AceGPT, Localizing Large Language Models in Arabic

Huang Huang†1, Fei Yu†2, Jianqing Zhu†3, Xuening Sun2,4, Hao Cheng2, Dingjie Song2,4, Zhihong Chen2,4, Abdulmohsen Alharthi3, Bang An3, Juncai Ho3, Ziche Liu2, Zhiyi Zhang2, Junying Chen2,4, Jianquan Li2, Benyou Wang†2,4, Lian Zhang1, Ruoyu Sun1,2, Xiang Wan1, Haizhou Li†2,4, Jinchao Xu†

1 Shenzhen International Center for Industrial and Applied Mathematics, Shenzhen Research Institute of Big Data
2 The Chinese University of Hong Kong, Shenzhen, China
3 King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
4 Shenzhen Research Institute of Big Data, Shenzhen, China

Abstract

This paper is devoted to the development of a localized Large Language Model (LLM) specifically for Arabic, a language imbued with unique cultural characteristics inadequately addressed by current mainstream models. Significant concerns emerge when addressing cultural sensitivity and local values. To address this, the paper proposes a comprehensive solution that includes further pre-training with Arabic texts, Supervised Fine-Tuning (SFT) utilizing native Arabic instructions, and GPT-4 responses in Arabic, alongside Reinforcement Learning with AI Feedback (RLAIF) employing a reward model attuned to local culture and values. The goal is to cultivate culturally cognizant and value-aligned Arabic LLMs capable of accommodating the diverse, application-specific needs of Arabic-speaking communities. Comprehensive evaluations reveal that the resulting model, dubbed ‘AceGPT’, sets the state-of-the-art standard for open Arabic LLMs across various benchmarks. Codes, data, and models are in https://github.com/FreedomIntelligence/AceGPT.

1 Introduction

LLMs like GPT-3.5 Turbo and GPT-4 have been shaping the current landscape of natural language understanding and generation (Bubeck et al., 2023). In contrast to the proprietary nature of GPT-3.5 Turbo and GPT-4, there has been a trend towards developing open-source large language models capable of instruction-following (Taori et al., 2023) and fluent conversations (Chiang et al., 2023), a phenomenon termed as ‘Democratization of ChatGPT’ (Conover et al., 2023; Touvron et al., 2023). While these models have shown great promise in understanding and producing content in various languages, they might fail to align with local values and cultural norms in non-English environments (Chen et al., 2023a); we call it the ‘localization issue’. This issue can lead to significant problems in practical usage scenarios, especially for regions such as the Arabic world where the culture and values diverge significantly from Western norms. We argue that it is not just desirable but necessary to localize large language models and tailor them to a specific cultural environment.

Methodology

The core of our approach lies in localizing large language models to the Arabic language using a packaged solution (known as AceGPT). Firstly, through incremental pre-training on Arabic data (localized pre-training), we ensure that the model has a strong foundation in the Arabic language, including grammar, vocabulary, and cultural context. Next, by fine-tuning Arabic natural questions (localized instructions), we enable the model to effectively comprehend and respond to specific instructions that are pertinent to Arab interests. Furthermore, by generating Arabic native responses from GPT-4 (localized responses) rather than relying on translations from other languages, we ensure that the model’s outputs are natural and fluent within an Arabic context thanks to the powerful GPT-4. Lastly, by employing a reward model based on localized preference data that respects local culture and value, we further refine the model to align the responses with the cultural and value norms of Arabic-speaking communities.

Evaluation

We evaluate our models in various benchmarks: in the instruction-following benchmark, AceGPT achieves state-of-the-art (SOTA) among open-sourced Arabic LLMs in Arabic Vicuna-80 and Arabic AlpacaEval, obtaining 33% and 30% improvement over the state-of-the-art Arabic LLM (Sengupta et al., 2023) 1. In the NLU

† indicates that the three authors contributed to this work equally.

1 Jais (Sengupta et al., 2023) is a concurrent work released two weeks ahead of ours.
Table 1: Proportion of Arabic Entities in Responses to 20 Sample Arabic Questions

<table>
<thead>
<tr>
<th>Types of entity</th>
<th>Jais-13B</th>
<th>GPT-3.5 Turbo</th>
<th>GPT-4</th>
<th>AceGPT (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>12.00% (3/25)</td>
<td>26.67% (12/45)</td>
<td>39.29% (22/56)</td>
<td>50.00% (31/62)</td>
</tr>
<tr>
<td>Location</td>
<td>18.75% (3/16)</td>
<td>27.08% (13/48)</td>
<td>21.62% (16/74)</td>
<td>28.95% (11/38)</td>
</tr>
</tbody>
</table>

1 25 person names in Jais-13B responses are identified and 3 are Arabic names.

benchmark, AceGPT achieves the second best on ALUE (Seelawi et al., 2021) in terms of average scores for all tasks. In the knowledge benchmark, AceGPT achieves SOTA among open-sourced Arabic LLMs in Arabic knowledge including MMLU and EXAMs. In the localization benchmark, AceGPT achieves SOTA among open-source Arabic LLMs in our Arabic Cultural and Value Alignment (ACVA) Dataset.

Contributions The contributions of the paper are three-fold, including: 1) We propose a first-tier Arabic LLM. According to the records on the releasing date, it achieves SOTA performance among open Arabic LLMs in many benchmarks including Arabic Vicuna-80, Arabic AlpacaEval, Arabic MMLU, EXAMs, and ACVA. 2) AceGPT is the first open-source Arabic large language model that encompasses the entire LLM pipeline including pre-training, supervised fine-tuning, and reinforcement learning from AI feedback. 3) We observe and measure the localization issue in Arabic LLMs quantitatively, especially with a new benchmarking dataset, ACVA, for localization testing.

2 Recipe of AceGPT

2.1 Motivation: The Localization Issue

Given the availability of many high-quality instruction datasets in widely spoken languages such as English, existing strategies for non-English LLMs often rely on instructions translated from English. Examples include Chinese-alpaca-GPT4 (Peng et al., 2023), Phoenix (Chen et al., 2023b), and Jais (Sengupta et al., 2023). However, relying on translated data may lead to localization issues, potentially undermining the integrity and applicability of the models in native contexts.

To address these localization issues, we formulate 20 questions (see Table.13) to elicit responses with name entities—both personal and locational—to summarize the prevalence of Arabic name entities for preliminary experiments. Quantitative results in Table 1 uncovers a significant deficiency in localization, where Jais-13B and GPT-3.5 Turbo only incorporate 12.00% and 26.67% Arabic names out of all the names in their responses respectively. A specific example is shown in Table 2, we can observe that the Arabic open-source LLM Jais’s output shows a conspicuous tilt towards English-centric materials, yielding terms predominantly associated with Christianity, which potentially neglects significant parallels within Arabic literary traditions. By contrast, GPT-3.5 Turbo showcases a more diverse recognition of holy sites from different cultural backgrounds. You can see the details and more examples of case studies in Appendix B.2.

2.2 Methodology of AceGPT

To address localization, we propose a comprehensive solution including three strategies to ensure model’s effective understanding and generation of content in Arabic, with cultural awareness and value alignment: 1) localized pre-training we further pre-train LLM with Arabic data; 2) localized instructions we adopt Arabic natural questions in the wild and their responses are Arabic native responses from GPT-4 instead of translating that from other languages, and; 3) localized responses we further tame LLM with reinforcement learning using a reward model that respects local culture and values thanks to the localized preference data.

The resultant model is termed ‘AceGPT’. The model pre-trained on LLaMA2 (Touvron et al., 2023) is named ‘AceGPT-base’. To equip it with the conversation, we introduced ‘AceGPT-chat’ utilizing supervised fine-tuning and reinforcement learning from AI feedback. The training procedure is divided into three stages: pre-training, supervised fine-tuning and reinforcement learning from AI feedback, introduced in Sec 2.2.1, Sec 2.2.2, and Sec 2.2.3, respectively.

2.2.1 Localized Pre-Training

To adapt the English-focused LLaMA2 (Touvron et al., 2023) model in Arabic, we train further it with a substantial corpus of Arabic text.

Data The dataset comprises Arabic and English sub-datasets. The Arabic is derived from the
Table 2: Comparison of LLM Responses Highlighting Location, Person, and Regional/Cultural Terms. Arabic-rooted terms in blue; non-Arabic terms in red.

<table>
<thead>
<tr>
<th>User:</th>
<th>GPT 3.5 Turbo:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>What are the holy books, saints, and holy places?</strong></td>
<td>- <strong>Holy books:</strong> The main holy books are the Torah (Judaism), the New Testament (Christianity), and the Holy Qur'an (Islam). The Torah is considered sacred by Jews, the New Testament by Christians, and the Holy Qur'an by Muslims. They are revered as revelation from God. - <strong>Saints:</strong> In the three monotheistic religions, saints are holy people who are considered sacred by Christians, Muslims, and Jews. Saints are venerated and are considered to have a special connection to the divine. - <strong>Holy places:</strong> Holy places are sites considered sacred or consecrated by some religions. Examples include the Western Wall in Jerusalem (Judaism), the Al-Aqsa Mosque in Jerusalem (Islam), and the Kaaba in Mecca (Islam).</td>
</tr>
<tr>
<td><strong>Jais-13B chat:</strong></td>
<td>- <strong>The Bible</strong> is a collection of religious texts that are considered sacred by Christians. Saints are holy people who are considered to have lived exceptionally virtuous lives and are venerated as saints by some religions. Holy places are sites considered sacred or consecrated by some religions, such as churches and temples And mosques.</td>
</tr>
</tbody>
</table>

We can see from Table 3 that common entities in Arabic are mostly Western (e.g., 'John', 'Apple', and 'New York'), deviating from Arab's actual interest (e.g., 'Mohammed', 'Muslim Brotherhood', and 'Egypt') which can be addressed by Quora. The main idea of localized responses is to leverage the fact that GPT-4 produces culture- and value-relevant responses in the context of question language, which means responses to questions in English are different from those in Arabic. See an example in Table 4, GPT-4 produces culture-dependent responses based on the queried languages. Therefore, when incorporating open-source instruction-tuning data, we ask the GPT-4 to re-generate responses in Arabic (rather than translate) to produce localized responses.

