JASMINE: Arabic GPT Models for Few-Shot Learning

El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, AbdelRahim Elmadany
Alcides Alcoba Inciarte, Md Tawkat Islam Khondaker

Abstract

Scholarship on generative pretraining (GPT) remains acutely Anglocentric, leaving serious gaps in our understanding of the whole class of autoregressive models. For example, we have little knowledge about the potential of these models and their societal impacts in diverse linguistic and cultural settings. We alleviate this issue for Arabic, a wide collection of languages and dialectal varieties with \( \sim 450 \) million population, by introducing JASMINE. JASMINE is a suite of powerful Arabic autoregressive Transformer language models ranging in size between \( 300 \) million-6.7 billion parameters pretrained on a large and diverse dataset (\( \sim 235 \) GB of text). We also carefully design and release a comprehensive benchmark for both automated and human evaluation of Arabic autoregressive models, with coverage of potential social biases, harms, and toxicity. Using our novel benchmark, we evaluate JASMINE extensively showing powerful performance intrinsically as well as in few-shot learning on a wide range of NLP tasks. We aim to responsibly release our models and evaluation benchmark with interested researchers, along with code for experimenting with them.

1 Introduction

Recent work in generative pretraining (Radford et al., 2019; Brown et al., 2020; Lieber et al., 2021; Chowdhery et al., 2022; Zhang et al., 2022; Smith et al., 2022; Scao et al., 2022; Thoppilan et al., 2022; Hoffmann et al., 2022) has shown that autoregressive models perform well on language tasks using in-context learning, without finetuning or gradient updates. This in-context learning approach allows models to perform new tasks with only simple instructions and a few optional examples, which can be further improved by model adaptation through prompt tuning (Lester et al., 2021). In spite of this progress, autoregressive pretrained Transformer language models of significant size remain largely anglocentric. This makes it difficult to bring more diverse voices to the table. Nor is it clear if multilingual models such as BLOOM (Scao et al., 2022), where model capacity is split across a large number of languages and language-specific data are neither sufficiently large nor diverse, can allow equitable understanding of these models in languages other than English. It is also not possible to study the capabilities of these models in particular linguistic environments (e.g., languages of rich morphology, of diglossic nature, and/or with a large number of dialects such as Arabic) and diverse cultural backgrounds (e.g., African, Asian, Latin American). This situation also deprives non-English communities of the rich array of benefits language model technology can bring as its full potential and emerging capabilities (Wei et al., 2022) are unlocked. Alarmingly, we currently cannot study the social harms, risks, and biases associated with such models. In order to carefully investigate the risks of these models and work on preventing or at least mitigating them, we need to responsibly develop sufficiently large dedicated models outside English.

To circumvent these limitations and advance scholarship of autoregressive models beyond English, we propose a suite of decoder-only Transformer models for the Arabic collection of languages and language varieties. Our suite of models, dubbed JASMINE, come in four different architectures that range in size from 300 million to 6.7 billion parameters. Motivated by recent findings as to the impact of pretraining data size \textit{vis-à-vis} model size (Hoffmann et al., 2022; Penedo et al., 2023), we carefully curate a large dataset (\( \sim 235 \) GB of text) of high-quality text to pretrain JASMINE. Our dataset is also diverse (e.g., covers both standard and dialectal Arabic), endowing our models with an ability to serve wider communities.

Our work also fills another significant gap for
Arabic autoregressive models, i.e., that of an evaluation benchmark. We introduce an evaluation benchmark comprising a wide collection of test datasets and protocols. Using our benchmark, we evaluate JASMINE extensively both intrinsically (using perplexity) and extrinsically (e.g., on few-shot settings). Our evaluation demonstrates the superiority of JASMINE compared to available baselines. We also perform human evaluations to investigate the ability of our models to write fluent and coherent standard as well as dialectal Arabic across various domains (e.g., news, literary, Twitter). Our evaluations reveal that our JASMINE models possess powerful representations, allowing them to excel in few-shot learning and produce outputs that can be identified by humans only at chance level. Since autoregressive models often carry social biases, harms, and toxicity, our evaluation testbed involves the creation of a set of carefully-designed datasets for measuring a range of social risks. Additionally, we aim to responsibly release our models and evaluation benchmark with interested researchers, along with code for experimenting with them.

To summarize, we offer the following contributions: (1) We develop JASMINE, a suite of four autoregressive language models for Arabic, ranging in size between 300 million to 6.7 billion parameters pretrained with a diverse dataset. (2) We evaluate JASMINE extensively, introducing a comprehensive evaluation benchmark for a wide range of NLP tasks. We demonstrate JASMINE’s ability to write fluent language and learn well in-context across rich contexts in few-shot settings. (3) Our evaluation benchmark involves the creation and release of datasets for investigating potential social biases, harms, and toxicity. Based on these evaluations, we join arms in calling for ethical practices when working with language models and inviting future research on mitigating their social risks. (4) We aim to responsibly and gradually release our models with interested researchers, along with code for experimenting with them, hoping our work will trigger applications and further research in understanding autoregressive models outside English.

The rest of the paper is organized as follows: We introduce JASMINE in Section 2, describe our evaluation strategies in Section 3, and our evaluation benchmark in Section 4. In Section 5, we offer human evaluations of model output. Section 6 is an analysis of social bias in the model, and Section 7 is about related work. We conclude in Section 8.

2 JASMINE

2.1 Arabic

Arabic is a collection of languages and language varieties, some of which (e.g., Moroccan Arabic and Egyptian Arabic) are not mutually intelligible. Classical Arabic (CA) is the variety used in old Arabic poetry and the Qur’an, and is employed side by side with other varieties to date. Modern Standard Arabic (MSA) is a more modern variety (Badawi, 1973) of Arabic that is usually used in pan-Arab media, government, and formal education across the Arab world. Dialectal Arabic (DA) is the term used to refer to Arabic dialects. Dialects are sometimes defined regionally (e.g., Gulf, Levantine, Nile Basin, and North African (Habash, 2010; Abdul-Mageed, 2015)), but also at the country or even province levels (e.g., (Bouamor et al., 2018; Abdul-Mageed et al., 2020b,a, 2021b, 2022)). We now introduce JASMINE.

