Benchmarking LLaMA-3 on Arabic Language Generation Tasks

Md Tawkat Islam Khondaker^{λ}, Numaan Naeem^{ξ}, Fatimah Lyba Khan^{ξ}, AbdelRahim A. Elmadany^{λ}, Muhammad Abdul-Mageed^{λ,ξ}

^{\lambda}Deep Learning & Natural Language Processing Group, The University of British Columbia ^{\lambda}Department of Natural Language Processing & Department of Machine Learning, MBZUAI {tawkat@cs, muhammad.mageed, AbdelRahim}@ubc.ca

{numaan.naeem, fatimah.khan}@mbzuai.ac.ae

Abstract

Open-sourced large language models (LLMs) have exhibited remarkable performance in a variety of NLP tasks, often catching up with the closed-sourced LLMs like ChatGPT. Among these open LLMs, LLaMA-3-70B has emerged as the most recent and the most prominent one. However, how LLaMA-3-70B would situate itself in multilingual settings, especially in a rich morphological language like Arabic, has yet to be explored. In this work, we focus to bridge this gap by evaluating LLaMA-3-70B on a diverse set of Arabic natural language generation (NLG) benchmarks. To the best of our knowledge, this is the first study that comprehensively evaluates LLaMA-3-70B on tasks related to Arabic natural language generation. Our study reveals that LLaMA-3-70B lags behind the closed LLMs like ChatGPT, both in modern standard Arabic (MSA) and dialectal Arabic (DA). We further compare the performance of LLaMA-3-70B with our smaller and dedicated finetuned Arabic models. We find that both LLaMA-3-70B and ChatGPT are outperformed by comparatively smaller dedicated Arabic models, indicating the scope for potential improvement with Arabic-focused LLMs.

1 Introduction

In the last few years, the emergence of large language models (LLMs) (OpenAI, 2022, 2023; AI@Meta, 2024; Gemini et al., 2023; Anthropic, 2023; Jiang et al., 2023) has emerged as powerful systems to accomplish diverse set of natural language processing (NLP) tasks. Such success of these models primarily stems from pretraining them on large datasets with the next token prediction objective. The current state-of-the-art performance is however, dominated mostly by 'closed' models (OpenAI, 2023; Gemini et al., 2023; Anthropic, 2023). That is, little to no information about them is known. This includes details about model architectures, pretraining data, languages involved, and training configurations. Such LLMs are also expensive both to pretrain and deploy. To alleviate these concerns, 'open' LLMs such as LLaMA-3 (Touvron et al., 2023a), Mistral (Jiang et al., 2023), and OLMo (Groeneveld et al., 2024) were introduced to facilitate research and (non-) commercial deployment.

Rigorous evaluations (Qin et al., 2023; Gilardi et al., 2023; Laskar et al., 2023) have already been conducted to analyze LLMs performance on a wide range of English benchmarks. However, a limited number of studies can be found when it comes to evaluating LLMs on non-English languages (Bang et al., 2023a; Lai et al., 2023; Kasai et al., 2023; Huang et al., 2023a). This limitation becomes more acute, in the case of Arabic; a rich morphological language comprised of a wide range of dialects and varieties. Khondaker et al. (2023) evaluate closed models like ChatGPT and GPT-4 on a wide range of Arabic NLP tasks, while Kadaoui et al. (2023) experiment with different commercially available closed models (e.g., Bard) on various Arabic machine translation tasks. However, the performance of the open-sourced models on Arabic natural language generation (NLG) benchmarks is still an unexplored territory to this date.

To address this gap, in this study, we benchmark LLaMA-3-70B (AI@Meta, 2024), the most recent and the best open LLM, on a diverse set of 20 Arabic NLG tasks. Notably, our study reveals that LLaMA-3-70B lags behind closed-sourced LLMs like ChatGPT. We further show that even smaller, dedicated, finetuned Arabic models can outperform LLaMA-3-70B.

