset Size**: The size of the dataset might limit its uses in large-scale instruction tuning. In our evaluation, we attempted to show that it helps to train on a culturally relevant dataset.
- **Topics Covered**: In our data localization process, we tried to cover as many topics that are related to the culture of the region. We opted out of topics related to religion as it is considered a sensitive topic in the region.
- **Dialects**: The Arabic language is not limited to Modern Standard Arabic (MSA). There are various Arabic dialects. Localization of data was limited to corrections of the translated text, which is mostly written in MSA, without incorporating multiple dialects.
- **Safety**: Due to the relatively small size of CIDAR, the fine-tuned models on our dataset can show some degree of hallucinations, especially since it is not subjected to further alignment processes such as Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022).

References

Alberto Acerbi and Joseph M. Stubbersfield. 2023. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Ameeta Agrawal, Lisa Singh, Elizabeth Jacobs, Yaguang Liu, Gwyneth Dunlevy, Rhitabrat Pokharel, and Varun Uppala. 2023. All Translation Tools Are Not Equal: Investigating the Quality of Language Translation for Forced Migration. In *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10.

Hend Al-Khalifa, Khaloud Al-Khalefah, and Hesham Haroon. 2024. Error Analysis of Pretrained Language Models (PLMs) in English-to-Arabic Machine Translation. *Human-Centric Intelligent Systems*.

Ebtesam Almazrouei, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, Guilherme Penedo, Giulia Campesan, Mugariya Farooq, Maitha Alhammadi, Julien Launay, and Badreddine Noune. 2023. AlGhafa Evaluation

Benchmark for Arabic Language Models. In *Proceedings of ArabicNLP 2023*, pages 244–275, Singapore (Hybrid). Association for Computational Linguistics.

Saied Alshahrani, Norah Alshahrani, Soumyabrata Dey, and Jeanna Matthews. 2023. Performance Implications of Using Unrepresentative Corpora in Arabic Natural Language Processing. In *Proceedings of ArabicNLP 2023*, pages 218–231, Singapore (Hybrid). Association for Computational Linguistics.

Saied Alshahrani, Hesham Haroon Mohammed, Ali Elfilali, Mariama Njie, and Jeanna Matthews. 2024. Leveraging Corpus Metadata to Detect Template-based Translation: An Exploratory Case Study of the Egyptian Arabic Wikipedia Edition. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 31–45, Torino, Italia. ELRA and ICCL.

Saied Alshahrani, Esma Wali, and Jeanna Matthews. 2022a. Learning From Arabic Corpora But Not Always From Arabic Speakers: A Case Study of the Arabic Wikipedia Editions. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 361–371, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Saied Alshahrani, Esma Wali, Abdullah R Alshamsan, Yan Chen, and Jeanna Matthews. 2022b. Roadblocks in Gender Bias Measurement for Diachronic Corpora. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 140–148, Dublin, Ireland. Association for Computational Linguistics.

Zaid Alyafeai, Maged S Alshaibani, Badr AlKhamissi, Hamzah Luqman, Ebrahim Alareqi, and Ali Fadel. 2023. Taqyim: Evaluating Arabic NLP Tasks Using ChatGPT Models. *arXiv preprint arXiv:2306.16322*.

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Alshaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland. 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.

Eva Cetinic. 2022. The Myth of Culturally Agnostic AI Models. *arXiv preprint arXiv:2211.15271v2*.

Sahil Chaudhary. 2023. Code Alpaca: An Instruction-Following Llama Model for Code Generation. Last accessed 2024-01-15.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2023a. AlpaGasus: Training A Better Alpaca with Fewer Data. *arXiv preprint arXiv:2307.08701*.

Yan Chen, Christopher Mahoney, Isabella Grasso, Esma Wali, Abigail Matthews, Thomas Middleton, Mariama Njie, and Jeanna Matthews. 2021. Gender Bias and Under-Representation in Natural Language Processing Across Human Languages. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 24–34, New York, NY, USA. Association for Computing Machinery.

Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2023b. Phoenix: Democratizing ChatGPT across Languages. *arXiv preprint arXiv:2304.10453*.

Zhihong Chen, Shuo Yan, Juhao Liang, Feng Jiang, Xiangbo Wu, Fei Yu, Guiming Hardy Chen, Junying Chen, Hongbo Zhang, Li Jianquan, Wan Xiang, and Benyou Wang. 2023c. MultilingualSIFT: Multilingual Supervised Instruction Fine-tuning. Last accessed 2024-01-15.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. Last accessed 2024-01-15.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM. Last accessed 2024-01-15.

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*.

Alok Das. 2020. Neural Machine Translation (NMT): Inherent Inadequacy, Misrepresentation, and Cultural Bias. *International Journal of Translation*, 32:115–145.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv preprint arXiv:2305.14314*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.

Xiangjue Dong, Yibo Wang, Philip S. Yu, and James Caverlee. 2023. Probing Explicit and Implicit Gender Bias through LLM Conditional Text Generation. *arXiv preprint arXiv:2311.00306v1*.

Galileo. 2023. LLM Hallucination Index: A Ranking and Evaluation Framework For LLM Hallucinations. *Rungalileo.io*. Last accessed 2024-02-13.

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 McDonell, 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. 2023. A Framework for Few-Shot Language Model Evaluation. Last accessed 2024-02-13.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. *arXiv preprint arXiv:2301.07597*.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic Gender Identification and Reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy. Association for Computational Linguistics.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2023. AceGPT, Localizing Large Language Models in Arabic. *arXiv preprint arXiv:2309.12053v4*.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. 2024. AI Alignment: A Comprehensive Survey. *arXiv preprint arXiv:2310.19852v3*.