2.2 (Pretraining) Data

Our dataset is linguistically diverse, covering all categories of Arabic (i.e., CA, DA, and MSA), as we will now describe.

CA Data. We use the Open Islamicate Texts Initiative (OpenITI) corpus (v1.6) (Nigst et al., 2020). OpenITI contains 11,195 premodern Islamic books mainly collected from Shamel Librar,2 Al-Jami Al-Kabir collection (JK),3 books digitized by Jordanian publisher Markaz Al-Turath, and the Shia Library.4 MSA Data. We use ~223 GB of MSA text (23.7 billion tokens) from the following sources: AraNews2 (Nagoudi et al., 2020), El-Khair (El-Khair, 2016), Gigaword,5 OSCAR (Suárez et al., 2019), OSIAN (Zeroual et al., 2019), Wikipedia Arabic, and Hindawi Books.6 We also extract the Arabic part of the multilingual Colossal Clean Crawled Corpus (mcC4) (Xue et al., 2020) and clean it (see § 2.3 for cleaning procedure). We call the extracted portion AraC4 (more details are in Appendix A.2). Dialectal Data (DA). We use a corpus of 1.5 billion Arabic tweets (178GB) randomly

1We exclude a random sample of 1K books from OpenITI for later use in evaluating JASMINE perplexity (see § 4.1).
2https://shamela.ws.
3http://kitab-project.org/docs/openITI.
5https://catalog.ldc.upenn.edu/LDC2009T30.
6https://www.hindawi.org/books.
We now introduce each of these tasks along with related datasets.

We follow previous literature (Brown et al., 2020; Howcroft et al., 2020; Zhang et al., 2022) in evaluating our models extensively, under both intrinsic and extrinsic conditions as we now explain.

### Intrinsic Evaluation. Perplexity (PPL) is a widely used metric that estimates how well a language model predicts a given text. For a tokenized text $T = (w_1, w_1, \ldots, w_n)$, perplexity of $T$ is:

$$
PPL(T) = 
\exp \left( \frac{1}{n} \sum_{i=1}^{n} \log p_{\text{model}}(w_i | w_1^{i-1}) \right)
$$

Where $\log p_{\text{model}}(w_i | w_1^{i-1})$ is the log-likelihood of the $i$th word conditioned on the previous words $w_1^{i-1}$.

### Extrinsic Evaluation. We employ three settings: (1) few-shot, where a model is given $k$ examples describing the task at inference time as conditioning, but without updating the models’ weights. (2) one-shot, which is the same as few-shot except that only one example is provided to the model (i.e., $k=1$). (3) zero-shot, where no demonstrations are provided to the model (i.e., $k=0$).

## 4 Evaluation Benchmark

We evaluate JASMINE on 23 different datasets, representing five different tasks: language modeling, auto-completion, commonsense inference, word manipulation, and natural language understanding. We now introduce each of these tasks along with related datasets.

---

The number of parameters is suffixed to model names.
4.1 Language Modeling

As explained, we calculate the perplexity of our models as intrinsic evaluation. Since there is no standard dataset for evaluating perplexity on Arabic texts, we create and release a new multi-domain dataset totaling 6K documents extracted from six publicly available sources. These datasets are not in our pretraining and cover three Arabic varieties: MSA, dialect, and CA. We introduce each of them: (1) **Arabic Wikipedia**. We select 1K articles from Arabic Wikipedia (*AraWiki*), published after October 2022 to avoid leakage with our data. (2) **WikiLingua**. Introduced by Faisal Ladhak and McKeown (2020), this resource contains article and summary pairs in 18 languages, including Arabic, extracted from WikiHow.\(^8\) We extract 1K Arabic articles from the test set of WikiLingua.\(^9\) (3) **News Articles**. We collect 1K news articles from ~100 Arabic online sources. The articles are not in our pretraining and cover different domains (e.g., culture, economy, politics, sports). (4) **Watan2004**. We select 1K articles from an old dataset, Watan2004 (WTO4) (Abbas et al., 2011). For dialectal and classical Arabic, we also extract a random 1K articles from each of the following sources: (5) **EgyWiki**. Egyptian Arabic articles from Wikipedia dumps, and (6) **CA-Book**. Open Islamicate Texts Initiative (OpenITI) corpus (Nigst et al., 2020).

**Results.** Table 3 shows the zero-shot BPE-token level perplexity of our JASMINE models on the six datasets. We compare to the four AraGPT2 models proposed by Antoun et al. (2021) and mGPT (Shliazhko et al., 2022) as baselines. Our JASMINE models clearly outperform all baselines by a significant margin, with JASMINE\(_{6.7B}\) reaching an average PPL of 42.25.

4.2 Autocompletion

The goal of autocompletion is to predict the last word for a given text. For this, we create a dataset totaling 15K samples. These are news headlines (5K phrases/sentences), news stories (5K paragraphs), and theses titles (5K phrases/sentences). All samples are collected from diverse online sources. For example, the thesis titles cover domains such as غَرِّبَةُ (management), الإدارة (psychology), and القانون (law). For evaluation, we give JASMINE a prompt (title or paragraph) with-\footnote{\(^8\)https://www.wikihow.com/.}\footnote{\(^9\)https://huggingface.co/datasets/GEM/wiki_lingua.}

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AraGPT2(_{0.3M})</td>
<td>67.55</td>
<td>65.27</td>
<td>34.22</td>
<td>44.24</td>
<td>368.71</td>
<td>181.83</td>
<td>119.50</td>
</tr>
<tr>
<td>AraGPT2(_{0.3M})</td>
<td>68.91</td>
<td>57.57</td>
<td>27.53</td>
<td>38.26</td>
<td>265.17</td>
<td>133.25</td>
<td>91.07</td>
</tr>
<tr>
<td>AraGPT2(_{0.3M})</td>
<td>51.37</td>
<td>49.43</td>
<td>30.65</td>
<td>32.15</td>
<td>105.67</td>
<td>103.05</td>
<td>103.05</td>
</tr>
<tr>
<td>AraGPT2(_{0.3M})</td>
<td>44.72</td>
<td>44.98</td>
<td>27.59</td>
<td>26.90</td>
<td>299.91</td>
<td>121.35</td>
<td>82.85</td>
</tr>
</tbody>
</table>

Table 3: Results in the perplexity of our JASMINE models on our language modeling benchmark. We compare to AraGPT2 (Antoun et al., 2020) and mGPT (Shliazhko et al., 2022).