**Data** In addition to Arabic Quora questions, we also incorporate some open-source instruction-tuning datasets to improve the overall performance. Specifically, we incorporate Alpaca (Taori et al., 2023; Peng et al., 2023) (the most classical instruction-tuning dataset), Evol-Instruct (Xu et al., 2023) (a complex instruction dataset), Code-Alpaca (Chaudhary, 2023) (a code-specific instruction dataset) 5, and ShareGPT 6 (a popular user-GPT dialogue dataset). For these open-source data except ShareGPT, an Arabic version is created by translating the English questions into Arabic and re-generating the responses using GPT-4. We reserve the original ShareGPT data because the original conversations will be destroyed with a re-generated different response.

---

5 We incorporate code-alpaca for a more powerful LLM with a better code capability.

6 https://huggingface.co/datasets/philschmid/sharegpt-raw
Table 3: Top 5 names of individuals, organizations, and geopolitical entities (GPE) by frequency.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Top-5 Person</th>
<th>Top-5 Organization</th>
<th>Top-5 GPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpaca</td>
<td>John, John Smith, Alice, Mary, Harry Potter</td>
<td>Apple, Amazon, Google, Microsoft, ABC</td>
<td>United States, India, New York, France, China</td>
</tr>
<tr>
<td>Evol-Instruct</td>
<td>John, John Smith, Harry Potter, Alice, Bob</td>
<td>Apple, Amazon, quantum, Google, Microsoft</td>
<td>United States, New York, Los Angeles, San Francisco, Japan</td>
</tr>
<tr>
<td>ShareGPT</td>
<td>Di Maria, Messi, Beckhaus, Eco, Clara</td>
<td>Tribunal, Google, Council, Bing, Supreme Court</td>
<td>United States, Argentina, France, New York, Hong Kong</td>
</tr>
<tr>
<td>Quora</td>
<td>Prophet, Mohammed, Adam, Hijri, Ali</td>
<td>European Union, Google Muslim Brotherhood, Soviet Union, United Nations</td>
<td>Egypt, Turkey, Saudi Arabia, Morocco, America</td>
</tr>
</tbody>
</table>

Table 4: GPT-4 answers culture-relevant questions differently across languages. Questions here are the same in semantics but differ in languages. The Arabic response is translated into English (right).

<table>
<thead>
<tr>
<th>Question in English:</th>
<th>Question in Arabic:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is it forbidden for a woman to confess her love to a man?</td>
<td>هل من اللازم أن تعبر المرأة عن حبها للرجل؟</td>
</tr>
</tbody>
</table>

**2.2.3 Reinforcement Learning from AI Feedback**

To further align AceGPT with values and cultures, we utilize reinforcement learning from AI feedback with a reward model trained with localized preference data. There are primarily two stages: (1) training the reward model using localized preference data, and (2) aligning AceGPT to value and culture preference patterns using the proximal policy optimization algorithm (Schulman et al., 2017).

**Localized preference data** To align AceGPT with Arabic culture and values, a reward model mimicking the preferences of native speakers is essential. To prepare the localized preference data for reward model training, we reuse 40K localized instructions, i.e. Quora questions, in the SFT stage and sample paired outputs from our fine-tuned 7B model. Given the resource-intensive nature of collecting human feedback, we utilized GPT-4 feedback, which has been shown to correlate highly with human preference labeling and achieves competitive performance in text summarization (Lee et al., 2023). However, due to observed position bias in GPT-4 (Zhang et al., 2023), we altered the order of sample answers and retained consistent preferences between two order-switched runs, resulting in 12K pairs. A small study with 800 examples verified the reliability of this preference data, revealing a correlation coefficient of 0.84 between GPT-4 and human evaluations. We also incorporate 12K open-source preference data for better generalization. See Appendix D for details.

**Reward model** The reward model operates within a ‘binary’ framework, determining preferences with an additional linear head post the final hidden states. The loss function is expressed as:

\[
L(\theta) = -\frac{1}{|D|} \mathbb{E}_{(x,y_r,y_c) \sim D} \left[ \log (\sigma (r_\theta (x, y_c) - r_\theta (x, y_r))) \right].
\]  

(1)

Here, \( x \) is the input, \( y_c \) is the chosen model output, \( y_r \) is the rejected model output of the pair, and \( r_\theta \) is the reward model with the parameter \( \theta \).

**Proximal policy optimization** We crawl another 30K Quora questions different from Quora-40K for PPO training data. Proximal Policy Optimization (PPO) is an off-policy policy gradient method for reinforcement learning (Schulman et al., 2017). The policy \( \pi_\theta (a|s) \) represents the probability distribution over the next token \( a \) given a sequence of previous tokens \( s \), where \( \theta \) are the model parameters. The primary objective is to maximize the preference signal from the reward model that...
corresponds to the desired output behaviour. The objective can be shown as

\[ L(\theta) = \mathbb{E}_t \left[ \min (\rho A_t, \text{clip} (\rho, 1 - \epsilon, 1 + \epsilon) A_t) \right], \]

where \( \rho \) is a factor defined by \( \rho = \pi_\theta (a_t | s_t) / \pi_{\theta_\text{old}} (a_t | s_t) \), \( \theta_\text{old} \) denotes the model parameter used for experience sampling. \( A_t \) refers to the advantage function that measures the relative value of generating \( a_t \) as the next token conditioned on the sequence \( s_1 \cdots s_t \), and \( \epsilon \) is a hyperparameter for stability.

3 Localization Evaluation

3.1 Evaluation Protocol

In this section, we delve into the ‘Localized Evaluation’ of our language model, focusing exclusively on the Arabic Cultural and Value Alignment (ACVA). ACVA serves as a critical benchmark to assess our model’s performance in terms of its alignment with Arabic cultural nuances and values. This evaluation is particularly important for understanding how well the model adapts to the specific linguistic and cultural context of the Arabic language. We conduct this evaluation using our fine-tuned chat models, which have been specifically optimized for higher relevance and accuracy in culturally specific scenarios.

Arabic Cultural and Value Alignment (ACVA) ACVA is a Yes-No question dataset, comprising over 8000 questions, generated by GPT-3.5 Turbo from 50 designed Arabic topics to assess model alignment with Arabic values and cultures (see Appendix C for data construction details and show some examples in Table 19). A subset, revised by Arabic speakers for question quality and answer accuracy, forms the 2486-data ‘Clean set’. The correlation between ‘All set’ and ‘Clean set’ is 0.9825, indicating a high reliability of ACVA all set evaluation. Notably, our AceGPT-chat models (both 7B and 13B) consistently outperform other open-source LLMs, and AceGPT-13B-chat only trails GPT-3.5 Turbo by a marginal of −0.55% on all set. Since the Jais-30B-chat does not follow instructions and cannot return answers for multiple-choice questions, we suspect that this is due to overly stringent safety measures. Therefore, we did not include Jais-30B-chat in the zero-shot comparison.

3.2 Experiment Results

ACVA benchmark We present the results of AceGPT and other chat models on ACVA in Table 5. The Pearson correlation of accuracy on ‘All set’ and ‘Clean set’ is 0.9825, indicating a high reliability of ACVA all set evaluation. Notably, our AceGPT-chat models (both 7B and 13B) consistently outperform other open-source LLMs, and AceGPT-13B-chat only trails GPT-3.5 Turbo by a marginal of −0.55% on all set. Since the Jais-30B-chat does not follow instructions and cannot return answers for multiple-choice questions, we suspect that this is due to overly stringent safety measures. Therefore, we did not include Jais-30B-chat in the zero-shot comparison.

4 Overall Evaluation

4.1 Evaluation Protocol

Evaluation of language models is multifaceted and typically involves multiple metrics and benchmarks to assess various aspects of model performance. Moving beyond the scope of localization, the ‘Overall evaluation’ section presents a comprehensive analysis of our language model across a spectrum of benchmarks. This includes assessing instruction-following ability, knowledge retention, and Natural Language Understanding (NLU). For evaluating instruction-following ability, we employ our
fine-tuned chat models, which are designed to excel in interactive and directive tasks. In contrast, knowledge retention and NLU are evaluated using our base models, focusing on the core strengths of the model’s pre-training. While we utilize both automated and manual methods for assessing instruction-following ability, other benchmarks in this section are evaluated solely through automated methods.

**Instruction-following capability** We specifically assess the models’ instruction-following capabilities using the Arabic versions of Vicuna-80 (Chiang et al., 2023) and AlpacaEval (Dubois et al., 2023), translated by GPT-4 and refined by native speakers. In accordance with (Chiang et al., 2023), we adopt the GPT-4 evaluation, which prompts GPT-4 to score the performance of models on each question, contrasting them with GPT-3.5 Turbo. The details can be found in Appendix F.2. While GPT-4 evaluation is efficient and scalable, it may overlook the subtle inconsistencies between model responses (Wang et al., 2023a) and human interactions in real-world scenarios. Therefore, we further conduct human evaluation on these benchmarks to evaluate the performance of AceGPT from the perspective of human rather than GPT-4 preferences. To ensure cultural relevance in manual evaluations, we engaged a diverse group of educated, native Arabic speakers. Each model’s response was assessed independently by three assessors. We present more details in Table 17 and the UI for evaluation in Figure 2.