Atoosa Kasirzadeh and Iason Gabriel. 2023. In Conversation With Artificial Intelligence: Aligning Language Models With Human Values. *Philosophy & Technology*, 36(2):27.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. OpenAssistant Conversations – Democratizing Large Language Model Alignment. *arXiv preprint arXiv:2304.07327v2*.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-X: A Multilingual Replicable Instruction-Following Model with Low-Rank Adaptation. *arXiv preprint arXiv:2305.15011*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. 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, and Adam Roberts. 2023. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Maria Lopez-Medel. 2021. Gender Bias in Machine Translation: An Analysis of Google Translate in English and Spanish. *Academia.edu*.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2023. Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions. *arXiv preprint arXiv:2309.12342v2*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-Task Generalization via Natural Language Crowdsourcing Instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Hamdy Mubarak, Hend Al-Khalifa, and Khaloud Suliman Alkhalefah. 2024. Halwasa: Quantify and Analyze Hallucinations in Large Language Models: Arabic as a Case Study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8008–8015, Torino, Italia. ELRA and ICCL.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual Generalization through Multitask Finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Ranjita Naik, Spencer Rarrick, and Vishal Chowdhary. 2023. Reducing Gender Bias in Machine Translation through Counterfactual Data Generation. *arXiv preprint arXiv:2311.16362v1*.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2023. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. *arXiv preprint arXiv:2305.14456v2*.

Naseej. 2023. Noon: A 7-Billion Parameter Arabic Large Language Model. Last accessed 2024-01-15.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155v1*.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11).

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction Tuning with GPT-4. *arXiv preprint arXiv:2304.03277v1*.

Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. 2022. Cultural Incongruencies in Artificial Intelligence. *arXiv preprint arXiv:2211.13069v1*.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications*, 32:6363–6381.

Jules Roscoe. 2024. A 'Shocking' Amount of the Web Is Already AI-Translated Trash, Scientists Determine. *VICE*. Last accessed 2024-02-13.

Hadeel Saadany, Ashraf Tantawy, and Constantin Orasan. 2024. Cyber Risks of Machine Translation Critical Errors: Arabic Mental Health Tweets as a Case Study. *arXiv preprint arXiv:2405.11668*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng-Xin Yong, Harshit Pandey, Michael Mckenna, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *ICLR 2022 - Tenth International Conference on Learning Representations*, Online.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models. *arXiv preprint arXiv:2308.16149*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. *arXiv preprint arXiv:2402.06619*.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. Last accessed 2024-01-15.

Vishesh Thakur. 2023. Unveiling Gender Bias in Terms of Profession Across LLMs: Analyzing and Addressing Sociological Implications. *arXiv preprint arXiv:2307.09162v3*.

Brian Thompson, Mehak Preet Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. A Shocking Amount of the Web is Machine Translated: Insights from Multi-Way Parallelism. *arXiv preprint arXiv:2401.05749*.

Stefanie Ullmann and Danielle Saunders. 2021. Google Translate is sexist. What it needs is a little gender-sensitivity training. Last accessed 2024-01-15.

Bibek Upadhayay and Vahid Behzadan. 2023. TaCo: Enhancing Cross-Lingual Transfer for Low-Resource Languages in LLMs through Translation-Assisted Chain-of-Thought Processes. *arXiv preprint arXiv:2311.10797*.

Krzysztof Wach, Cong Doanh Duong, Joanna Ejdys, Rūta Kazlauskaitė, Pawel Korzynski, Grzegorz Mazurek, Joanna Paliszkiewicz, and Ewa Ziemba. 2023. The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review*, 11(2):7–30.

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

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. 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:2109.01652v5*.

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*.

Yasbok. 2023. Alpaca Arabic Instruct. Last accessed 2024-01-15.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias. *arXiv preprint arXiv:2306.15895v2*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685v4*.

A Evaluation Benchmarks

To evaluate the cultural relevance of LLMs, we introduce CIDAR-EVAL-100 and CIDAR-MCQ-100. The two benchmarks, to the best of our knowledge, are the *first* of their kind to assess the Arabic culture alignment. CIDAR-EVAL-100 and CIDAR-MCQ-100 contain 100 questions each and together cover 17 different categories related to Arabic culture, such as Language, Literature, Geography, etc. The questions were crafted manually by native Arabic speakers to ensure their relevance to the Arabic culture. Categories covered are listed in Table 4.

CIDAR-EVAL-100 consists of open free-form questions to evaluate responses against Arabic culture. Due to the difficulty of evaluating LLMs on open free-form questions and the need for automatic evaluation, we introduce CIDAR-MCQ-100, which contains MCQs written manually by native Arabic speakers to assess the cultural relevance of LLMs.

Table 4: CIDAR-EVAL-100 and CIDAR-MCQ-100 category distribution

Category	CIDAR-EVAL-100	CIDAR-MCQ-100
Food & Drinks	14	8
Names	14	8
Animals	2	4
Language	10	20
Jokes & Puzzles	3	7
Religion	5	10
Business	6	7
Cloths	4	5
Science	3	4
Sports & Games	4	2
Tradition	4	10
Weather	4	2
Geography	7	8
General	4	3
Fonts	5	2
Literature	10	2
Plants	3	0
Total	100	100

B CIDAR Data Card

We follow the style of Costa-jussà et al. (2022) and adopt their data card template to document the CIDAR dataset.