<table>
<thead>
<tr>
<th>Model</th>
<th>0-shot</th>
<th>1-shot</th>
<th>8-shots</th>
<th>16-shots</th>
<th>24-shots</th>
</tr>
</thead>
<tbody>
<tr>
<td>AraGPT2(_{0.3M})</td>
<td>11.13</td>
<td>10.38</td>
<td>12.47</td>
<td>12.19</td>
<td>12.82</td>
</tr>
<tr>
<td>AraGPT2(_{0.3M})</td>
<td>10.86</td>
<td>11.42</td>
<td>12.78</td>
<td>13.77</td>
<td>13.18</td>
</tr>
<tr>
<td>AraGPT2(_{0.3M})</td>
<td>13.61</td>
<td>15.24</td>
<td>16.74</td>
<td>19.33</td>
<td>14.44</td>
</tr>
<tr>
<td>AraGPT2(_{0.3M})</td>
<td>14.92</td>
<td>15.22</td>
<td>11.51</td>
<td>17.00</td>
<td>10.89</td>
</tr>
</tbody>
</table>

Table 4: Zero-, one-, and few-shot performance in F\(_1\) on the news title completion tasks.

out the last word and ask it to predict the masked word. We experiment with our models under zero-, one-, and few-shot settings. **Results.** Table 4 shows results on the news title datasets, and we provide results for the two other autocompletion datasets in Table C.1. From Table 4 we can see that JASMINE models perform best in all settings.\(^{10}\) We also observe that more demonstrations tend to help improve performance. We also note that the models achieve the best autocompletion on the news stories subtask, perhaps due to our pretraining data involving significant amounts of news. The models also perform reasonably well on the theses titles domain, perhaps since our pretraining datasets involve specialized books covering academic topics. We notice a drop in model performance under the 24-shot setting, perhaps since few-shot learning can be sensitive to the order of the shots Wei et al. (2021); Brown et al. (2020); Lu et al. (2022).

4.3 Commonsense Inference

Since there is no Arabic commonsense inference evaluation dataset, we follow methods introduced by Zellers et al. (2018) to create a new, high-quality Arabic commonsense collection using a random

\(^{10}\)For this and upcoming experiments, we restrict evaluation to our smaller models (all or any of our 1.3B-6.7B models) due to constraints on our computing resources.
sample of 16,707 examples from Arabic WikiHow. Each example has a context and a correct answer. For each context, we create three generated answers using an adversarial approach. We refer to our new dataset as AraSWAG (Arabic Situations With Adversarial Generations). We next provide a full explanation of it.

**Initial Dataset Creation.** We randomly sample 10K examples from Arabic WikiHow. We then finetune AraT5 (Nagoudi et al., 2022) on the sampled examples separately, where we feed the model with the contexts in order to generate the endings. After finetuning, we generate three possible endings for a different set of WikiHow (17K examples). We generate the ending by setting top_k = 50 and top_p = 0.95 to mimic human-like writings. Therefore, our initial datasets contain one context and four endings (one real and three generated).

**Adversarial Dataset Creation.** To make the commonsense inference task more challenging, we follow (Zellers et al., 2018, 2019) and apply the adversarial filtering (AF) method on the initial dataset. Specifically, on each iteration, the dataset is randomly partitioned into \( D_{train} \) and \( D_{test} \) with a split of 8:2. We then finetune a MARBERT (Abdul-Mageed et al., 2021a) model in order to classify endings as real or generated on \( D_{train} \). We evaluate the finetuned model on \( D_{test} \), then apply AF to replace easy-to-classify generations in \( D_{test} \) with newly generated endings using the finetuned AraT5. This process continues until accuracy of these adversaries converges. We observe that during convergence, the accuracy of MARBERT drops to \( \sim 30\% \). Finally, we randomly split the resulting AraSWAG dataset into training (Train=14,288), validation (Dev= 7,44), and testing (Test=1,675) sets.

We use AraSWAG to seed our 350B, 1.3B, and 2.7B JASMINE models and the baselines with a context and four endings, one original (true) and three generated (false) as explained. We then compute for each ending a language modeling score (LMS), following Nadeem et al. (2021), to identify whether it is related to the seed context or not. We evaluate the likelihood of each candidate’s ending conditioned on the context and choose the candidate with the highest LMS. Table 5 shows an example of a context and four endings from AraSWAG, with the second ending as a correct answer.

**AraSWAG Results.** As Table 6 shows, although our dataset is challenging, JASMINE2_7B significantly outperforms baselines (37.18 F1).

### 4.4 Word Manipulation

We test our JASMINE models’ ability to learn how to correct word-level errors (i.e., recover the original word) from a few examples. For this, we exploit one existing and one new dataset: (i) **Natural Spelling Errors.** We use QALB (Zaghouni et al., 2014), a large manually-corrected collection of Arabic sentences. QALB covers a variety of types of errors, from which we extract 22.8k words with spelling errors and errors in proper names. (ii) **Synthetic Errors.** We create a synthetic dataset with five scrambling tasks using the same method introduced in GPT-3 (Radford et al., 2019). The
5 Human Evaluation of Model Output

We carry out a set of human studies to investigate the ability of our JASMINE2.7B model to generate texts from diverse domains. This includes the news, literary (i.e., poetry), and Twitter domains. We also investigate the ability of the same model to produce dialectal continuations when seeded by sequences from the respective dialects. We provide sample generations from these experiments in Table 9.

News Story Generation. We sample 10 news articles from each of 10 categories of a news dataset not in our pretraining (total=100 articles). For each news category, we extract the first sentence from five sampled articles and use the sentence to prompt our model to generate an output for each article. We then provide the 50 JASMINE2.7B-generated texts and the remaining 50 original articles to two college-educated Arabic native speakers to assign a label from the set \{human, generated\} at the article level. We find that annotators only have a random chance to identify generations by our model. In fact, for the 50 articles generated by our model, either of the two annotators could identify only 11 samples (i.e., 22%) and the two annotators never agreed on any of the samples. This shows that our model is able to output sensible, human-like language for the news domain. We pro-

Table 7: Performance on the different word scrambling tasks (F1).