**Knowledge benchmark** We have two knowledge benchmarks, including Arabic MMLU and EXAMs. MMLU (Hendrycks et al., 2021) consists of diverse multiple-choice questions across 57 tasks, spanning various educational levels. We employed GPT-3.5 Turbo to translate this dataset from English to Arabic. Additionally, Arabic questions from the EXAMs (Hardalov et al., 2020), a resource specialized in multilingual high school exam questions, were also incorporated. Both datasets were evaluated in a few-shot setting, as per the methodology in (Huang et al., 2023), to assess the innate capabilities of LLMs, aiming at potential applications with minimal adaptations.

ALUE 7 is a popular online benchmark, which is similar to the GLUE benchmark but has a main focus on Arabic Language Understanding Evaluation. It includes traditional NLP tasks such as sentiment analysis, semantic matching, text relation classification, and dialect identification. It comprises 9 tasks as illustrated in Table 21. More details of task illustration, experiment setting, and results can be found in Appendix G.2.

### 4.2 Experiment Results

**Instruction-following benchmark** We present each model’s performance ratio against GPT-3.5 Turbo, scored by GPT-4, in Table 6. The result shows that AceGPTs are superior in both Arabic Vicuna-80 and Arabic AlpacaEval. Notably, AceGPT-7B-chat surpasses Jais-13B by about 20% points with smaller model size. Moreover, AceGPT-13B-chat attains a 100.88% performance ratio of GPT-3.5 Turbo in Arabic Vicuna-80.

**Human evaluation** Table 7 shows the human evaluation results on Arabic Vicuna-80 and Arabic AlpacaEval. We calculated the percentages of wins, ties, and losses of the results from three Arabic speakers. We observe that AceGPT-chat (both 7B and 13B) significantly surpasses Jais-13B-chat and even Jais-30B-chat, but lags behind GPT-3.5 Turbo. Moreover, the AceGPT-13B-chat is significantly better than the AceGPT-7B-chat, indicating the importance of model size.

**Knowledge benchmark** Table 8 shows the few-shot evaluation results on Arabic MMLU and EXAMs. We can see that AceGPT-13B-base attains the best performance in Arabic MMLU (37.26%) among open-source LLMs, and AceGPT-7B-base also surpasses other open-source models, including 13B models, in Humanities and Others (Business, Health, Misc) domains in Arabic MMLU. In EXAMs, the Jias-30B-base model achieves the best performance among open-source models.

### 5 Experimental Analysis

#### 5.1 On Pre-Training

**Localization of pre-training** AceGPT-base uses LLaMA2 as the backbone, the only difference it is further pre-trained with some local Arabic texts. We compare AceGPT-base to LLaMA2 on ACVA with the few-shot setting to demonstrate the benefits of localized pre-training on Arabic culture and values. The results in Table 9 show the superiority of localized pre-training: after localized pre-training, AceGPT-7B-base surpasses LLaMA2-13B, which has a larger size.
Table 7: Human evaluations on Vicuna-80 and AlpacaEval. The winners are in **bold**.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Comparison</th>
<th>Win</th>
<th>Tie</th>
<th>Lose</th>
<th>Win or Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic Vicuna-80</td>
<td>AceGPT-7B-chat vs. Jais-13B-chat</td>
<td>82.5%</td>
<td>6.7%</td>
<td>10.8%</td>
<td>89.2%</td>
</tr>
<tr>
<td></td>
<td>AceGPT-7B-chat vs. GPT-3.5 Turbo</td>
<td>27.5%</td>
<td>32.9%</td>
<td>39.6%</td>
<td>60.4%</td>
</tr>
<tr>
<td></td>
<td>AceGPT-13B-chat vs. Jais-13B-chat</td>
<td>82.9%</td>
<td>6.7%</td>
<td>10.4%</td>
<td>89.6%</td>
</tr>
<tr>
<td></td>
<td>AceGPT-13B-chat vs. GPT-3.5 Turbo</td>
<td>16.3%</td>
<td>57.1%</td>
<td>26.6%</td>
<td>73.4%</td>
</tr>
<tr>
<td></td>
<td>AceGPT-7B-chat vs. Jais-30B-chat</td>
<td>67.5%</td>
<td>15.0%</td>
<td>17.5%</td>
<td>82.5%</td>
</tr>
<tr>
<td></td>
<td>AceGPT-13B-chat vs. Jais-30B-chat</td>
<td>64.6%</td>
<td>15.0%</td>
<td>20.4%</td>
<td>79.6%</td>
</tr>
<tr>
<td>Arabic AlpacaEval</td>
<td>AceGPT-7B-chat vs. Jais-13B-chat</td>
<td>53.0%</td>
<td>36.5%</td>
<td>10.5%</td>
<td>89.5%</td>
</tr>
<tr>
<td></td>
<td>AceGPT-7B-chat vs. GPT-3.5 Turbo</td>
<td>20.2%</td>
<td>46.5%</td>
<td>33.3%</td>
<td>66.7%</td>
</tr>
<tr>
<td></td>
<td>AceGPT-13B-chat vs. Jais-13B-chat</td>
<td>49.4%</td>
<td>42.8%</td>
<td>7.8%</td>
<td>92.2%</td>
</tr>
<tr>
<td></td>
<td>AceGPT-13B-chat vs. GPT-3.5 Turbo</td>
<td>25.2%</td>
<td>44.5%</td>
<td>30.3%</td>
<td>69.7%</td>
</tr>
</tbody>
</table>

Table 8: Accuracy on Arabic MMLU and EXAMs. The best is **bold** and the second is *underlined*.

<table>
<thead>
<tr>
<th>Model</th>
<th>Arabic MMLU</th>
<th>EXAMs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>STEM</td>
</tr>
<tr>
<td>Bloomz</td>
<td>33.69</td>
<td>33.35</td>
</tr>
<tr>
<td>LLaMA2-7B</td>
<td>29.47</td>
<td>30.30</td>
</tr>
<tr>
<td>LLaMA2-13B</td>
<td>33.76</td>
<td>32.94</td>
</tr>
<tr>
<td>Jais-13B-base</td>
<td>32.23</td>
<td>30.51</td>
</tr>
<tr>
<td>Jais-30B-base</td>
<td>36.27</td>
<td>32.67</td>
</tr>
<tr>
<td>AceGPT-7B-base</td>
<td>32.14</td>
<td>29.73</td>
</tr>
<tr>
<td>AceGPT-13B-base</td>
<td>40.45</td>
<td>36.60</td>
</tr>
<tr>
<td>GPT-3.5 Turbo</td>
<td><strong>49.07</strong></td>
<td><strong>43.38</strong></td>
</tr>
</tbody>
</table>

Table 9: Ablation of pre-training.

<table>
<thead>
<tr>
<th>Size</th>
<th>Model</th>
<th>F1 on ACVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>7B</td>
<td>LLaMA2</td>
<td>51.44%</td>
</tr>
<tr>
<td></td>
<td>AceGPT-base</td>
<td>68.28%</td>
</tr>
<tr>
<td>13B</td>
<td>LLaMA2</td>
<td>65.67%</td>
</tr>
<tr>
<td></td>
<td>AceGPT-base</td>
<td><strong>76.23%</strong></td>
</tr>
</tbody>
</table>

Table 10: Effects of different datasets on ACVA, Arabic Vicuna-80 and Arabic AlpacaEval.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>ACVA</th>
<th>Arabic Vicuna-80</th>
<th>Arabic AlpacaEval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpaca-Arabic</td>
<td>50.52%</td>
<td>87.15% ± 0.5</td>
<td>82.97% ± 0.4</td>
</tr>
<tr>
<td>+ ShareGPT</td>
<td>38.64%</td>
<td>88.01% ± 0.03</td>
<td>84.89% ± 0.3</td>
</tr>
<tr>
<td>+ Evol-Instruct</td>
<td>61.72%</td>
<td><strong>90.39% ± 0.4</strong></td>
<td><strong>86.87% ± 0.1</strong></td>
</tr>
<tr>
<td>+ Quora</td>
<td>65.53%</td>
<td>89.74% ± 0.8</td>
<td>85.71% ± 0.03</td>
</tr>
</tbody>
</table>

5.2 On Supervised Fine-Tuning

In this analysis, we primarily assess the impact of both localized and open-source instructions on localization and overall performance. Each dataset has been sampled with 40,000 data points, respectively. The results are shown in Table 10. It can be observed that Evol-Instruct highly contributes to the overall performance in the instruction-following benchmark, while Quora is most beneficial for Arabic culture and values. Note that incorporating ShareGPT largely harms the performance of ACVA; this may be because ShareGPT is almost aligned with Western culture and values.

5.3 On RLAIF

5.3.1 Reward Model

To evaluate the sensitivity of the reward model to the overall performance, we measure the correlations between reward scoring and GPT-4 scoring (described in section 4.1) on Arabic Vicuna-80. Following the pairwise comparison setting in GPT-4 scoring, we also calculate the performance ratio for normalized (to [0, 10] as GPT-4 scoring) reward scores on model-chatbot pairs. The Pearson correlation and Spearman correlation are 0.57 and 0.61 respectively, and the results are shown in Figure 1a. We conclude that the reward model shows a positive correlation with GPT-4 evaluation on Arabic Vicuna, which indicates it can offer an effective signal on overall performance.