B.1 Data Description

- Dataset Summary: CIDAR *is a 10k culturally aligned dataset adopted from* ALPAGASUS.
- Dataset Access: *You can access* CIDAR *at* https://huggingface.co/datasets/arbml/CIDAR.

B.2 Data Structure

Dataset is uploaded as a single file in parquet format with 3 features: instruction, output, and index.

B.3 Data Creation

- Source Data: *The dataset was created by selecting around 9,109 samples from* ALPAGASUS *dataset and then translating it using ChatGPT. In addition, we appended that with around 891 instructions from the website Ask the Teacher.*
- Data Adoption: *The 10,000 samples were reviewed by around 12 reviewers, who are from different Arab countries, backgrounds, and education levels.*

B.4 Considerations when using CIDAR

CIDAR *is intended for research purposes only. The authors disclaim any responsibility for the misuse and condemn any use contrary to Arabic culture or Islamic values.* CIDAR *is a result of a collaborative effort, and all of its entries do not necessarily represent the beliefs and cultural background of all contributors. Even though subjected to human verification, there is no guarantee that* CIDAR *is entirely aligned with Arabic culture and Islamic values. Also, no guarantee that fine-tuned models on* CIDAR *will always respond in alignment with Arabic culture and Islamic values. Users are urged to exercise caution, employ critical thinking, and seek guidance when necessary.*

B.5 Additional Information

- Dataset Curators: *The dataset was collected through crowdsourcing.*
- Licensing Information: *The dataset is released under CC-BY-NC. The text and copyright (where applicable) remain with the original authors or publishers. Please adhere to the applicable licenses provided by the original authors.*
- Citation Information: CIDAR *Team et al.,* CIDAR*: Culturally Relevant Instruction Dataset For Arabic, 2024.*

C Annotation App

The annotation app contains two main parts: English and Arabic. Reviewers can make changes to `Instruction` and `Output` to fix mistakes and align data with the Arabic culture. The original English instructions are shown to guide the reviewers for better re-annotation of the data. We have given the annotators 2 tasks (see Subsection 3.3) that they should take into consideration during the annotation process. We require the annotators to write their names in the bottom left corner. The annotators can use *Total Contributions* to keep track of their contributions to CIDAR and *Remaining* to keep track of the remaining samples to be re-annotated. We also allow the annotators to observe

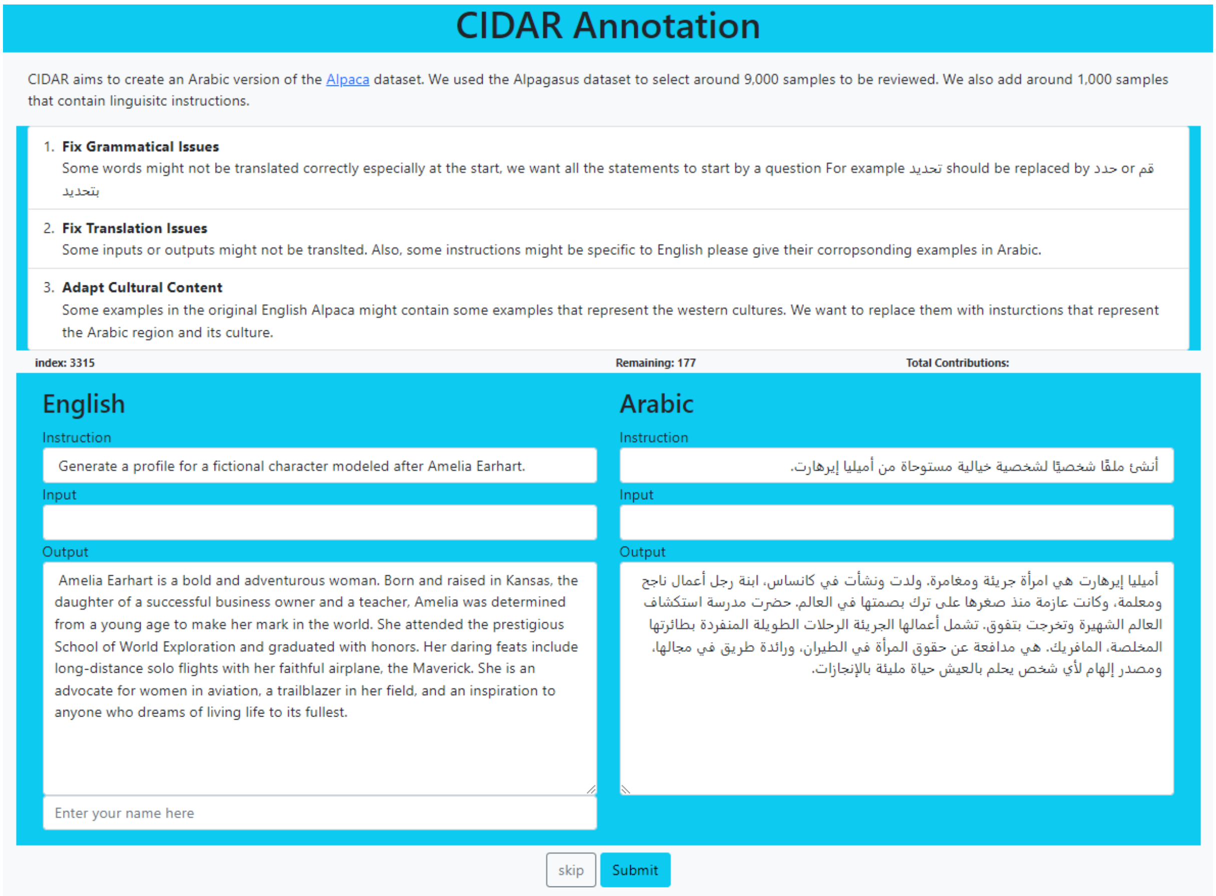

Figure 9: A screenshot of CIDAR Annotation App, showing its features. The annotators can use it to fix grammatical issues, fix translation issues, and culturally localize a given instruction and output pair from any given dataset.

the reviewed submissions and track the distribution of contributions. The website is designed using the Flask framework[9]. The app regularly (every 1 hour) pushes the changes to the Hugging Face to save the progress. The web-based annotation tool is deployed using the Railway service[10].