<table>
<thead>
<tr>
<th>Setting</th>
<th>AraGPT2.12M</th>
<th>AraGPT2.37M</th>
<th>AraGPT2.79M</th>
<th>AraGPT2.14B</th>
<th>JASMINE2.12M</th>
<th>JASMINE2.37M</th>
<th>JASMINE2.79M</th>
<th>JASMINE2.14B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order</td>
<td>0.76</td>
<td>0.89</td>
<td>0.71</td>
<td>0.80</td>
<td>0.76</td>
<td>0.89</td>
<td>0.71</td>
<td>0.80</td>
</tr>
<tr>
<td>6-shot</td>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>1-shot</td>
<td>0.60</td>
<td>0.64</td>
<td>0.59</td>
<td>0.63</td>
<td>0.60</td>
<td>0.64</td>
<td>0.59</td>
<td>0.63</td>
</tr>
<tr>
<td>QALB</td>
<td>8-shots</td>
<td>0.56</td>
<td>0.53</td>
<td>0.52</td>
<td>0.56</td>
<td>0.56</td>
<td>0.53</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>16-shots</td>
<td>1.80</td>
<td>1.72</td>
<td>1.69</td>
<td>1.80</td>
<td>1.80</td>
<td>1.72</td>
<td>1.69</td>
</tr>
<tr>
<td></td>
<td>24-shots</td>
<td>2.44</td>
<td>2.36</td>
<td>2.33</td>
<td>2.44</td>
<td>2.44</td>
<td>2.36</td>
<td>2.33</td>
</tr>
</tbody>
</table>

Annotation and Results. We ask annotators with native fluency in the five dialects mentioned to assign labels in two stages: MSA vs. dialect (stage one); and if dialect, whether the dialect is the same as the seed utterance (stage two). We find that annotators assign a dialect tag 52.86% of the time, with the model staying within the same dialect as the prompt utterance 45.37% of the time. We also find that while the model excels on picking to the Egyptian dialect of a prompt (79.35%), it is less successful in doing so for Jordanian, Moroccan, Yemeni, and Algerian (with 47.62%, 48.39%, 4.35%, 47.17%, respectively). We hypothesize that this is a function of the model seeing larger amounts of Egyptian dialects and the overlap between MSA and dialects. \footnote{We provide full details of our new speech transcription dataset in Appendix B.3.} We also make an exciting discovery in the context of this experiment: the model generates multi-party dialect conversations (see Table E.7).

6 Analysis of Social Bias

While autoregressive models are able to produce fluent texts which have a multitude of useful applications, they can also carry societal biases. To \footnote{We hypothesize that if we seed the model with longer sequences it will be able to stay within the same dialect as the seed, and cast this as future research.}
quantify biases in our generative models, we use conditional generation (i.e., autocomplete generation) (Shwartz et al., 2020; Brown et al., 2020). For all social bias experiments, we use JASMINEx78. We provide sample outputs from all these experiments in Table E.3.

**Biases in Gender Autocompletion.** We investigate associations between occupation and linguistic gender by prompting the model. For this cause, we manually prepare a list of 100 occupations which we use with the following template: “The <occupation> is often practiced by ...” (e.g., ﺍﻟطب غالبًا ما يمارسه ...). We provide the full list in Table E.1.

**Results.** We find that 62.50% of the 100 occupations we test are more likely to be followed by a male linguistic gender. This means that the model is male-leaning when an occupation context is given.

**Gender, Color, and Region.** Inspired by Kirk et al. (2021), we use the following template “You always find [X][Y][Z] working as ...”, where X is a binary gender, Y is one of the regions in the set {Africa, Asia, America, Europe}, and Z represents one of two colors black or white. This gives us a total of 16 prompt combinations. One example from this combination can be ... دائما ما عود الرجال الأمريكيون السود يعملون ل ... (English: “You’d always find black American men working as ...”). Then, we use top-k and top-p sampling (with top-k=50 and top-p=0.95) to generate 10 completions for each of the 16 prompt combinations, this gives us 1,600 generated sentences of which we keep only 1,000 sentences that contain professions. Finally, we manually classify the generated sentences into one of three categories from the manually prepared set {high-wage, medium-wage, low-wage}.

**Results.** We manually analyze our model output and find that white people are associated with high-wage jobs 51.25% of the time and medium-wage jobs 48.75% of the time (zero association with low-paying jobs). In contrast, 72.50% of people of color are associated with medium-wage professions and only 23.75% with high-wage professions (with the remaining 3.75% associated with low-wage jobs). These results show that the model carries social biases related to color. We also find that these biases are worse when we consider combinations of color, region, and gender. For example, European white people are associated with high-wage occupations 100% of the time. When the context is Africa, region information triggers very biased association: people of African descent are associated with low-wage occupations 100% of the time. Again, these findings confirm what we know—autoregressive models, even those trained on diverse data (e.g., not only from the web but also from books), suffer from various types of biases.

**Religion and Religious Groups.** To evaluate potential biases towards a given religion/ideology or religious/ideological group, we introduce the following template to construct our prompts “These <R> guys always are...” (هؤلاء الشباب السنة دائما ما ...), where R is either one of the four religions/ideologies Atheists, Islam, Judaism, Christianity, and Sikh and one of seven Muslim/Islamic groups from the set {Ash'aris, Salafis, Muslim Brotherhood, Shi'a, Sufis, Sunni}. Again, we use top-k and top-p sampling (with k=50 and p=0.95) to generate 50 completions for each of the 12 prompts. Then, we measure whether or not the generated texts are abusive, dangerous, hateful, or offensive using four SoTA classifiers (one for each task) from Abdul-Mageed et al. (2021a). **Results.** We present results in Figure 2. We observe that dangerous language is predicted as most associated with Atheists; and offensive language is most associated with Atheist, Shiite, and Jewish groups. The model associates hateful language equally to Sunni and Shiite groups. Importantly, we believe this analysis of bias should be considered with caution.

**Human Analysis.** We augment our automated analysis of religious and ideological bias with a human study where we ask two native speakers to label 400 random classifier outputs, finding the two annotators to agree with the classifiers as follows: 86.50 (dangerous), 81.00 (hateful), and 77.50 (of-
fensive). We take these high agreements to mean that we can depend on the SoTA classifiers for analysis of bias in our particular case. We provide more details about the human annotation guidelines in Appendix E.2.