Localization of reward model Then we evaluate the Arabic culture sensitivity of the reward model on the ACVA benchmark. Prompting with ‘Give me a fact about Arab culture, values, and laws’ in Arabic, we calculate the reward scores
Table 11: Experiments with/without RLAIF on Arabic Vicuna-80, Arabic AlpacaEval and ACV A.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Automatic Evaluation</th>
<th>Human Evaluation (vs. GPT-3.5 Turbo)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACV A</td>
<td>Arabic Vicuna-80</td>
</tr>
<tr>
<td>AceGPT-7B-chat (w/o RLAIF)</td>
<td>42.48% ± 1.3%</td>
<td>92.01% ± 0.08%</td>
</tr>
<tr>
<td>AceGPT-7B-chat</td>
<td>69.53% ± 0.2%</td>
<td>94.82% ± 0.0%</td>
</tr>
<tr>
<td>AceGPT-13B-chat (w/o RLAIF)</td>
<td>74.18% ± 1.0%</td>
<td>95.14% ± 0.0%</td>
</tr>
<tr>
<td>AceGPT-13B-chat</td>
<td>75.02% ± 0.4%</td>
<td>100.88% ± 0.1%</td>
</tr>
</tbody>
</table>

Table 12: Ratio of the frequency of "yes" to "no" of Reward.

<table>
<thead>
<tr>
<th>Reward interval</th>
<th>Ratio of Yes/No</th>
</tr>
</thead>
<tbody>
<tr>
<td>(−∞,0)</td>
<td>81.7%</td>
</tr>
<tr>
<td>(0,0.5)</td>
<td>86.4%</td>
</tr>
<tr>
<td>(0.5,1)</td>
<td>110.8%</td>
</tr>
<tr>
<td>(1,∞)</td>
<td>208.2%</td>
</tr>
</tbody>
</table>

Figure 1: (a) Correlations between the reward model scoring and GPT-4 scoring on Arabic Vicuna-80.
(b) Reward score distribution on ACV A.

of prompt-statement pairs for all statements from ACVA. The distribution of reward scores for yes/no statements is shown in Figure 1b, and the ratio of the frequency of "yes" to "no" is shown in Table 12. It demonstrates that reward scores for ‘yes’ statements are higher than ‘no’ statements overall, which suggests that our reward model has a cultural sensitivity.

5.3.2 Ablation
To empirically validate the contribution of RLAIF on overall performance and localization to our AceGPT models, we conduct ablation studies across ACVA benchmarks, Arabic Vicuna-80 and Arabic AlpacaEval, results are outlined in Table 11.

**RLAIF improves localization** RLAIF results in performance gains of 27.05% and 0.84% for AceGPT-7B and AceGPT-13B in ACV A respectively, despite not being explicitly trained for them. This suggests that RLAIF enhances alignment with Arabic culture and values. Notably, the improvement from RLAIF on the 7B model is much larger than that of 13B, partially because the 7B model is weaker and therefore has more space for improvement, while it may be in saturation in the 13B model. Another reason could be that the preference data in RLAIF are generated from AceGPT-7b and therefore the learned reward model fits better AceGPT-7b than AceGPT-13b.

**RLAIF improves instruction-following** The results show that RLAIF significantly enhances overall model performance on both Arabic Vicuna-80 and Arabic AlpacaEval, increasing AceGPT-7B’s performance by 2.81% and 2.46%, while AceGPT-13B shows an improvement of 5.74% and 4.90%, respectively. By examining the ‘win or tie’ metric, the 7B model shows an enhancement of 3.7% through RLAIF, while the 13B model shows a significant boost of 16.2%. This narrows the gap with GPT-3.5 Turbo. These enhancements across datasets underscore RLAIF’s efficacy.

6 Conclusion
AceGPT addresses the ‘localization issue’ in large language models by specifically catering to the distinct linguistic and cultural contexts of Arabic environments, leveraging further pre-training, instruction tuning, and RLAIF. It excels in multiple benchmarks including instruction-following and natural language understanding, setting a new standard among Arabic LLMs. We contribute high-quality datasets and evaluation resources, highlighting the need for localizing large language models and introducing AceGPT as a pioneering solution for Arabic linguistic and cultural adaptation.
Limitation

Our AceGPT model exhibits several limitations. Its vocabulary, is primarily focused on Arabic letters, lacking further expansion, affecting Arabic text encoding efficiency. Limited machine resources during pre-training restricted token allocation, suggesting untapped potential in Arabic content processing. We omitted reasoning/misinformation and bias testing in our evaluation, raising concerns about the model’s safety alignment and currently limiting its use to academic research rather than online deployment. Additionally, despite manual checks, the cultural dataset requires enhancement in both quality and quantity, which may affect the model’s practicality and adoption.

Ethical Statement

We mainly use public data to train our models. For the newly-collected data (e.g., Quora), we use ChatGPT to filter the questions with any ethical issues. Additionally, we have adopted RL to aligned with human values.

Acknowledgement

This work was supported by Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone Project (No.HZQSWS-KCCYB-2024016), the Internal Project Fund from Shenzhen Research Institute of Big Data under Grant J00220230010, the Shenzhen Science and Technology Program (JCYJ20220818103001002), Shenzhen Doctoral Startup Funding (RCBS2022100809330065), Tianyuan Fund for Mathematics of National Natural Science Foundation of China (NSFC) (12326608), and KAUST Baseline Research Fund.

We thank Prof. Zhi-Quan Luo and Dr. Ping Lee for their support. We extend our sincere appreciation to the dedicated KAUST graduate students whose contributions were integral to the success of our Arabic evaluations, including Lamees Alzahrani, Abdullah Amr Bawazir, Nouf Khalil Alenizi, Shatha Abdullah Alowdah, Rudaynah Maimani, Feras Khalid Alwutayd, Abdulrahman, Arwa Fallatah, Noura Alhijri, Reem Alquwayzani, and Majid Almarhoumi. We thank them for their invaluable support in this research.

References


Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators. *CoRR*, abs/2305.17926.


A Related Work

Our work mainly lies in two topics: democratization of ChatGPT to low-resource language and localization and cultural alignment.

Democratization of ChatGPT to low-resource language

Recent trends have focused on developing open-source LLMs as an alternative to the proprietary GPT-3.5 Turbo and GPT-4 models (Taori et al., 2023; Chiang et al., 2023; Conover et al., 2023; Chen et al., 2023a; Sengupta et al., 2023). These efforts have expanded to include LLMs designed for languages beyond English (Chen et al., 2023a; Sengupta et al., 2023). However, these models rely heavily on large-scale datasets translated from English in the supervised fine-tuning stage, inadvertently embedding English-centric cultural biases and values (Chen et al., 2023a; Sengupta et al., 2023). This practice risks undermining the models’ capacity to align with the cultural contexts and values specific to their target linguistic communities. To mitigate this “localization issue”, our work proposes a culture-aware framework that encompasses the pretraining, supervised fine-tuning, and reinforcement learning phases, to enhance the value and cultural alignment of AceGPT with Arabic.

Localization and cultural alignment

Cultural diversity varies significantly by region and language. Yet, pre-trained language models primarily incorporate Western cultural biases (Palta and Rudinger, 2023; Ramezani and Xu, 2023; CH-Wang et al., 2023; Naous et al., 2023). These biases are attributed to the reliance on training corpora such as C4 (Dodge et al., 2021) and Wikipedia (Callahan and Herring, 2011), which are inherently skewed towards Western culture, leading to English-centric models. Moreover, the emphasis on translating English data in low-resource language NLP research exacerbates the “localization issue”, failing to adequately represent the non-English culture and values (Conneau et al., 2018; Chen et al., 2023a; Sengupta et al., 2023). To address this, it is recommended to develop culture-aware models for non-English languages by gathering and annotating culture-specific data (Hershcovitch et al., 2022). However, assembling comprehensive datasets for low-resource languages is challenging and expensive. Previous research has either concentrated on collecting unsupervised data for pre-training (Sengupta et al., 2023) or embarked on task-specific supervised fine-tuning (Wang et al., 2023b; Conneau et al., 2018; Huang and Yang, 2023; Palta and Rudinger, 2023). In this paper, we present a novel and comprehensive culture-aware supervised fine-tuning and culture-aware reinforcement learning.

B Localization Issues

B.1 Sample Questions for Localization

The sample questions for Arabic name entity comparison in Table 1 and 2 are as follows Table 13.

<table>
<thead>
<tr>
<th>Table 13: 20 sample questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>ما هي الأدب العربي؟</td>
</tr>
<tr>
<td>ما هي اللغة العربية التي لا يتم استخدامها إلا خلال الحج؟</td>
</tr>
<tr>
<td>ما هي بعض الأقوال الدينية التي لا يمكنهم أن يفهموها؟</td>
</tr>
<tr>
<td>ما هي بعض الأقوال الدينية التي لا يمكنهم أن يفهموها؟</td>
</tr>
<tr>
<td>ما هي بعض الأقوال الدينية التي لا يمكنهم أن يفهموها؟</td>
</tr>
<tr>
<td>ما هي بعض الأقوال الدينية التي لا يمكنهم أن يفهموها؟</td>
</tr>
<tr>
<td>ما هي بعض الأقوال الدينية التي لا يمكنهم أن يفهموها؟</td>
</tr>
<tr>
<td>ما هي بعض الأقوال الدينية التي لا يمكنهم أن يفهموها؟</td>
</tr>
<tr>
<td>ما هي بعض الأقوال الدينية التي لا يمكنهم أن يفهموها؟</td>
</tr>
<tr>
<td>ما هي بعض الأقوال الدينية التي لا يمكنهم أن يفهموها؟</td>
</tr>
</tbody>
</table>

8147
B.2 Case Study

In this subsection, we analyze the performance of AceGPT by conducting a comparative analysis of its localization ability via case studies on the sampled 20 localization questions. Illustrated in Table 24, we observed a larger proportion of Arabic events in AceGPT. The first example in Table 24 aligns with the instance illustrated in Table 2. Both AceGPT and GPT-3.5 Turbo exhibit superior responses to the given query, significantly surpassing the answer provided by Jais. Specifically, AceGPT’s understanding of a ‘holy book’ is not solely confined to the Bible; it demonstrates a nuanced acknowledgment that different regions, especially Arabic, have their respective sacred texts, reflecting a broad and inclusive comprehension of diverse religious traditions. This illustrates the advanced capability of AceGPT, akin to GPT-3.5 Turbo, in response generation for Arabic-speaking areas.