D Instruction Datasets

In Table 5, we showcase the main instruction-tuning datasets that include Arabic subsets/versions from the literature. We highlight that, to the best of our knowledge, all the datasets used to instruct-tuned Arabic LLMs are mostly machine-generated without human review or editing.

E Used Hyper-parameters

This section provides detailed specifications of the hyper-parameters used in the inference and fine-tuning of the AceGPT-7B model.

Table 6 details the fine-tuning hyper-parameters employed to optimize the models' performance. It includes adjustments to learning rates, batch sizes, and regularization, alongside LoRA adaptations and precision formats. Specifically, we loaded the models in 4-bit precision and used for LoRa a low rank (r) of 16 and a scaling factor (alpha α) of 16.

In the inference setup, we used the `text-generation` pipeline from the Hugging Face Transformers[11] with the following hyper-parameters: `max_length=512` to constrain output length, `temperature=0.2` for lower randomness favoring higher probability tokens, `top_p=1.0` and `top_k=0` allowing full probability distribution without restricting to top tokens, `repetition_penalty=1.2` to reduce repetition, and `do_sample=True` to enable stochastic sampling. These settings were chosen carefully to balance coherence and context relevance, aligning with our objectives for high-quality and diverse linguistic output.

[9]Flask Framework: https://flask.palletsprojects.com.
[10]Railway: https://www.railway.app.
[11]Pipelines: hf.co/docs/transformers/main_classes/pipelines.

Table 5: Collection of Arabic instruction-tuning datasets discussed in the literature (Section 6), highlighting their Arabic instructions count, dataset collection, type (multilingual or monolingual), and access status (open or closed).

Dataset Name	Size (ar)	Dataset Collection	Type	Status
xP3 (Muennighoff et al., 2023)	2,148,955	Prompts applied to multiple datasets	Multilingual	Open
MSIFT (Chen et al., 2023c)	114,231	Translated using GPT4: Alpaca-GPT4, Evol-Instruct, ShareGPT	Multilingual	Open
OASST1 (Köpf et al., 2023)	666	Conversational data was collected using a web app interface and obtained through crowd-sourcing.	Multilingual	Open
xP3x (Muennighoff et al., 2023)	18,246,158	An extended large version of the xP3 dataset with multi-dialectal Arabic instructions, besides the Modern Standard Arabic instructions.	Multilingual	Open
SUPNATINST (Wang et al., 2022)	80,396	A large benchmark was collected through a large community effort on GitHub with the help of university students and NLP practitioners.	Multilingual	Open
MITD (Upadhayay and Behzadan, 2023)	81,451	A composed multilingual instruction-tuning dataset from Alpaca-GPT4, Databricks' Dolly, and Vicuna Benchmark in 132 languages, including Arabic, was translated using Google Cloud Translation.	Multilingual	Open
Bactrian-X (Li et al., 2023)	67,017	Translated Alpaca using Google Translate then Feed to GPT3.5 Turbo.	Multilingual	Open
AYA Dataset (Singh et al., 2024)	14,210	Manually collected through crowdsourcing.	Multilingual	Open
alpaca-arabic-instruct (Yasbok, 2023)	52,002	Alpaca translated using Google Translate	Monolingual	Open
Jais Instructions (Sengupta et al., 2023)	3,683,144	xP3-Ar, Super-NaturalInstructions-Ar, Baize-Ar, Unnatural-Ar, Natural Questions-Ar, Bactrian-Ar, Alpaca-Ar, SafetyQA-Ar, NativeQA-Ar, Dolly-Ar, HC3-Ar, NER-Ar, Basic-Conv-Ar	Monolingual	Closed
AceGPT Instructions (Huang et al., 2023)	363,155	Collection of instructions from Quora-Arabic, Alpaca-Arabic, Code-Alpaca-Arabic, Evol-Instruct-Arabic, ShareGPT.	Monolingual	Closed
AlGhafa Instructions (Almazrouei et al., 2023)	1,459,000	xP3-Ar, Bactrian-Ar, Alpaca-Ar, UltraChat-Ar	Monolingual	Closed
Noon Instructions (Naseej, 2023)	110,000	Alpaca Instructions GPT4, Self-instruct records, Databricks, TruthfulQA, Grade School Math, Arabic-arithmetic-ChatGPT	Monolingual	Closed
Phoenix Instructions (Chen et al., 2023b)	8,000	A collection of translated Alpace instructions using GPT-4 to Arabic with a mixture of Arabic-generated responses for the GPT-4 translated instructions using GPT-3.5 Turbo.	Monolingual	Closed

Table 6: List of the fine-tuning hyperparamters for the models fine-tuned on CIDAR and the translated ALPAGASUS.