Our contributions can be summarized as follows:

- 1. We present the first benchmark of opensourced LLM, LLaMA-3-70B on a wide range of 20 Arabic NLG tasks.
- 2. We perform a systematic evaluation of LLaMA-3-70B and demonstrate how the

model situate itself compared to closed-source models like ChatGPT.

- We additionally conduct evaluation on Arabic dialects to show the inferiority of LLaMA-3-70B compared to ChatGPT.
- 4. Through our empirical analyses, we further demonstrate that the LLaMA-3-70B significantly lags behind much smaller Arabicfocused finetuned models. We believe such findings can motivate future studies on developing stronger open-sourced Arabic LLMs.

2 Related Work

This section reviews recent developments in Large Language Models (LLMs), focusing on their applications in Machine Translation, Question Answering, and Multilinguality. It particularly emphasizes tasks involving non-English languages such as Arabic. Please see Appendix A for a comprehensive guide.

2.1 Performance analysis on English

Open-sourced LLMs have advanced in machine translation yet faced challenges in domain-specific and low-resource scenarios. Innovations like LlamaIT framework (Zheng et al., 2024) reduce the need for detailed input examples, while the TER framework (Feng et al., 2024) enhances translation accuracy through systematic self-correction. Realtime adaptive translation capabilities have been demonstrated to be especially useful with limited in-domain data (Moslem, 2024), and improvements in instruction adherence through a two-stage finetuning algorithm (Zan et al., 2024) significantly enhance translation accuracy in low-resource languages. In question-answering (QA), open LLMs demonstrate significant potential, (Gramopadhye et al., 2024) enhance medical QA with a Chain of Thought approach, (Jiang et al., 2024) generate document-question pairs. Additionally, (Arefeen et al., 2024) introduce LeanContext, which uses reinforcement learning to tailor the amount of contextual information needed, (Kim et al., 2024) further refine open-domain (QA) through the SuRe framework, which supports answer validation. However, despite these advancements, (LLMs) still face challenges in achieving human-like performance. Majumdar et al. (2024) note that current models underperform in interactive environments such as Embodied (QA), where understanding and interacting with real-world contexts remains a substantial hurdle. For text classification, Guo et al. (2024) surpass traditional supervised models in healthrelated classifications using social media data. Similarly, Abburi et al. (2023) employ LLMs to distinguish between AI-generated and human-written texts, (Miah et al., 2024) confirm the effectiveness of LLMs in multilingual sentiment analysis, translating and processing texts with high accuracy across different languages. However, despite these successes, challenges persist as Chae and Davidson (2023) point out that while (LLMs) perform well in minimal-example scenarios, they still require extensive fine-tuning to achieve optimal accuracy.

2.2 Performance on Multilingual Data

In assessing the capabilities of open-sourced LLMs within multilingual contexts, several studies have illuminated both the strengths and the limitations these models exhibit across diverse language environments. Shen et al. (2024) highlight significant safety challenges, particularly when LLMs process lower-resource languages, where they tend to generate unsafe or irrelevant responses. Alam et al. (2024) address similar challenge across multilingual, multimodal, and dialectal settings, emphasizing the importance of instruction tuning and reinforcement learning. Meanwhile, Li et al. (2024) propose integrating code elements during training to strengthen multilingual structured reasoning, showcasing substantial improvements in tasks requiring scientific commonsense across various languages. Additionally, Zhang et al. demonstrate the effectiveness of Self-Distillation from Resource-Rich Languages (SDRRL) in leveraging the capabilities of LLMs trained on resource-rich languages to enhance performance more broadly. Liu et al. (2024) challenge the over-reliance on translation, advocating for native language prompting to capture cultural nuances better, and achieve more authentic multilingual functionality. Expanding on in-context learning, Cahyawijaya et al. (2024) explore its application in low-resource languages, emphasizing minimal training data's potential to yield effective language processing and examine how LLMs handle multilingual processing within their architectures,. Together, these studies form a comprehensive examination of LLMs in multilingual contexts, revealing promising strategies for improvement and persistent challenges that necessitate further research.