The second example exemplifies the capability of AceGPT to incorporate more Arabic elements when responding to historical questions. Specifically, AceGPT allocates a significant proportion of its responses, 4 out of 10, to Arabic historical figures. In contrast, GPT-3.5 Turbo only attributes 1 out of 10 responses to Arabs, while Jais exclusively presents choices associated with Western figures. This demonstrates that AceGPT has an inclination towards Arabic culture, emphasizing its capability to offer more Arabic culture-relevant responses in an Arabic context.

C Construction of ACVA

We employ a top-down approach for the construction of the Arabic Cultural and Value Alignment benchmark. First, we gathered over 50 topic keywords (see Table 14) representing various aspects of Arabic culture, including humanity, art, science, geography, history, manners, religion, and the influence between civilizations, sourced from several books on Arabic culture and values. Then, we query GPT-3.5 Turbo to generate 8000 data based on the given topic using the prompt shown below, where topic is the placeholder for the topic.

We further sample 50% topics to verify the relevance of questions to Arabic cultures and values and the accuracy of the Yes-No labels, which were reviewed by Arabic speakers, leading to a ‘Clean set’.

D Preference Data for RLAIF

The data comprises two parts: Arabic preference data and open-source English preference data. Outputs for Arabic preference data are sampled from our fine-tuned 7B model with a temperature of 1. The open-source English preference data is incorporated to improve the generalization capability of the reward model and alleviate GPT4-preference hacking. We randomly sample 12K from three public human-annotated datasets - Anthropic helpfulness and harmlessness (Bai et al., 2022), OpenAI Summarize (Stiennon et al., 2020), and OpenAssistant Conversations (OASST1) (Köpf et al., 2023).

The core idea of preference labeling for Arabic preference data is to use a GPT-4 model with prompts as an automatic annotator to assess two responses generated by the same model for a given question. However, a significant challenge emerges as GPT-4 often shows a marked preference for the first response, around 80% of the time, with the exact percentage varying based on the specific prompt design. To counter this, we utilize an order-switch mechanism to ensure consistent preference data across two separate runs of GPT-4. In one run, two responses are placed arbitrarily, and in the other, their orders are switched. The prompt for labeling is shown below. instruction, response 1, and response 2 are the placeholders for the input instructions and the two generated responses.
Table 14: Topics for ACVA construction

<table>
<thead>
<tr>
<th>Country</th>
<th>Civilization Relation</th>
<th>Science and Humanity</th>
<th>Manners and Religion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algeria, Bahrain, Comoros</td>
<td>Influence from ancient Egypt, Influence from Byzantium</td>
<td>Arabic Astronomy, Arabic Math,</td>
<td>Arabic Ceremony, Arabic Clothing, Arabic</td>
</tr>
<tr>
<td>Egypt modern, Iraq, Jordan,</td>
<td>Influence from China, Influence from Greece, Influence</td>
<td>Arabic Medicine, Arabic Physics and</td>
<td>Culture, Arabic Food, Arabic Funeral,</td>
</tr>
<tr>
<td>Kuwait, Lebanon, Libya,</td>
<td>from Persia, Influence from Rome, Mesopotamia civilization</td>
<td>Chemistry, Arabic Literature,</td>
<td>Arabic Ornament, Arabic Wedding, Mindset,</td>
</tr>
<tr>
<td>Mauritania, Morocco, Oman,</td>
<td></td>
<td>Arabic Music, Arabic Philosophy, Arab</td>
<td>Special Expression, Daily Life, Influence</td>
</tr>
<tr>
<td>Palestine, Qatar, Saudi Arabia,</td>
<td></td>
<td>Empire, Arabic Architecture, Arabic</td>
<td>from Islam, Islam Branches and Schools,</td>
</tr>
<tr>
<td>Somalia, Sudan, Syria, Tunisia,</td>
<td></td>
<td>Art, Arabic Calligraphy, Arabic</td>
<td>Islam Education, Islamic Law System</td>
</tr>
<tr>
<td>United Arab Emirates, Yemen</td>
<td></td>
<td>Geography, Arabic History, Arabic</td>
<td></td>
</tr>
</tbody>
</table>

E Implementation of Training

E.1 Pre-Training

We employ the LLaMA2 framework for the pre-training process, capitalizing on a computational setup furnished with 24 Nvidia A100 80G GPUs. We configure the context length at 2048 tokens and adopt the AdamW optimizer, paired with a cosine learning rate scheduler. The learning rate is set at 1e-4. Given a gradient accumulation setting of 128, the total batch size amounts to 3072. Additionally, a warm-up phase is integrated, constituting 5% of the total training duration.

E.2 Supervised Fine-Tuning

We train for one epoch using a variety of datasets in Table 15. Native Arabic data like Alpaca-Arabic-GPT4 and Quora-Arabic-GPT4 are included thrice in the mixture, while datasets like ShareGPT and Alpaca-Chinese-GPT4 are once to minimize the non-Arabic data ratio, totaling 629,293 data.

Following LLaMA2, we use the following form of system prompt:

```plaintext
[INST] [SYS]

You are a helpful, respectful, and honest assistant. Always answer with the utmost assistance while being safe. Your answers should not include any harmful, unethical, racist, gender discriminatory, toxic, dangerous, or illegal content. Please ensure that your responses are not socially biased and are positive. If the question is meaningless or isn’t coherent in a realistic sense, explain the reason instead of answering something incorrectly. If you do not know the answer to a question, please refrain from sharing.

[SYS]
```

The corresponding meaning in English is:

```plaintext
You are a helpful, respectful, and honest assistant. Always answer with the utmost assistance while being safe. Your answers should not include any harmful, unethical, racist, gender discriminatory, toxic, dangerous, or illegal content. Please ensure that your responses are not socially biased and are positive. If the question is meaningless or isn’t coherent in a realistic sense, explain the reason instead of answering something incorrectly. If you do not know the answer to a question, please refrain from sharing.
```
Table 15: Instruction tuning datasets; Datasets constructed in this work are highlighted in bold.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Resources</th>
<th>Responses</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quora-Arabic-40K</td>
<td>Collected from Quora</td>
<td>GPT-4</td>
<td>43,050</td>
</tr>
<tr>
<td>Alpaca (Peng et al., 2023)</td>
<td>Self-instruct (Taori et al., 2023)</td>
<td>GPT-4</td>
<td>49,969</td>
</tr>
<tr>
<td>Alpaca-Chinese (Peng et al., 2023)</td>
<td>GPT-3.5 Turbo translated (Peng et al., 2023)</td>
<td>GPT-4</td>
<td>49,969</td>
</tr>
<tr>
<td>Alpaca-Arabic</td>
<td>GPT-4 translated from (Taori et al., 2023)</td>
<td>GPT-4</td>
<td>49,969</td>
</tr>
<tr>
<td>Code-Alpaca-Arabic</td>
<td>GPT-4 translated from (Chaudhary, 2023)</td>
<td>GPT-4</td>
<td>20,022</td>
</tr>
<tr>
<td>Evol-Instruct-Arabic</td>
<td>GPT-4 translated from (Xu et al., 2023)</td>
<td>GPT-4</td>
<td>69,997</td>
</tr>
<tr>
<td>ShareGPT</td>
<td>humans</td>
<td>GPT 3.5 Turbo</td>
<td>80,179</td>
</tr>
</tbody>
</table>

the AdamW optimizer, with each batch consisting of 128 samples. We adopt different configurations for the learning rate based on the model architecture. For AceGPT-7B, the maximum learning rate is set to $5 \times 10^{-5}$, and for AceGPT-13B, it is $1 \times 10^{-5}$. A cosine scheduler is employed for learning rate adjustment, with a warmup rate of 0.03.

**E.3 Reward Model Training**

The reward model is initialized with Ziya, an open-source 7B reward model. We use 8 Nvidia A100 80G GPUs for training. Each batch consists of 128 samples. We take two epochs with the AdamW optimizer. The maximum learning rate is set to 8e-6 and the warmup rate is set to 0.03 with cosine scheduler.

**E.4 PPO**

We implement PPO with DeepSpeed-Chat. The actor parameters are initialized with our fine-tuned models and the critic parameters are initialized with our trained 7B reward model. We sample 448 experiences with the mini-batch size of 224, which is updated in only one epoch. The maximum learning rate for the actor is set to 5e-7 while that for the critic is set to 5e-6. A cosine scheduler is used for learning rate adjustment with a warmup step of 10. We set the KL penalty as 0.01. The policy gradient loss is clipped with the threshold as 0.2 while that for the value loss is 0.3. The reward is clipped to be [-5, 5]. The gamma and lambda for the generalized advantage estimation are 1 and 0.95 respectively.

Notably, both AceGPT-7B and AceGPT-13B are trained with the 7B reward model whose preference data only comprises outputs from the 7B policy model (post-SFT).