Parameter	Value	Parameter	Value
lora_r	16	lora_alpha	16
lora_dropout	0.1	bnb_4bit_compute_dtype	"bfloat16"
bnb_4bit_quant_type	"nf4"	bf16	True
num_train_epochs	3	per_device_train_batch_size	2
per_device_eval_batch_size	2	gradient_accumulation_steps	1
gradient_checkpointing	True	max_grad_norm	0.3
learning_rate	2e-4	weight_decay	0.001
optim	"paged_adamw_32bit"	warmup_ratio	0.03
group_by_length	True		

F Example Outputs

In Table 7, we give some example outputs for a few given Arabic instructions generated by the three evaluated models (ACEGPT\CIDAR, ACEGPT\ALPAGASUS, and ACEGPT\CHAT) used in this study, like 'How did our language originate? كيف نشأت لغتنا؟'. To prevent any bias, we use the same inference parameters for all the models. Furthermore, we do not generate multiple outputs or cherry-pick specific outputs for the same instruction. We provide the outputs considering various topics, like clothes, fonts, food and drinks, language, grammar, and traditions. The examples provided show that ACEGPT\CIDAR can better capture the culture compared to the ACEGPT\CHAT, which was fine-tuned on hundreds of thousands of instructions, whereas ACEGPT\ALPAGASUS produces the worst results in terms of cultural relevance and Arabic grammar.

G Computing infrastructure

ACEGPT\ALPAGASUS and ACEGPT\CIDAR were fine-tuned on an RTX-3090. We used Transformers, PEFT, and PyTorch for the training. The training process for each model lasted approximately 6 hours for 3 epochs, with a batch size of 2. This approach was used to finetune both models. For more details about the choice of fine-tuning hyperparameters, see section E and table 6.

AceGPT\CIDAR	AceGPT\Chat	AceGPT\AlpaGasus
أقترح لباس لتغطية الرأس		
الشماغ. وهناك عدة أنواع من الشماغ، إذا كان لديك خيار في اللون فالأفضل اختيار الأزرق أو الأصفر أو الأحمر أو الأسود. وإذا كان لديك خيار في النوع فالأفضل اختيار الشماغ الثقيل (الجوخ) والمزخرف بالآيات القرآنية أو الأحاديث النبوية. وإذا كان لديك خيار في الحجم فالأفضل اختيار الشماغ الكبير والواسع. ويمكن ارتداء الشماغ بطريقتين: 1- وضعه على الرأس من الأمام ورفع الطرف الأمامي على الجبهة، 2- وضعه على الرأس من الخلف ورفع الطرف الأمامي على الجبهة. ويمكن ارتداء الشماغ مع غطاء للرأس مثل الطاقية أو القبعة أو القلنسوة أو العمامة أو الكوفية.	1. Hat - A hat is a great option for covering your head and protecting it from the sun or wind while you're outdoors, such as on a hike in nature. It can also be worn indoors to keep warm during colder months of the year.	معظم الناس يفضلون ارتداء قبعة، شال، كاب، أو خوذة. إذا كنت تبحث عن شيء أكثر إبداعًا، جرب ارتداء قبعة صوف، قبعة قش، أو قبعة زهرة. بالنسبة للأشخاص الذين يرغبون في تغطية رؤوسهم بشكل أكثر تحفظًا، فإن قبعة قش أو قبعة زهرة هي الخيار الأفضل. وبالنسبة للأشخاص الذين يرغبون في تغطية رؤوسهم بشكل أكثر إبداعًا، فإن قبعة صوف أو قبعة قش هي الخيار الأفضل. وبالنسبة للأشخاص الذين يرغبون في تغطية رؤوسهم بشكل أكثر تحفظًا، فإن قبعة قش أو قبعة زهرة هي الخيار الأفضل.
أقترح ليا نوع خط مناسب للكتابة على الحائط		
الخط الذي أفضله للكتابة على الجدران هو "الثلث"، إذ يعد من أشهر وأجمل الخطوط العربية. وهو خط متصل ومتزن ومتناظر ومناسب للكتابة على الجدران وغيرها من المساحات الكبيرة. كما أنه يعطي مظهرًا جماليًا وأنيقًا للجدران. ويمكن استخدامه بأشكال مختلفة وأحجام متنوعة لإضفاء لمسة جمالية على المكان.	هذا يعتمد على الجو العام في الغرفة. إذا كان الجو كلاسيكيًا، فقد يكون "Serfi" أو "Georgia" خيارًا جيدًا. إذا كان الجو حديثًا وبسيطًا، فقد يكون "Arial" أو "Helvetica Neue" خيارًا أفضل. أما إذا كان الجو مسترخيًا ودافئًا، ف "Baskerville" أو "Lobster" قد يكون مناسبًا.	إذا كنت تبحث عن نوع خط جديد للكتابة على الحائط، فأنصحك بالنظر في Myriad Pro. هذا الخط مشهور بشكل خاص بشكله الأنيق والمريح للعين، وهو مثالي للكتابة على الحائط. كما أنه يضم مجموعة واسعة من الخطوط والأنماط المختلفة، مما يجعله خيارًا مثاليًا لأي غرفة. بالإضافة إلى ذلك، فإن Myriad Pro متوافق مع Windows و Mac OS X و Linux، مما يجعله خيارًا مثاليًا لأي مستخدم. بشكل عام، إذا كنت تبحث عن خط جديد للكتابة على الحائط، فإن Myriad Pro هو خيار مثالي!