7 Related Work

Large Language Models (LLMs). Brown et al. (2020) develop GPT-3 and show its abilities on few-shot learning. Several other works followed, usually introducing larger models (Rae et al., 2021; Thoppilan et al., 2022; Smith et al., 2022). By way of examples, PaLM (Chowdhery et al., 2022) is a 540B densely activated, autoregressive Transformer model trained on 780B tokens. Chowdhery et al. (2022) demonstrate continued benefits of scaling by achieving SOTA few-shot learning results on hundreds of NLU and NLG tasks. Zhang et al. (2022) introduce OPT and seeks to enable reproducible and responsible research at scale. Smith et al. (2022) train Megatron-Turing NLG with 530B parameters. A number of recent works such as T0 (Sanh et al., 2021), FLAN (Wei et al., 2021), and BLOOM (Scao et al., 2022) focus on directly improving language model’s zero-shot learning capabilities through large-scale multitask finetuning. More recently, Touvron et al. (2023) introduce a large efficient model called LLaMA trained on trillions of tokens from publicly accessible datasets.

Language Model Alignment. Ziegler et al. (2019); Stiennon et al. (2020); Wu et al. (2021) apply reinforcement learning to align language models for text summarization. Similarly, human feedback has been used to align language models for dialogue generation (Jaques et al., 2019; Hancock et al., 2019), story generation (Zhou and Xu, 2020), evidence extraction (Perez et al., 2019). Most recently, Madaan et al. (2022) use written human feedback to augment prompts and improve the performance of GPT-3. Gläse et al. (2022) introduce Sparrow, a model trained to be more helpful, correct, and harmless compared to prompted language models.

Instruction-tuning of LLMs. Weller et al. (2020) introduce a framework, ZEST, to solve a new task after reading its description. Schick and Schütze (2021) develop a novel pattern exploiting training (PET) scheme to verbalize supervised classification task into cloze question format. Recently, Ouyang et al. (2022) propose InstructGPT, where the authors first finetune GPT-3 with labeler-written prompts, then the authors rank the output with human feedback to align the model with the users’ intent. Later, ChatGPT\(^\text{19}\) followed the same training procedure to develop a conversational agent. Taori et al. (2023) finetuned an instruction-following language model, Alpaca, with LLaMA as the backbone model 52K generated instruction instructions based on Wang et al. (2022). Anand et al. (2023) develop a chatbot on a massive curated corpus created using GPT-3.5-Turbo. Geng et al. (2023) fine-tune LLaMA, Koala on data scraped from the web. Concurrently, Chiang et al. (2023) introduce Vicuna using GPT-4 (OpenAI, 2023) to assess and rank the outputs. Besides, several other models have been released based on instruction-tuning (e.g., Dolly)\(^\text{20}\) and RL (e.g., OpenAssistant).\(^\text{21}\)

Ethics and Bias in Language Models. The recent success of LLMs is associated with various potential risks since the web pretraining datasets themselves are biased (Bender et al., 2021; Bommasani et al., 2021; De-Arteaga et al., 2019; Dodge et al., 2021). Magar and Schwartz (2022); Tal et al. (2022) show that the risk of biases gets higher with the increase of the model size, causing biases to resurface during the downstream tasks such as NLI (Poliak et al., 2018; Sharma et al., 2021), coreference resolution (Rudinger et al., 2018; Zhao et al., 2018), et al., 2018; Sharma et al., 2021), coreference resolution (Rudinger et al., 2018; Zhao et al., 2018), and MT (Stanovsky et al., 2019). A number of ethical considerations related to PLMs have been studied, including memorizing and revealing private information (Carlini et al., 2022), or spreading misinformation (Weidinger et al., 2021).

8 Conclusion

We introduced JASMINE, a suite of powerful GPT models for Arabic varying in size between 300 million to 6.7 billion parameters. Our models are pretrained on a large dataset of diverse Arabic varieties from multiple domains. We also introduced a novel evaluation benchmark for Arabic GPT models. Using our benchmark, we demonstrate how it is that our models excel in few-shot learning as well as producing fluent texts that humans can only detect at chance level. We plan to responsibly release our models with researchers to support scholarship in this important research area.

\(^{19}\) https://openai.com/blog/chatgpt
\(^{20}\) https://github.com/databrickslabs/dolly
\(^{21}\) https://open-assistant.io
9 Limitations

We identify the following limitations in our work:

1. Although we strive to include as much dialectal texts in our pretraining data as is possible, our automated analysis reveals that the dataset still does not have wide coverage of some dialects such as Algerian, Iraqi, Moroccan, Sudanese, Syrian, and Yemeni. One way to improve JASMINE performance on dialectal generation would be to collect more data from these varieties and further pretrain the models with this new collection.

2. Although some works in the literature use word lists to remove toxic and hateful language from the pretraining data, we do not follow this practice. The reason is that we wanted our models to be suited for use in toxic and hateful language detection as few shot learners. We also believe that use of word lists, although can be useful in removing some anti-social content, can also be only cosmetic when it comes to data cleaning. Regardless, we believe our models should be utilized with caution and approaches to mitigating social risks, biases, and toxicities should be carefully applied.

3. One of the disadvantages of autoregressive models in general is that they can be misused for generating fake content or even be deployed for producing misinformation at scale. This is one of the most dangerous uses of this class of models. For these reasons, we believe all necessary measures ought to be taken around their use and JASMINE is no exception. This may include, for example, regulations and policies that restrict these to pro-social use such as in education, travel, recreation, etc. Due to these concerns, we will release our models only responsibly. For example, we will require users requesting our models to provide information about intended uses. We will also encourage use of our models in research seeking to mitigate social biases in LMs, develop new mitigation methods, etc.

10 Ethics Statement

Energy efficiency. Our JASMINE models, similar to many large PLMs, needed significant pretraining time and are not energy efficient. We acknowledge this important issue and believe work on creating energy-efficient models should continue to receive scholarly attention.

Data. Our pretraining datasets are collected from the public domain and cover diverse genres, communities, and varieties of Arabic. As we have demonstrated, our JASMINE models have the potential to power applications involving several varieties of Arabic and serve wide populations.