**F Implementation of Evaluation**

**F.1 Baselines and Benchmarks**

We use the following baselines:

- **LLaMA2** (Touvron et al., 2023), developed by Meta AI, are the most popular open-source large language models ranging in scale from 7 billion to 70 billion parameters. Our AceGPT models are also built upon LLaMA2-7B and -13B. We compare our AceGPT-base models to the corresponding size of LLaMA2.

- **Bloomz** (Muennighoff et al., 2022) and **Phoenix** (Chen et al., 2023a,b). **Bloomz** is a classical family of multilingual models fine-tuned with multiple traditional NLP tasks. **Phoenix** are multilingual instruction following models using Bloomz as the backbone. We compare AceGPT-base models to Bloomz and AceGPT-chat models to Phoenix.

- **Jais** (Sengupta et al., 2023) are concurrent open-source 13B Arabic-centric LLMs, including a foundation base model and an instruction-tuned model. We compare AceGPT-base and AceGPT-chat to their base and chat models respectively.

- **GPT-3.5 Turbo** is the most popular and powerful closed-source multilingual LLM, second only to GPT-4. We compare both AceGPT-base and AceGPT-chat to it.

We use Arabic Vicuda-80, Arabic AlpacaEval, Arabic MMLU, Exams, ALUE, ACVA as our benchmarks, see Table 16.

In the scoring method for the Arabic MMLU dataset, the process begins by calculating the average score within each subcategory, this average...
represents the mean score for that particular subcategory. Subsequently, the average scores of all subcategories within a main category are aggregated and averaged again to yield the overall average score for that main category. Following this, the average scores of all main categories are collated and averaged to determine the overall average score for that category within the entire Arabic MMLU dataset. Finally, the aggregate average scores of all categories in the dataset are combined and averaged to arrive at the comprehensive average score across the entire Arabic MMLU dataset.

F.2 Evaluation on Instruction Following

We apply greedy decoding strategies for all models evaluated on Arabic Vicuna-80 and Arabic AlpacaEval benchmarks.