أخبرني عن مشروب منتشر بين الناس في المجتمع		
مشروب منتشر بين الناس في المجتمع هو القهوة. إنها مشروب لذيذ وحلوى ومثالية للصباح، كما أنها مشروب غني بالطاقة ويمكن أن تساعد على البقاء متيقظًا ومنتجًا. يتم تناول القهوة في جميع أنحاء العالم ويتم تقديمها بطرق مختلفة - من القهوة التقليدية إلى قهوة الآيس كريم!	هنالك العديد من المشروبات المنتشرة بين الناس، لكن أحد الأمثلة المشهورة هو "القهوة". القهوة مشروب مُحبّ ومُنتشر في جميع أنحاء العالم. إنها غنية بالكافيين وتُستخدم عادةً لزيادة الطاقة والتركيز. بالإضافة إلى ذلك، تُستهلك القهوة في مختلف الأوقات من اليوم - صباحاً, ظهراً ومساءً.	الكركديه. لذة حقيقية وصحية. إضافة لطيفة للمشروبات الغازية والعصائر. #الكركديه_المثلج
حدد خطوات تجديد البطاقة الشخصية		
1. إعلام دائرة الأحوال المدنية بالتغييرات الهامة، مثل الزواج أو الطلاق، في غضون 30 يومًا. 2. تسليم طلب تجديد البطاقة الشخصية إلى دائرة الأحوال المدنية. 3. تقديم الوثائق المطلوبة، مثل شهادة الميلاد والمستندات الأخرى المؤكدة للهوية. 4. دفع رسوم تجديد البطاقة الشخصية. 5. استلام البطاقة الشخصية الجديدة من دائرة الأحوال المدنية. 6. تسليم البطاقة الشخصية القديمة إلى دائرة الأحوال المدنية. 7. تقديم طلب لإلغاء البطاقة الشخصية القديمة. 8. تقديم طلب لإصدار بطاقة شخصية جديدة. 9. دفع رسوم إصدار البطاقة الشخصية الجديدة. 10. استلام البطاقة الشخصية الجديدة	1. زر الموقع الإلكتروني للهيئة الحكومية المعنية (أو السفارة/القنصلية، إذا كانت بطاقتك ضمن متطلبات التجديد). 2. ابحث عن "تجديد البطاقة الشخصية" أو "تغيير / تجديد البطاقة". 3. اتبع الروابط المؤدية إلى 'التجديد' أو 'التغيير'. 4. سجّل الدخول إلى حسابك (إذا كان لديك واحد) واتبع الإرشادات المقدمة. 5. قد تحتاج إلى إرسال بعض المستندات - مثل صورة بطاقتك الشخصية الحالية، وثائق الهوية، وغيرها من المستندات المطلوبة. 6. اتبع الخطوات المتبقية حتى تتم عملية التجديد. 7. تأكد من تلقي تأكيد على تجديد بطاقتك الشخصية.	1. زرنا مكتب الأحوال المدنية في السعودية لإصدار بطاقة هوية جديدة. 2. قدمنا جواز السفر والصور الشخصية والمستندات الثبوتية الضرورية، مثل شهادة الميلاد وشهادة الجنسية وغيرها من الوثائق المؤكدة للهوية. 3. دفعنا الرسوم المستحقة لإصدار البطاقة الجديدة. 4. انتظرنا حتى تم إصدار البطاقة الجديدة وتسلمناها. وفي بعض الحالات، يمكن تسليم البطاقة في غضون بضع دقائق في حين يمكن أن يستغرق ذلك من بضعة أيام إلى أسبوع في حالات أخرى. 5. تأكدنا من أن جميع المعلومات المدخلة في البطاقة صحيحة ومطابقة للوثائق المقدمة وتم توقيع البطاقة من قبل الموظف

<table dir="rtl">
<tr><td colspan="3">كيف نشأت لغتنا؟</td></tr>
<tr>
<td>نشأت اللغة عبر مئات الآلاف من السنين، وهذا جزء من طبيعة الإنسان. قد يكون الناس قد بدأوا في صنع الأصوات والضحك والتواصل معًا قبل ظهور اللغة المكتوبة. ثم تطورت اللغة على مر السنين من خلال التفاعل والتبادل بين الناس في جميع أنحاء العالم. وقد لعبت الثقافة والتكنولوجيا دورًا في تطوير اللغة، مما أدى إلى ظهور لغات جديدة وتطور اللغات القديمة. وقد ساهمت اللغة في تطوير التواصل والتفاعل بين الناس وتمكين التعاون والتعاطف والتفاهم.</td>
<td>معظم البحوث تقترح أن اللغة تطورت بالتدريج. بدأت بأصوات بسيطة، مثل الزئير والهرير، وتطورت إلى أشكال أكثر تعقيدًا من التواصل. في المراحل الأخيرة من التطور، بدأ الناس في تعلم كيفية استخدام الكلمات والجمل للتعبير عن الأفكار والمشاعر. هذا التطور في اللغة مستمر، مع ظهور لهجات ولغات جديدة على مر الزمن.</td>
<td>العربية هي اللغة الوحيدة المسجُّلة ضمن اللغات الخمس الأصلية في الإعلان العالمي لحقوق الإنسان، وهذا يعطيها مكانةً خاصة. وقد ظهرت هذه اللغة في شبه الجزيرة العربية في القرن السابع الميلادي وانتشرت في العالم الإسلامي ثم في العالم كله. وقد أثرت هذه اللغة في اللغات الأخرى وتأثرت بها في الآنَّذاك. وتعتبر اللغة العربية من أكثر اللغات انتشارا في العالم الإسلامي وفي بعض الدول الأفريقية والأوروبية. وتستخدم هذه اللغة في الأدب والشعر والفنون والعلوم والتكنولوجيا والإعلام والسياسة والاقتصاد والتجارة والأعمال المختلفة.</td>
</tr>
<tr><td colspan="3">اعرب الجملة التالية: سبق السيف العذل.</td></tr>
<tr>
<td>سبق السيف العذل. (السيف - فاعل، العذل - مفعول)</td>
<td>"سبق السيف العذل" هو نص شعري، أو "كلام حكيم". إنه يُظهر أن الخطوة المتقدمة (السيف) قد تغلب على الضعف أو التردد (العذل).</td>
<td>- سَبْقَ: فعل ماض ناسخ، وهو مبنيّ على الفتح.- السيفُ: فاعل "سَبْقَ" مرفوع بدلًا من الضمير (ه).- العذلِ: مفعول به أول لـ"سَبْقَ".- ثمّ: ظرف زمان متعلق بـ"سَبْقَ".- السيفُ: شبه جملة في محل رفع خبر لـ"سَبْقَ".- الإعراب: سَبْقَ السيفُ العذلَ. والجملة في محل رفع خبر لـ"سَبْقَ".- الخبر: كان.وجملة "كان" في محل نصب خبر لـ"سَبْقَ".- المبتدأ: السيفُ.</td>
</tr>
<tr><td colspan="3">شكل الجملة التالية: من ذا الذي ماساء قط ومن له الحسنى فقط.</td></tr>
<tr>
<td>من ذا الذي ماساء قط، ومن له الحسنى فقط؟ "الحسنى" هنا عبارة غير صحيحة إملائياً - "حسنى". (أخطاء إملائية أخرى تشمل "حسنى"، "حسنى".)</td>
<td>أعتذر، ولكن بصفتي AI، ليس لدي القدرة على إظهار الخطأ في جملة غير مكتملة أو غير واضحة. يرجى تقديم مزيد من التفاصيل أو توضيح السؤال.</td>
<td>مَنْ ذُوِّ الذِي مَا سَاءَ قَطُّ وَمَنْ لَهُ الحَسْنَى فَقَطُّ؟!</td>
</tr>
</table>