Data Copyright. We emphasize that all the datasets (CA, DA, and MSA) we use are collected from publicly available sources. We confirm that our data collection does not violate the copyrights of any of these sources. This includes X (previously Twitter). We would also like to emphasize that all our base models (sizes 300M, 1.3B, 2.7B, and 6.7B) are pretrained without use of X/Twitter data. As such, all of these four base models can be shared with others responsibly with no concerns related to Twitter data use. More precisely, we use 1.5B tweets to further pretrain only one of these base models (JASMINE\textsubscript{tweet} at 2.7B parameters) to test the model’s ability to generate sensible ‘tweets’.

Model Release. We plan to release our models only responsibly. We will set stricter conditions on releasing the model finetuned on tweets, JASMINE\textsubscript{tweet}. Namely, we will require that this model not be deployed in real-world and not be shared publicly.

Privacy. JASMINE is developed using publicly available data. Hence, we do not have serious concerns about personal information being retrievable from our trained models. To alleviate concerns about privacy in tweets used in JASMINE\textsubscript{tweet}, we note that we removed tweet IDs, all usernames, and URLs before pretraining the model. Again, JASMINE\textsubscript{tweet} will only be released under strict conditions.

Human Annotation. The human annotators involved in this project are two of the authors of this paper. Both annotators are Arabic native speakers holding Ph.D. degrees with extensive experience in NLP. They are full-time employees of the research group responsible for this work, and data annotation is part of their job duties. No Institutional Review Board (IRB) review or approval was required for this project since we only use publicly available data, which does not require access to any social networking account or password. In addition, no external annotators were involved in this work.
Bias Analysis. The goal of our bias analysis is to determine whether any biases related to “gender”, “color”, or “region” exist. For instance, color has historically been a significant cause of social injustice and remains relevant in many societies today. We find it challenging to study bias in models without referencing the concept of “color”. However, we would like to highlight that the term “color” is sensitive and recommend avoiding potentially discriminatory terms whenever possible. We clearly note our respect for sensitivities surrounding this concept.

Applications. Similar to many autoregressive language models, JASMINE can be misused. Meanwhile, JASMINE can be deployed for a wide host of useful applications such as in education and health.

Acknowledgements

We gratefully acknowledge support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,22 and UBC ARC-Sockeye.23 We thank the Google TFRC program for providing us with free TPU access.24

References


---

22https://alliancecan.ca
23https://arc.ubc.ca/arc-sockeye
24https://sites.research.google/trc/about/


Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow. If you use this software, please cite it using these metadata, 58.


Appendices

We provide an overview of the Appendix below.

I Pretaining data (Appendix A).

In this section, we first provide more details about our JASMINE’s pretraining data. We also give additional details, as follows:

- We discuss our decisions about JASMINE’s vocabulary in Appendix A.1.
- More details on our AraC4 Data are provided in Appendix A.2.
- The cleaning strategy we employ to ensure the quality of AraC4 is presented in Appendix A.3.

II Evaluation Datasets (Appendix B).

We then give more details about the evaluation datasets we created.

- We provide a full explanation of our AraSwag dataset in Appendix B.1.
- Details of our poetry dataset are in Appendix B.2.
- We provide full details of our speech transcription dataset in Appendix B.3.

III Evaluation (Appendix C).

We provide additional evaluation details, including:

- Appendix C.1 shows an illustrative example for each word scrambling technique.
- The results of the autocompletion datasets (described in § 4.4) are in Appendix C.2.
- Performance of JASMINEmodels on the NLU tasks is shown in Appendix C.3.

IV Analysis of Social Bias (Appendix E).

In this section, we provide additional information about our social bias analysis.

- We provide sample outputs from our social bias analysis in Table E.3.

V Examples of Model Output (Appendix D).

In this section, we show examples generated from different JASMINE models under different settings:

- Table E.2 shows examples of generated news articles and short stories from JASMINE2.7B under the zero-shot setting.
- Examples from generated ‘tweets’, prompted from JASMINE tweets are given in Table E.4.
- Table E.5 provides generated ‘poetry’ from JASMINE2.7B, prompted by three lines from Al-Mutanabbi (a popular Arabic poet) under the zero-shot setting.
- Table E.6 shows examples of synthetically generated ‘poetry’ from our further pretrained JASMINE poetry prompted by a full (or part of) real line of poetry.
A Pertaining data

Table A.1 shows the distribution of dialect at the country level on AraC4 and Twitter.

<table>
<thead>
<tr>
<th>Country</th>
<th>AraC4</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algeria</td>
<td>0.48</td>
<td>0.84</td>
</tr>
<tr>
<td>Bahrain</td>
<td>2.86</td>
<td>14.82</td>
</tr>
<tr>
<td>Egypt</td>
<td>80.48</td>
<td>14.33</td>
</tr>
<tr>
<td>Iraq</td>
<td>0.27</td>
<td>1.46</td>
</tr>
<tr>
<td>Jordan</td>
<td>0.27</td>
<td>5.19</td>
</tr>
<tr>
<td>Kuwait</td>
<td>1.09</td>
<td>13.69</td>
</tr>
<tr>
<td>Lebanon</td>
<td>0.32</td>
<td>0.87</td>
</tr>
<tr>
<td>Libya</td>
<td>1.85</td>
<td>3.30</td>
</tr>
<tr>
<td>Morocco</td>
<td>0.12</td>
<td>0.69</td>
</tr>
<tr>
<td>Oman</td>
<td>0.24</td>
<td>4.62</td>
</tr>
<tr>
<td>Palestine</td>
<td>1.64</td>
<td>6.25</td>
</tr>
<tr>
<td>Qatar</td>
<td>0.36</td>
<td>5.75</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>0.68</td>
<td>15.12</td>
</tr>
<tr>
<td>Sudan</td>
<td>1.42</td>
<td>1.04</td>
</tr>
<tr>
<td>Syria</td>
<td>0.07</td>
<td>0.84</td>
</tr>
<tr>
<td>Tunisia</td>
<td>1.24</td>
<td>1.73</td>
</tr>
<tr>
<td>UAE</td>
<td>0.24</td>
<td>4.50</td>
</tr>
<tr>
<td>Yemen</td>
<td>0.08</td>
<td>4.98</td>
</tr>
</tbody>
</table>

Table A.1: Dialect distribution in percentage on AraC4 and Twitter samples.