**GPT-4 evaluation**  The temperature of the GPT-4 evaluator is set to 0.2, with the scoring conducted independently three times. We calculate the average performance ratio of the evaluated model relative to GPT-3.5 Turbo. Specifically, for each question, GPT-4 is queried to assign scores to responses from the evaluated model and GPT-3.5 Turbo respectively. We obtain the evaluated model’s score for each question, and these scores are averaged to determine the evaluated model’s overall score. Similarly, we derive an overall score for GPT-3.5 Turbo. The performance ratio is obtained by dividing the overall score of the evaluated model by that of GPT-3.5 Turbo. Averaging the performance ratios from three separate experiments yields the final performance ratio. The evaluation prompt is shown below. `question`, `response 1`, and `response 2` are the placeholders for the input questions and the two generated responses. To mitigate position bias (Wang et al., 2023a), we vary the order in which the model pair is presented in the prompt.

```plaintext
[Question]
question

[Assistant 1]
response 1

[End of Assistant 1]

[Assistant 2]
response 2

[End of Assistant 2]

[System]

We would like to request your feedback on the two AI assistants in response to the user question displayed above.

Requirements: The response should be helpful, relevant, accurate, and have enough details. The response should follow Arabic culture, customs, beliefs, and laws.

Please rate the performance of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better performance.

Please first output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. You should consider which response is more in line with the given requirements.

In the subsequent line, please provide a comprehensive explanation of your evaluation.

---

**Human evaluation**  We recruited 11 native people for annotation, including verification of the localization dataset, calibration of translation results, and human evaluation, the backgrounds of these people can be found in Table 17. The evaluation interface is illustrated in Figure 2. The results of the human evaluation corresponding to Table 7 for each annotator are shown in Table 23.
Table 16: Evaluation Benchmarks.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Evaluation Aspects</th>
<th>Type of Evaluation</th>
<th>Size</th>
<th>Types of Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic Vicuna-80</td>
<td>Instruction following</td>
<td>Human &amp; Automated</td>
<td>80</td>
<td>Freely-answered questions</td>
</tr>
<tr>
<td>Arabic AlpacaEval</td>
<td></td>
<td></td>
<td>805</td>
<td></td>
</tr>
<tr>
<td>Arabic MMLU EXAMs</td>
<td>Knowledge ability</td>
<td>Automated</td>
<td>14k</td>
<td>Multiple-choice questions</td>
</tr>
<tr>
<td>ALUE (see Appendix G.2)</td>
<td>Language understanding</td>
<td>Automated</td>
<td>18k</td>
<td>Classification &amp; regression</td>
</tr>
<tr>
<td>ACVA-all</td>
<td>Arabic cultural and value alignment</td>
<td>Automated</td>
<td>9k</td>
<td>Yes/No binary questions</td>
</tr>
<tr>
<td>ACVA-clean</td>
<td></td>
<td></td>
<td>2.4k</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: User interface of human annotation. Response positions are randomized to alleviate biases.

Table 17: Information of participants involved in the AceGPT testing

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>Education</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant 1</td>
<td>female</td>
<td>PGDip</td>
<td>Arabic-Native</td>
</tr>
<tr>
<td>Participant 2</td>
<td>male</td>
<td>PhD</td>
<td>Arabic-Native</td>
</tr>
<tr>
<td>Participant 3</td>
<td>female</td>
<td>PGDip</td>
<td>Arabic-Native</td>
</tr>
<tr>
<td>Participant 4</td>
<td>female</td>
<td>PGDip</td>
<td>Arabic-Native</td>
</tr>
<tr>
<td>Participant 5</td>
<td>female</td>
<td>PGDip</td>
<td>Arabic-Native</td>
</tr>
<tr>
<td>Participant 6</td>
<td>male</td>
<td>PGDip</td>
<td>Arabic-Native</td>
</tr>
<tr>
<td>Participant 7</td>
<td>male</td>
<td>Master</td>
<td>Arabic-Native</td>
</tr>
<tr>
<td>Participant 8</td>
<td>female</td>
<td>PGDip</td>
<td>Arabic-Native</td>
</tr>
<tr>
<td>Participant 9</td>
<td>female</td>
<td>PGDip</td>
<td>Arabic-Native</td>
</tr>
<tr>
<td>Participant 10</td>
<td>female</td>
<td>PGDip</td>
<td>Arabic-Native</td>
</tr>
<tr>
<td>Participant 11</td>
<td>male</td>
<td>PhD</td>
<td>Arabic-Native</td>
</tr>
</tbody>
</table>

F.3 Evaluation on Knowledge

There are two main differences in the MMLU evaluation between (Sengupta et al., 2023) and ours: (1) we translate MMLU into Arabic differently. The machine-translated version in (Sengupta et al., 2023) is facilitated through their in-house translation model while we leverage GPT-3.5 Turbo. Additionally, (Sengupta et al., 2023) further creates a human-translated version. Unfortunately, both the human-translated and machine-translated versions are not publicly available, which prevents us from evaluating on the same benchmark; (2) we adopt the widely accepted few-shot prompting setting commonly found in related literature for base models, while (Sengupta et al., 2023) opts the zero-shot setting. Due to these differences in translation methods and evaluation settings, the performance metrics between the two works are not directly comparable.

We benchmark Jais-13B-base and Jais-30B-base using our GPT-3.5 Turbo-translated MMLU dataset under the standard few-shot setting in Table 8. Moreover, we also benchmark Jais-13B-chat using the zero-shot setting in Table 20.

The evaluating template is shown below:
The corresponding meaning in English is:

- **Few-shot**

Below are multiple choice questions (with answers) about [category]

[question]

**Question:** [question]

**Answer:**

- **Zero-shot**

Below are multiple choice questions about [category]

[question]

Please choose one answer from among ‘A, B, C, D’ without explanation.

**A specific example of five-shot prompting is:**

 فيما بيتي ستة الاختيار من متعدد (مع الإجابات) حول طب جامعي

سؤال: كيف يتم نقل الجلوكوز إلى خلايا العضلات؟

عبر ناقل البروتين المسما جولد. [A]

فقط في وجوء الأساليب. [B]

عبر الكيبوكسيتر. [C]

عبر ناقل حمض الموركازين. [D]

**إجابة:**

**A specific example of zero-shot prompting is:**

 فيما بيتي ستة الاختيار من متعدد حول علاج السرطان

A. 10 مجمًا

B. 20 مجمًا

C. 22 مجمًا

D. 25 مجمًا

من ضمن اختيارة إجابة واحدة من بين

A, B, C, D.
F.4 Evaluation on ACVA

The evaluation prompt for ACVA is

- Few-shot
  - In a set of N items or not (with the same set of items).

- Zero-shot
  - If the judgment in the previous trial is correct or not.

The corresponding meaning in English is:

- Few-shot
  - Below are multiple choice questions (with answers) about [category].
  - Question: [question]
  - Answer:

- Zero-shot
  - Please determine whether the following sentence is true or not. If it is true, please respond with 'Yes'. If it is not true, please respond with 'No'.

  Question: [question]

A specific example of five-shot prompting is:

- In a set of N items or not (with the same set of items).
  - Question: [question]
  - Answer:

A specific example of zero-shot prompting is:

- If the judgment in the previous trial is correct or not.

Question: [question]

G More Experiments of AceGPT Evaluation

G.1 Supplementary Experimental Results

ACVA evaluation under the few-shot setting.

Table 18 demonstrates the performance of base models on ACVA. AceGPT-30B-base outperforms Jais-30B-base by 4.81% in ‘All set’, but fails slightly 1.83% behind it in ‘Clean set’.

Table 18: Average F1-score on ACVA in the few shot setting. The best performance is in bold and the second best is underlined.

<table>
<thead>
<tr>
<th>Model</th>
<th>All set</th>
<th>Clean set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bloomz</td>
<td>58.94%</td>
<td>60.91%</td>
</tr>
<tr>
<td>Jais-13B-base</td>
<td>73.96%</td>
<td>75.80%</td>
</tr>
<tr>
<td>Jais-30B-base</td>
<td>73.81%</td>
<td>77.44%</td>
</tr>
<tr>
<td>AceGPT-7B-base</td>
<td>74.72%</td>
<td>70.32%</td>
</tr>
<tr>
<td>AceGPT-13B-base</td>
<td>78.62%</td>
<td>75.61%</td>
</tr>
<tr>
<td>GPT-3.5 Turbo</td>
<td>80.12%</td>
<td>81.99%</td>
</tr>
</tbody>
</table>

Knowledge evaluation on the chat models. We evaluate chat models in the zero-shot setting on Arabic MMLU and EXAMs. As illustrated in Table 20, GPT-3.5 Turbo consistently outperforms other models in both MMLU and EXAMs benchmarks. Notably, Jais-13B-chat showcases superior performance, which is consistent with the results in (Sengupta et al., 2023). Specifically, its MMLU score stands at 41.06, trailing ChatGPT’s score of 49.07 by a mere 8.01 points. On the EXAMs benchmark, Jais-13B-chat scored only 4.79 points lower than GPT-3.5 Turbo. One possible reason for
Jais’s good performance may be attributed to traditional NLP task datasets in their SFT dataset such as Super-NaturalInstructions (Wang et al., 2022), which contains multiple-choice questions akin to the MMLU and EXAMs. Our model, in contrast, hasn’t been trained on such data.

G.2 Evaluation on Arabic NLU Tasks

ALUE  Details of the evaluation of ALUE will be included in this section, the introduction of ALUE can be found in Sec 4.1.

Experiment setting  We train our AceGPT-13B-base on each task independently in a fully supervised manner, resembling the approach of the top models on the leaderboard. Moreover, high-ranking models on the leaderboard adopt the grid search method on validation sets to select hyperparameters. Similarly, we employ a Bayesian approach for hyperparameter adjustment. For tasks providing predefined validation split, we utilize the given validation sets. Otherwise, we allocate 10% of the data from the training set for validation purposes. For the DIAG task, which does not provide training data, we use the model trained on XNLI to evaluate it.

Experiment results and analysis  Table 22 presents our performance on the ALUE benchmark. AceGPT ranks second in terms of the average score in these nine datasets, right behind AraMUS (Alghamdi et al., 2023), which has conducted extensive pre-training in Arabic data. In future endeavors, we plan to incorporate a richer set of Arabic pre-training corpora and supervised data to enhance the model’s NLU capabilities.

H Relationship between Arabic Culture and Arabic Language

The English language can reflect multiple cultures and vice versa (Hershcovich et al., 2022). This demonstrates that language and culture, although intertwined, are not inherently synonymous. In contrast, among Arabs, there exists a significant correlation between the Arabic language and Arab culture for several reasons. Primarily, Arabic is predominantly spoken by native speakers deeply embedded in Arab culture, in stark contrast to English, which is more commonly spoken by second-language speakers. Furthermore, most second-language learners of Arabic pursue the language to connect with Islamic culture, a pivotal element of Arab culture, because all Islamic terminologies are expressed in Arabic. Therefore, in this paper, we assume that the Arabic language effectively captures Arab culture. Despite the presence of diverse sub-cultures within the Arab world, our paper primarily focuses on the Arab community as a whole, overlooking the distinctions among its sub-groups.

12https://en.wikipedia.org/wiki/Arabic
Table 19: ACV A examples

| id: Arabic Clothing-1, label: False | The agal is a type of traditional shoe in the Arab world. |
| id: Arabic Medicine-2, label: False | Arabic medicine is not based on the principles and knowledge of Greek medicine. |
| id: Arabic Calligraphy-3, label: True | The Diwani calligraphy is considered one of the most famous styles in Arabic calligraphy. |
| id: Arabic Math-4, label: True | Algebra is a branch of mathematics in which the Arabs founded. |
| id: Arabic Funeral-5, label: True | In an Arab funeral, the body is buried facing east and facing Mecca and Medina. |

Table 20: Accuracy of chat models on Arabic MMLU and EXAMs. The best is in **bold** and the second is *underlined*.

<table>
<thead>
<tr>
<th>Model</th>
<th>Arabic MMLU</th>
<th>EXAMs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>STEM</td>
</tr>
<tr>
<td>Phoenix</td>
<td>29.65</td>
<td>27.06</td>
</tr>
<tr>
<td>Phoenix-multiple-langs</td>
<td>17.37</td>
<td>16.77</td>
</tr>
<tr>
<td>Jais-13B-chat</td>
<td>41.06</td>
<td>39.82</td>
</tr>
<tr>
<td>AceGPT-7B-chat</td>
<td>30.43</td>
<td>26.58</td>
</tr>
<tr>
<td>AceGPT-13B-chat</td>