2.3 LLMs with Low Resource Languages

Efforts such as the HELM project and BIG-Bench have expanded the evaluation of Large Language Models (LLMs) to encompass low-resource languages, shedding light on the models' generalizability across a diverse range of tasks (Liang et al., 2022; Srivastava et al., 2022). Despite these extensive evaluations, studies including those by (Bang et al., 2023b; Ahuja et al., 2023) consistently report that LLMs underperform in low and extremely low-resource languages, highlighting a persistent challenge in achieving parity for non-English languages.

The representation of languages like Arabic in major datasets such as Common Crawl is notably limited, posing significant obstacles in LLM research. Even though models like Bloom are trained on data from numerous languages, they exhibit restricted effectiveness in non-English contexts, particularly for complex NLP tasks (Le Scao et al., 2022). A detailed evaluation by (Khondaker et al., 2023) further illuminates this issue, showing that while ChatGPT demonstrates strong performance in English, it is often surpassed by models that are fine-tuned specifically for Arabic. Their comprehensive analysis, which includes both automated and human evaluations on over 60 datasets, constitutes a significant exploration of ChatGPT's capabilities with Arabic dialects and Modern Standard Arabic (MSA), emphasizing the model's limitations in handling the nuanced characteristics of Arabic languages. Our research adds to these findings by assessing open LLM, LLaMA-3-70B across multiple Arabic datasets and tasks, providing deeper insights into the capabilities and limitations of LLaMA-3-70B in managing the distinct linguistic features of Arabic. This exploration helps to elucidate further the disparities in performance between general and specialized models in handling lowresource languages.

3 Dataset

We conduct a comprehensive evaluation of LLaMA-3-70B and ChatGPT on Arabic Natural Language Generation (NLG) tasks. This evaluation encompasses 20 distinct tasks, organized into 10 clusters, compiled from Dolphin benchmark (Nagoudi et al., 2023).

Machine Translation. This cluster evaluates translation performance across various languages, focusing on capturing nuanced meanings and contexts. Key tasks include translating from the six official UN languages to Modern Standard Arabic (MSA), using the UN parallel corpus (Ziemski et al., 2016) and the MultiUN corpus for training (Eisele and Chen, 2010). Another task involves translating Arabizi, a Romanized version of Arabic dialects, into English and French using the Darija (Outchakoucht and Es-Samaali, 2021) and NArabizi datasets (Seddah et al., 2020). Additionally, the cluster includes translating texts from six Arabic dialects into English utilizing the MDP corpus (Bouamor et al., 2018), with zero-shot training supported by additional MSA-English sentences.

Code-Switching. This task cluster addresses translating Arabic dialect texts, including codeswitching, into the corresponding foreign languages. Specifically, it includes translating 300 code-switched Arabic-French tweets from Twitter users in Algeria, Morocco, and Tunisia, and 300 code-switched Arabic-English posts from users in Egypt, Jordan, and Palestine. These translations are conducted in a zero-shot setting, utilizing 50,000 MSA-English and MSA-French sentences from the AraOPUS-20 dataset for monolingual training (Nagoudi et al., 2022b).

Text Summarization (TS). The Text Summarization cluster leverages five datasets featuring Arabic and multilingual content to enhance and evaluate summarization capabilities. The datasets include MassiveSum, which offers a vast collection of multilingual news articles designed for summarization (Varab and Schluter, 2021); XLSum, providing summaries across 44 languages with a focus on abstractive techniques (Hasan et al., 2021); Cross-Sum, which supports summarization over 1500 language pairs, enabling crosslingual abstractions (Bhattacharjee et al., 2021); ANT, a specialized corpus for extractive summarization in Arabic using news sources (Chouigui et al., 2021); and MarSum, targeting summarization in the Moroccan Arabic dialect with advanced AI methodologies (Gaanoun et al., 2022).

News Title Generation (NTG). This task focuses on crafting suitable and grammatically correct titles for news articles. It employs the Arabic (NTG) dataset, which is dedicated to creating titles specifically for Arabic news content (Nagoudi et al., 2022c), and the XLSum dataset, used for generating titles across multiple languages (Hasan et al., 2021).