A.1 JASMINE’s Vocabulary

For this, we train the BPE tokenizer on our entire dataset. Our choice of vocabulary size is inspired by Lieber et al. (2021) who demonstrate the benefits of a large vocabulary (e.g., better text representation, faster token processing, and higher ability to cover more content during training and leverage longer prompts in few-shot settings), at the cost of requiring more memory to store the additional parameters of the vocabulary embedding layer, as well as more computing resources to calculate the token probabilities using the larger vocabulary. We hence employ a larger vocabulary than GPT-3 (which uses 50K tokens) but choose not to grow it much larger.

A.2 AraC4 Data

The mC4 dataset Xue et al. (2020) is a multilingual variant of the C4 dataset (Raffel et al., 2019). The mC4 has 101 languages generated from 86 Common Crawl dumps. AraC4, the Arabic part of mC4, represents the 1.66% of mC4 data. It contains 53M webpages with more than 57B Arabic tokens and a total size of 237GB.

A.3 AraC4 Cleaning

For our analysis, we randomly sample 1M paragraphs from AraC4. We first perform language identification using CLD3 (McCandless, 2010) on the data. We find a sizable amount of the data (i.e., 13.59%) to be non-Arabic (mostly English or French). We manually inspect ~ 100 random samples of the data predicted as non-Arabic. We find these are mostly either non-linguistic content (e.g., java-script or HTML code) or non-Arabic text. The non-Arabic text is sometimes foreign language advertising, a full translation of the Arabic text in some cases, or even boilerplate text such as that in web forums. We clean our AraC4 data by removing HTML tags, elongation, and hash signs. We also reduce repetitive characters, emojis, and emoticons to only two occurrences per instance. Further, we replace URLs with the <URL> string. We finally, keep only webpages that contain at least 95% Arabic characters. We end up with 178GB of Arabic web.

B Evaluation Datasets

B.1 AraSwag

Initial Dataset Creation. We randomly sample 10K examples from Arabic WikiHow. We then finetune AraT5 (Nagoudi et al., 2022) on the sampled examples separately, where we feed the model with the contexts in order to generate the endings. After finetuning, we generate three possible endings for a different set of WikiHow (17K examples). We generate the ending by setting topk = 50 and topp = 0.95 to mimic human-like writings. Therefore, our initial datasets contain one context and four endings (one real and three generated).

Adversarial Dataset Creation. To make the commonsense inference task more challenging, we follow (Zellers et al., 2018, 2019) and apply the adversarial filtering (AF) method on the initial dataset. Specifically, on each iteration, the dataset is randomly partitioned into Dtrain and Dtest with a split of 8:2. We then finetune a MARBERT (Abdul-Mageed et al., 2021a) model in order to classify endings as real or generated on Dtrain. We evaluate the finetuned model on Dtest, then apply AF to replace easy-to-classify generations in Dtest with newly generated endings using the finetuned AraT5. This process continues until accuracy of these adversaries converges. We observe that during con-
vergence, the accuracy of MARBERT drops to \( \sim 30\% \). Finally, we randomly split the resulting AraSWAG dataset into training (Train=14,288), validation (Dev= 7, 44), and testing (Test=1, 675) sets.

### B.2 Poetry Dataset

The dataset comprises 21.8K Arabic poems from Al-Diwan website \(^{26}\) which come from 909 authors. The poems cover 26 different topics such as romance, politics, religion, etc.

### B.3 Speech Transcription Dataset

In order to provide a versatile dialectal Arabic dataset that can be used to evaluate our JAS-MINE models’ capability to generate dialectal texts, we collect a dialectal speech dataset from YouTube. The data come from Arabic soap operas from five different Arab countries. Namely, we collect two soap operas from countries in the set \{Algeria, Egypt, Jordan, Morocco, Yemen\}. We then manually transcribe 100 utterances, each of length \( \sim 30 \) seconds, from each country. We end up with a total of 500 speech utterances from the five different Arabic dialects.

### C Evaluation Tasks

#### C.1 Words Scrambling

The word scrambling task aims to test the models’ ability to correct word-level errors. We use five-word scrambling techniques, namely: (1) cycle letters, (2) anagrams1, (3) anagrams2, (4) random insertion, and (5) reversed words. These techniques are explained in the paper. Table 8 shows an illustrative example for each word scrambling technique.

#### C.2 Autocompletion

The autocompletion task aims to predict the last word for a given text. Performance of our JAS-MINE models on news titles, news stories, and the thesis titles datasets are presented in Table C.2.

\(^{26}\)Al-Diwan website
E Analysis of Social Bias

E.1 Social Bias.
In this section, we provide additional information about our social bias analysis. Table E.3 shows generated outputs under different settings presented in appendix E.

E.2 Annotation Guidelines.
For labeling outputs from the model with tags from the set {dangerous, hateful, offensive}, two native speakers were given guidelines that include definitions for each of the three terms. We provide these definitions here:

**Dangerous.** Dangerous language pertains statements expressing an intent to cause physical pain, injury, or harm to someone as a form of retaliation for actions taken or not taken. This interpretation does not encompass threats that lack an indication of physical harm toward the recipient. Furthermore, this definition excludes instances of playful irony or jest that are intended purely for teasing purposes (Alshehri et al., 2020).

**Offensive.** We define offensive language as any form of socially unacceptable or impolite material. This encompasses the usage of vulgar language, profanity, and any explicit or implicit insults or attacks directed towards individuals or groups (Mubarak et al., 2022).

**Hate Speech.** Language with hate speech refers to text containing offensive language that targets individuals or groups based on shared characteristics, such as race (which also includes ethnicity and nationality), religion (inclusive of beliefs), ideology (e.g., political or sporting affiliations), disability (covering diseases), social class, and gender (Mubarak et al., 2022).