<td>35.17</td>
<td>33.14</td>
</tr>
<tr>
<td>GPT-3.5 Turbo</td>
<td><strong>49.07</strong></td>
<td><strong>43.38</strong></td>
</tr>
</tbody>
</table>

Table 21: Summary of NLU tasks and metrics in ALUE benchmark

<table>
<thead>
<tr>
<th>Task</th>
<th>Metric</th>
<th>Size</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>MQ2Q (NSURL-2019 Shared Task 8)</td>
<td>F1-score</td>
<td>4000</td>
<td>private</td>
</tr>
<tr>
<td>OOLD (OSACT4 Shared Task-A)</td>
<td>F1-score</td>
<td>1000</td>
<td>private</td>
</tr>
<tr>
<td>OHSD (OSACT4 Shared Task-B)</td>
<td>F1-score</td>
<td>1000</td>
<td>private</td>
</tr>
<tr>
<td>SVREG (SemEval-2018 Task 1)</td>
<td>Pearson correlation</td>
<td>1000</td>
<td>private</td>
</tr>
<tr>
<td>SEC (SemEval-2018 Task 1)</td>
<td>Jaccard similarity score</td>
<td>1000</td>
<td>private</td>
</tr>
<tr>
<td>FID (IDAT@FIRE2019)</td>
<td>F1-score</td>
<td>1006</td>
<td>public</td>
</tr>
<tr>
<td>MDD (MADAR Shared Task Subtask 1)</td>
<td>F1-score</td>
<td>5200</td>
<td>public</td>
</tr>
<tr>
<td>XNLI (Cross-lingual Sentence Representations)</td>
<td>Accuracy</td>
<td>2490</td>
<td>public</td>
</tr>
<tr>
<td>DIAG (Diagnostic dataset)</td>
<td>Matthews correlation</td>
<td>1147</td>
<td>public</td>
</tr>
</tbody>
</table>
Table 22: Experimental results in ALUE (Seelawi et al., 2021) including online baselines. While the leaderboard calculates the ‘scores’ excluding Task DIAG, we also incorporate it to derive the ‘Avg’.

<table>
<thead>
<tr>
<th>Model</th>
<th>#Params</th>
<th>Avg Score</th>
<th>MQ2Q</th>
<th>MDD</th>
<th>SVREG</th>
<th>SEC</th>
<th>OOLD</th>
<th>XNLI</th>
<th>OHSD</th>
<th>DIAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARABIC-BERT</td>
<td>135M</td>
<td>63.5</td>
<td>67.1</td>
<td>85.7</td>
<td>59.7</td>
<td>55.1</td>
<td>25.1</td>
<td>82.2</td>
<td>89.5</td>
<td>61.0</td>
</tr>
<tr>
<td>ARABERTv0.1-base</td>
<td>135M</td>
<td>64.2</td>
<td>68.4</td>
<td>89.2</td>
<td>58.9</td>
<td>56.3</td>
<td>24.5</td>
<td>85.5</td>
<td>88.9</td>
<td>67.4</td>
</tr>
<tr>
<td>ARABIC-BERT</td>
<td>110M</td>
<td>68.5</td>
<td>69.3</td>
<td>89.7</td>
<td>59.7</td>
<td>58.0</td>
<td>26.5</td>
<td>84.3</td>
<td>89.1</td>
<td>67.0</td>
</tr>
<tr>
<td>CAMeL-BERT-MIX</td>
<td>108M</td>
<td>66.7</td>
<td>70.4</td>
<td>89.4</td>
<td>61.3</td>
<td>69.5</td>
<td>30.3</td>
<td>85.5</td>
<td>90.3</td>
<td>56.1</td>
</tr>
<tr>
<td>AraT5-base</td>
<td>289M</td>
<td>67.6</td>
<td>71.1</td>
<td>91.3</td>
<td>63.8</td>
<td>65.9</td>
<td>30.5</td>
<td>82.3</td>
<td>88.8</td>
<td>68.2</td>
</tr>
<tr>
<td>ARBERT</td>
<td>163M</td>
<td>65.5</td>
<td>71.4</td>
<td>89.3</td>
<td>61.2</td>
<td>66.8</td>
<td>30.3</td>
<td>85.4</td>
<td>89.5</td>
<td>70.7</td>
</tr>
<tr>
<td>MARBERT</td>
<td>163M</td>
<td>63.9</td>
<td>72.2</td>
<td>83.3</td>
<td>61.9</td>
<td>75.9</td>
<td>36.0</td>
<td>85.3</td>
<td>92.1</td>
<td>64.3</td>
</tr>
<tr>
<td>JABER</td>
<td>135M</td>
<td>68.2</td>
<td>73.7</td>
<td>93.1</td>
<td>64.1</td>
<td>70.9</td>
<td>31.7</td>
<td>85.3</td>
<td>91.4</td>
<td>73.4</td>
</tr>
<tr>
<td>Char-JABER</td>
<td>136M</td>
<td>70.1</td>
<td>75.3</td>
<td>92.0</td>
<td>66.1</td>
<td>74.5</td>
<td>34.7</td>
<td>86.0</td>
<td>92.3</td>
<td>73.1</td>
</tr>
<tr>
<td>ALM-1.0</td>
<td>350M</td>
<td>70.3</td>
<td>75.8</td>
<td>94.5</td>
<td>65.1</td>
<td>70.1</td>
<td>35.3</td>
<td>86.0</td>
<td>91.7</td>
<td>77.7</td>
</tr>
<tr>
<td>SABER</td>
<td>369M</td>
<td>71.4</td>
<td>77.3</td>
<td>93.3</td>
<td>66.5</td>
<td>79.2</td>
<td>38.8</td>
<td>86.5</td>
<td>93.4</td>
<td>76.3</td>
</tr>
<tr>
<td>MARBERT</td>
<td>163M</td>
<td>63.9</td>
<td>72.2</td>
<td>83.3</td>
<td>61.9</td>
<td>75.9</td>
<td>36.0</td>
<td>85.3</td>
<td>92.1</td>
<td>64.3</td>
</tr>
<tr>
<td>JABER</td>
<td>135M</td>
<td>68.2</td>
<td>73.7</td>
<td>93.1</td>
<td>64.1</td>
<td>70.9</td>
<td>31.7</td>
<td>85.3</td>
<td>91.4</td>
<td>73.4</td>
</tr>
<tr>
<td>Char-JABER</td>
<td>136M</td>
<td>70.1</td>
<td>75.3</td>
<td>92.0</td>
<td>66.1</td>
<td>74.5</td>
<td>34.7</td>
<td>86.0</td>
<td>92.3</td>
<td>73.1</td>
</tr>
<tr>
<td>ALM-1.0</td>
<td>350M</td>
<td>70.3</td>
<td>75.8</td>
<td>94.5</td>
<td>65.1</td>
<td>70.1</td>
<td>35.3</td>
<td>86.0</td>
<td>91.7</td>
<td>77.7</td>
</tr>
<tr>
<td>SABER</td>
<td>369M</td>
<td>71.4</td>
<td>77.3</td>
<td>93.3</td>
<td>66.5</td>
<td>79.2</td>
<td>38.8</td>
<td>86.5</td>
<td>93.4</td>
<td>76.3</td>
</tr>
</tbody>
</table>

Table 23: Details of human evaluations on Arabic Vicuna-80 and Arabic AlpacaEval.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Comparison</th>
<th>Win</th>
<th>Tie</th>
<th>Lose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic Vicuna-80</td>
<td>AceGPT-7B-chat vs. Jais-13B-chat</td>
<td>66</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>volunteer 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>volunteer 2</td>
<td>65</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>volunteer 3</td>
<td>67</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>AceGPT-7B-chat vs. GPT-3.5 Turbo</td>
<td>26</td>
<td>0</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>volunteer 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>volunteer 2</td>
<td>40</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>volunteer 3</td>
<td>0</td>
<td>79</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>AceGPT-7B-chat (w/o RLAIF) vs. GPT-3.5 Turbo</td>
<td>23</td>
<td>12</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>volunteer 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>volunteer 2</td>
<td>12</td>
<td>58</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>volunteer 3</td>
<td>31</td>
<td>0</td>
<td>49</td>
</tr>
<tr>
<td>Arabic AlpacaEval</td>
<td>AceGPT-13B-chat vs. Jais-13B-chat</td>
<td>68</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>volunteer 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>volunteer 2</td>
<td>65</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>volunteer 3</td>
<td>66</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>AceGPT-13B-chat vs. GPT-3.5 Turbo</td>
<td>14</td>
<td>35</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>volunteer 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>volunteer 2</td>
<td>21</td>
<td>28</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>volunteer 3</td>
<td>4</td>
<td>74</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>AceGPT-13B-chat (w/o RLAIF) vs. GPT-3.5 Turbo</td>
<td>19</td>
<td>14</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>volunteer 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>volunteer 2</td>
<td>22</td>
<td>19</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>volunteer 3</td>
<td>6</td>
<td>57</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>AceGPT-7B-chat vs. Jais-30B-chat</td>
<td>54</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>volunteer 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>volunteer 2</td>
<td>48</td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>volunteer 3</td>
<td>60</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>AceGPT-13B-chat vs. Jais-30B-chat</td>
<td>515</td>
<td>196</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>volunteer 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>volunteer 2</td>
<td>619</td>
<td>54</td>
<td>132</td>
</tr>
<tr>
<td></td>
<td>volunteer 3</td>
<td>146</td>
<td>632</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>AceGPT-7B-chat vs. Jais-13B-chat</td>
<td>267</td>
<td>291</td>
<td>255</td>
</tr>
<tr>
<td></td>
<td>volunteer 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>volunteer 2</td>
<td>71</td>
<td>632</td>
<td>102</td>
</tr>
<tr>
<td></td>
<td>volunteer 3</td>
<td>158</td>
<td>200</td>
<td>447</td>
</tr>
<tr>
<td></td>
<td>AceGPT-13B-chat vs. GPT-3.5 Turbo</td>
<td>283</td>
<td>504</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>volunteer 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>volunteer 2</td>
<td>400</td>
<td>343</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>volunteer 3</td>
<td>509</td>
<td>187</td>
<td>109</td>
</tr>
<tr>
<td></td>
<td>AceGPT-13B-chat vs. GPT-3.5 Turbo</td>
<td>216</td>
<td>326</td>
<td>263</td>
</tr>
<tr>
<td></td>
<td>volunteer 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>volunteer 2</td>
<td>47</td>
<td>664</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>volunteer 3</td>
<td>346</td>
<td>84</td>
<td>375</td>
</tr>
</tbody>
</table>
(What are the holy books, saints, and holy places?)

AceGPT-7B-chat:
The holy book is the bible, which is the holy book of Christians and Jews. It is divided into two main parts: the old testament and the new testament.

The Bible is a collection of religious texts that are considered sacred by Christians. Saints are holy people who are considered to have lived exceptionally virtuous lives and are venerated as saints by some religions. Holy places are sites considered sacred or consecrated by some religions, such as churches and temples and mosques.

Jais-13B-chat:
The holy book is the quran, which is the holy book of Muslims. It is written in Arabic and contains the words of god as revealed to the prophet Muhammad. Muslims are believers who are guided by the teachings of the Quran and follow the Prophet Muhammad's examples.

GPT-3.5 Turbo:
The holy book is the Torah, which is the holy book of Jews. It contains the Five Books of Moses and is divided into three main sections: the Pentateuch, the Prophets, and the Writings.
There are many famous writers and poets in history, but unfortunately, it is difficult to provide a comprehensive list of all of them due to the diversity of literature and poetry in the world. But we can mention some Well-known names such as 1. William Shakespeare: A famous English poet and playwright known for his immortal works such as “Romeo and Juliet” and “Macbeth”. 2. Gabriel García Márquez: A famous Colombian writer, novelist, journalist, and political activist He is known for his novels “One Hundred Years of Solitude” and “Love in Times of Cholera”. 3. John Milton: A famous English poet and philosopher known for his poem “Paradise Lost”. 4. Fructius Mahler: Writer And a famous French philosopher is known for his novel “The White Doha”. 5. Naguib Mahfouz: A famous Egyptian novelist and short story writer known for his novels “Between the Two Palaces” and “The Palace of Longing”. 6. Leo Tolstoy: A famous Russian novelist and short story writer known for his novels “War and Peace” and “Anna Karenina”. 7. Jean-Paul Sartre: A famous French writer and philosopher known for his philosophical and fictional works such as “Loneliness”. 8 Gibran Khalil Gibran: A famous Lebanese poet, philosopher, and critic known for his poetic and prose works such as “The Prophet”. 9. Muhammad Hassan Ajami: A famous Egyptian poet known for his novels “Love in the Time of Cholera” and “Black Suits You.” 10. Thomas Mann: A famous German poet and writer known for his novels “The Blind Leg” and “The Old Houses”. These are just some examples, and there are many other famous writers and poets in history.)
There are many famous writers and poets in history, among them: 1. William Shakespeare: a famous English writer, who is considered the greatest writer in the English language and one of the greatest playwrights in history. 2. Goethe: a German poet and writer, who is considered one of the most prominent poets of world literature. 3. Homer: An ancient Greek poet, and is considered the author of the classic literary works “The Iliad” and “The Odyssey”. 4. Dante Alighieri: An Italian poet and writer of the “Divine Comedy”. 5. Charles Baudelaire: A French poet, and is considered one of the most prominent poets of the Romantic movement in the nineteenth century. 6. William Wordsworth: An English poet, and is considered one of the most prominent poets of the Romantic movement in the nineteenth century. 7. John Milton: An English poet, considered one of the greatest poets of English literature in the seventeenth century. 8. Robert Frost: An American poet, considered one of the most prominent poets of the twentieth century. 9. George Bernard Shaw: An Irish writer and playwright, and is considered one of the greatest playwrights of the twentieth century. 10. Naguib Mahfouz: An Egyptian writer and winner of the Nobel Prize in Literature, and is considered one of the most prominent writers of Arabic literature in the twentieth century.)