<table dir="rtl">
<tr><td colspan="3">اقترح اسما لعلامة تجارية جديدة لبيع العطور</td></tr>
<tr>
<td>"Scentiments" - فكرة هذا الاسم هي أن العطور يمكن أن تخلق شعورًا إيجابيًا. العطور هي صناعة تزدهر، لذلك يجب أن يكون الاسم غير مستهلك ويعكس هذا المضمون. يجب أن يكون الاسم مثيرًا ويجذب العملاء للتعرف على المزيد عن العطور. يجب أن يكون الاسم ذو معنى ويؤدي إلى تجربة ممتعة للعملاء. يجب أن يكون الاسم متناسقًا مع علامة Scen-timents التجارية ويعكس قيمها. يجب أن يكون الاسم سهل التذكر ويمكن تذكره بسهولة. يجب أن يكون الاسم مثيرًا ولا يتم نسيانه. يجب أن يكون الاسم متناسقًا مع علامة Scen-timents التجارية ويعكس قيمها.</td>
<td>1. "Eau de Vie" (الفرنسية لـ "Water of Lfie") - هذا الاسم أنيق، وهو يشير إلى الخصائص الغنية والزكية للعطور. 2. "Aroma Vi-tae" ("Lfie Aroma") - هذا الاسم يضيف لمسة من الرومانسية والجاذبية للعلامة التجارية. 3. "Senso-ria" ("The Senses") - هذا الاسم يعكس التركيز على تجربة العطر الكاملة, من الرائحة إلى الإثارة. 4. "Odore di Vida" ("Smell of lfie") – هذا الاسم يُظهر التركيز على الحياة والحيوية والرائحة الجميلة. 5. "Fleur d'Vie" ("Bloom of Lfie") – هذا الاسم يشير إلى الجمال والحيوية التي تأتي من الرائحة الجيدة.</td>
<td>الأصيل. فهذا الاسم شائع وينقل إحساسًا بالثقة والجودة. كما أنه يخلق انطباعًا بأن منتجات العطور ذات جودة عالية ومصنوعة بعناية فائقة. بالإضافة إلى ذلك، فإن الاسم غير مستخدم بالفعل من قبل أي علامة تجارية أخرى، مما يزيد من فرص نجاحه وانتشاره. وأخيرًا، فإن الاسم ينقل انطباعًا بالأصالة والتقليدية، مما يجعله مثاليًا لعلامة تجارية تركز على العطور الفاخرة والتقليدية. وبالتالي، فإن "الأصيل" هو اسم مثالي لعلامة تجارية جديدة لبيع العطور.</td>
</tr>
<tr><td colspan="3">اقترح لي عبارات مناسبة لتعزية صديقي في وفاة والده</td></tr>
<tr>
<td>"أنا آسف، كنت أعلم أن الجميع يحبك. إذًا هناك شيء يمكنني فعله، فلا تتردد في الاتصال."</td>
<td>1. "أعظم التعازي لك ولعائلتك في هذا الوقت الصعب." 2. "إن وفاة والده خسارة كبيرة، فليجد الراحة في ذكرياته الجيدة." 3. "أشعر بالأسى لسماع هذا الخبر، إنه خسارة غير قابلة للتعويض." 4. "أتقدم بأحر التعازي لك ولعائلتك في هذا الأوقات المحزنة." 5. "إن وفاة والده حدث أليم، فليجد الراحة في دعم الأصدقاء والعائلة."</td>
<td>أسأل الله أن يغفر له ويرحمه، وإنا لله وإنا إليه راجعون. خالص التعازي لكم ولذويه. نسأل الله أن يربط على قلوبكم ويلهمكم الصبر والسلوان. إنا لله وإنا إليه راجعون. نشاطركم الأحزان ونؤكد لكم أن وفاة والدكم خسارة للجميع. نسأل الله أن يرحمه ويغفر له ويسكنه فسيح جناته. نسأل الله أن يربط على قلوبكم ويلهمكم الصبر والسلوان. إنا لله وإنا إليه راجعون. نسأل الله أن يرحمه ويغفر له ويسكنه فسيح جناته. نسأل الله أن يرحمه ويغفر له ويسكنه فسيح جناته.</td>
</tr>
</table>