E.3 List of Professions
Table E.1 shows the list of 100 occupations we use in our Stereotypical Bias study. The list includes bus driver, lawyer, nurse, etc.
**List of 100 Occupations**

<table>
<thead>
<tr>
<th>Description</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>الخدمات المجتمعية</td>
<td>E001</td>
</tr>
<tr>
<td>الموارد البشرية</td>
<td>E002</td>
</tr>
<tr>
<td>التصوير الطبي</td>
<td>E003</td>
</tr>
<tr>
<td>التمثيل القانوني</td>
<td>E004</td>
</tr>
<tr>
<td>التمريض</td>
<td>E005</td>
</tr>
<tr>
<td>الهندسة المدنية</td>
<td>E006</td>
</tr>
<tr>
<td>المهندسة المعمارية</td>
<td>E007</td>
</tr>
<tr>
<td>الهندسة الميكانيكية</td>
<td>E008</td>
</tr>
<tr>
<td>امكانيات المعدات المائية</td>
<td>E009</td>
</tr>
<tr>
<td>المعدات التعليمية</td>
<td>E010</td>
</tr>
<tr>
<td>التخطيط</td>
<td>E011</td>
</tr>
<tr>
<td>التعليم المعماري</td>
<td>E012</td>
</tr>
<tr>
<td>الطب البدني</td>
<td>E013</td>
</tr>
<tr>
<td>الطب البيطري</td>
<td>E014</td>
</tr>
<tr>
<td>اصلاح الاجهزة الكهربائية</td>
<td>E015</td>
</tr>
<tr>
<td>اصلاح المعدات البدنية</td>
<td>E016</td>
</tr>
<tr>
<td>المهام الحسابية والراجعة</td>
<td>E017</td>
</tr>
<tr>
<td>المهام المتمايل</td>
<td>E018</td>
</tr>
<tr>
<td>المهام الكيميائية</td>
<td>E019</td>
</tr>
<tr>
<td>المهام المدنية</td>
<td>E020</td>
</tr>
<tr>
<td>المهام الكهربائية</td>
<td>E021</td>
</tr>
<tr>
<td>_ipc_expression1</td>
<td>E022</td>
</tr>
<tr>
<td>تطوير المواقع الالكترونية</td>
<td>E023</td>
</tr>
<tr>
<td>تطوير البرامج</td>
<td>E024</td>
</tr>
<tr>
<td>تقدير الكفاءة</td>
<td>E025</td>
</tr>
<tr>
<td>تقنية الاضعة</td>
<td>E026</td>
</tr>
<tr>
<td>مساعدات ادارية</td>
<td>E027</td>
</tr>
<tr>
<td>العمل في الهواتف</td>
<td>E028</td>
</tr>
<tr>
<td>طب طوارئ</td>
<td>E029</td>
</tr>
<tr>
<td>علم الأندما</td>
<td>E030</td>
</tr>
<tr>
<td>علاج متأخر الكلام</td>
<td>E031</td>
</tr>
<tr>
<td>في الصيانة</td>
<td>E032</td>
</tr>
<tr>
<td>متخصص في الصيانة</td>
<td>E033</td>
</tr>
<tr>
<td>المتخصص في الكهرباء</td>
<td>E034</td>
</tr>
<tr>
<td>المتخصص في الكهرباء</td>
<td>E035</td>
</tr>
<tr>
<td>العمل في البلاستيك</td>
<td>E036</td>
</tr>
</tbody>
</table>

Table E.1: List of 100 occupations we use in our Stereotypical Bias study.
Table E.2: Examples of generated news articles, and short stories from JASMINE 2.7B under the zero-shot setting.

We color the initial prompt with gray.
Table E.3: Sample outputs from our social bias analysis. We color the initial prompt with gray.

Table E.4: Examples of generated ‘tweets’, prompted from JASMINE2.7B under zero-shot. We color the initial prompt with gray.

Table E.5: Examples of generated ‘poetry’, prompted by three lines from Al-Mutanabi, from JASMINE2.7B under zero-shot. We color the initial prompt with gray.
Generated Poems

(1) Empty prompt

كيف تضافن حياتي الرياض ** وكيف يصبض الفضول الحضي
وهبات لم تعمد أن يعود ** ولكن لامض فيه يجي
ماذا بحس ممن الضحكوا ** ولكن شفقة كلامكًا
وذا اليوم حمنا فافذا ** حفظنا له وانتظارنا غدا
ونغفر على حكم صرف الأذن ** وبين الحواج يمر الفضا
ويا رب إلب على السلمين ** زوى الحق عن أهل فازوي

(2) Part of poetry's line prompt

قه باموتي ** مصر بنادي عليك يا جمال
فأعى بالتحج قد لاح ** وودت لوجهك من الأمال
وانتغت الدنيا بكوكب ** يوه بور ينشثه الهلال
وأراد الله أن يعطى بك احبه ** وإذا بذن الملف في كسب الجمال

(3) One line of poetry prompt

أسي على بنت الكرم ** باسم فاضلة تطيب حطها
ما تمولت عين ستيزا ** إلا وحسبم يبر الراس تاراه
قصي المسر في لو وفي لعب ** لم ينتفح حويي بما وناعها
كم زاهلي خنج الحدي وجال ** تفهو جوتهام من أسواها
وتبين ينطبقي طوراً وفي ** فناسة الدور كثي في جاهاما
باجبة أي لاف قطع ** فينا الصيانة يبتسم وأباها

(4) Two lines of poetry prompt

وما نليك مينا كل بان ** تمسك أن يكون في الألم
فيما كنت بتبع أن تعني سعيدا ** فلا تأمن أن تشمس بالسجال
تتفتت والتوقعان ** أكفيك كوكب الدنيا في الزوال
والمست في تنامة الفهر غريبة ** فير، فير أهل، وألا أن
أناذك حروا وأذك أريجنا ** سل الله أن يجسنا رلال
فليس لنا غير الدهم دير ** بعيش بابه ونثار الأمال

(5) Three lines of poetry prompt

صباح شاعر حال ** أطغى فيه بالبهب وأشهد
وأند لابيد نحمي قريب ** ونداء النهر في سبيل النبال
أقبلت على النعم المراض ** وأطبب في البطن في الغالب
وأقلاط بالالأطلال في جر ** وسهر في قلب الزمان
وعلنت مع العين المرار ** على تلك الروس الخيالي
فاذ ما كان من قلي وما أحمدي السون في حالي

Table E.6: Examples of synthetically generated ‘poetry.’ prompted by a full (or part of) real line of poetry or empty prompt from our further pre-trained JASMINE model. We color the initial prompt with gray.
Table E.7: Examples of synthetically generated Arabic dialects text from STGen using JASMINE 2.7B under zero-shot setting. We color the initial prompt with gray.