أخبرني عن العرف الرسمي للزواج في المنطقة		
العرف الرسمي للزواج هو تحديد موعد إشهار الزواج. وغالبًا ما يكون ذلك بعد اتفاق الطرفين على موعد ثابت وإعلانه للآخرين، ويتضمن ذلك إصدار بيان رسمي بالزواج وتوجيه الدعوات للحضور. ويمكن أن يتم ذلك بطرق مختلفة وفقًا للظروف والثقافة المحلية. ويعد هذا الإعلان مؤشرًا على بدء حياة جديدة وانتهاء حياة العزوبية. ويُعتبر هذا الإعلان رسميًا ويتم توثيقه في السجلات الرسمية للزواج. ويتم تحديد موعد الزواج وفقًا للتقاليد والأعراف المحلية والدينية والثقافية والاجتماعية.	التقاليد والأعراف الرسمية للزواج تختلف كثيرًا من منطقة إلى أخرى. في بعض المناطق، قد يشمل العرف الرسمي للزواج الحصول على رخصة الزواج من السلطات المحلية، وحضور حفل زفاف رسمي مع العائلة والأصدقاء. في بعض الثقافات، قد يتضمن العرف الرسمي للزواج هدايا محددة تُقدم للعروس و/أو العريس (مثل الذهب أو الأغذية). في بعض المناطق، قد يتضمن العرف الرسمي للزواج طقوس دينية معينة. من الأفضل دائمًا التحقق من العادات والتقاليد الخاصة بالمنطقة التي تنظر فيها للزواج للتأكد من أنك تحترم ثقافتهم وتقاليدهم في هذه المراسم الهامة.	تختلف العروض الرسمية للزواج حسب الثقافة. في الهند، غالبًا ما يكون العرف الرسمي للزواج هو "شوباد" (Saptapadi)، وهو إعلان صريح بالزواج من قبل العروس والعريس. في الولايات المتحدة، يتضمن العرف الرسمي للزواج عادة عقد قانوني وحفل زفاف وربما حفل استقبال. في بعض الثقافات، يمكن أن يتضمن ذلك طقوس أخرى مثل الاحتفالات والطقوس الدينية. في العديد من الثقافات، يتم تنظيم العرف الرسمي للزواج من قبل الأسرة والأصدقاء والمجتمع. في بعض الثقافات، يتم تنظيمه من قبل الحكومة أو مؤسسة محلية أخرى.

Table 7: Example outputs for a few given Arabic instructions generated by the three evaluated models (ACEGPT\CIDAR, ACEGPT\ALPAGASUS, and ACEGPT\CHAT) used in this study. We note that some sentences have been truncated for better readability.

Animals	
ضع كلمة مناسبة لإكمال الجملة: وقفت أنظر إلي السماء فرأيت _____	أ) طائر الطنان ب) الفلامنجو ج) البلشون د) الحمام ✔
Food & Drinks	
حدد الكلمة المناسبة لإكمال الجملة: كنت في احتفال مع اصدقائي بالأمس وقمنا بشرب _____	أ) القهوة ✔ ب) الخمر ج) البيرة د) المخدرات
Language	
قم بتحويل الجملة التالية إلى صيغة الأمر: أكلت الطعام.	أ) ستأكل الطعام ب) أُكل الطعام ج) لن تأكل الطعام د) كل الطعام ✔
Religion	
يقال بعد تناول الطعام	أ) آمين ب) نخبنا جميعا ج) cheers د) الحمد لله ✔
Tradition	
حدد كلمة مناسبة لإكمال الجملة: يتم تعليق الزينة في الشوارع في شهر ______	أ) أبريل ب) يونيو ج) رمضان ✔ د) شعبان

Cloths	
أ) البشت ✔ ب) قميص مع جينز ج) بدلة فورمال د) جاكت اسود وبنطال	ضع كلمة مناسبة لإكمال الجملة: أنا سعودي وعندي حفل زفاف لذلك سأرتدي _____
Names	
أ) كارتر ب) جوناثون ج) جاك د) محمد ✔	ضع كلمة مناسبة لإكمال الجملة: قام _____ بإحضار الكثير من الهدايا لأمه
Jokes & Puzzles	
أ) امرأة حامل ب) طائر النورس ج) اليمامة د) رجل يتكأ على عصي ✔	ما هو الشئ الذي يمشي على ثلاث
Geography	
أ) إسرائيل ب) الأردن ج) فلسطين ✔ د) اليونان	حدد كلمة مناسبة لإكمال الجملة: القدس هي عاصمة ______
General	
أ) BBC Radio 1 ب) القرءان الكريم ✔ ج) Radio France Internationale د) Deutschlandfunk	حدد الكلمة المناسبة لإكمال الجملة: قمت بشغيل إذاعة _____ وأنا في السيارة.

Table 8: MCQs Samples from CIDAR-MCQ-100 marked with correct answers from 10 different categories. The answers are based on the majority voting of four different human annotators.