Question Answering (QA). The QA cluster com-

prises four distinct tasks utilizing seven publicly available datasets. Extractive QA employs the Arabic QA dataset ARCD and Arabic sections from multilingual QA test sets like MLQA, XQuAD, and TyDiQA, with models fine-tuned using the GoldP multilingual TyDiQA training set (Mozannar et al., 2019; Lewis et al., 2019; Artetxe et al., 2020). Retrieval QA involves LAReQA, adapted from XQuAD into a retrieval task focusing on its Arabic version AraQuAD-R (Roy et al., 2020). Open-Domain QA answers fact-based questions using the DAWQAS, an Arabic Why QA dataset (Ismail and Nabhan Homsi, 2018). Multi-choice QA uses the EXAMS dataset, a cross-lingual multi-choice QA collection that includes Arabic, evaluated in a zero-shot scenario (Hardalov et al., 2020).

Question Generation (QG). The Question Generation cluster focuses on creating questions from provided passages. The objective is for models to generate straightforward, yet pertinent questions along with their answers, based on the approach described in (Gehrmann et al., 2021).

Paraphrase. The Paraphrase task focuses on generating paraphrases of Arabic sentences while preserving their original meaning. It utilizes four datasets: AraPara, a multi-domain Arabic paraphrase dataset (Nagoudi et al., 2022a); ASEP, from the Arabic SemEval paraphrasing challenge (Cer et al., 2017); the Arabic Paraphrasing Benchmark (APB), a key dataset for testing Arabic paraphrase systems (Alian et al., 2019); and TaPaCo, the Arabic portion of a multilingual paraphrase corpus (Scherrer, 2020).

Transliteration. The transliteration task focuses on converting text from one writing system to another while maintaining the original pronunciation. It utilizes three word-level datasets: ANETA, for transliteration and classification of English-Arabic named entities (Ameur et al., 2019); ATAR, a parallel corpus for transliterating Jordanian Arabizi into Modern Standard Arabic (MSA) (Talafha et al., 2021); and NETransliteration, a comprehensive dataset for transliterating named entities from English to Arabic, Hebrew, Japanese (Katakana), and Russian, sourced from Wikidata (Merhav and Ash, 2018).

Text Rewriting. The text rewriting task focuses on recreating text in a specific target style while maintaining the original content. It includes converting Arabic dialectal text into Modern Standard Arabic (MSA) using the Dial2MSA corpus, which covers dialects like Egyptian, Maghrebi, Levantine, and Gulf (Mubarak, 2018), and adapting sentences to the opposite gender with the Arabic parallel gender corpus (Alhafni et al., 2022).

Diacritization. The diacritization task restores missing diacritics in Arabic texts, utilizing the Arabic diacritization dataset (Fadel et al., 2019).

4 Prompt design

In this study, we define a prompt as a customized set of instructions that shapes the functionality and enhances the output of a large language model, aligning with the framework proposed by (White et al., 2023). The design of a prompt is critical as it influences how the model interacts with users and the quality of its responses by establishing specific operational rules. Therefore, to ensure the model's responses meet desired outcomes, it's essential to articulate these rules and objectives.

In our initial experiments, we employed a variety of prompts in both English and Arabic to evaluate their efficacy across several tasks. These included dialect identification, detection of machinegenerated texts, and toxicity detection. Our observations indicated that prompts in English consistently outperformed their Arabic counterparts.

We developed a standardized English prompt template to optimize performance. This template systematically guides the model by (1) Assigning a specific *role* to the model, such as acting as a translator for natural language generation (NLG) tasks; (2) Naming the *task* to be undertaken, like machine translation; (3) Describing the expected *outcome*, such as accurately translating text from one language to another; (4) Incorporating k-shot learning *examples* if applicable, and; (5) Providing a *test input* to generate an output.

We showcase the structure of our prompts in Figure 1. This structured approach allows us to fully utilize the capabilities of LLMs by ensuring they operate within well-defined parameters that meet the complexities of the tasks at hand.

5 Experiments

Setup. We run experiments under *zero*-shot, 3-shot, and 5-shot settings. For the evaluation, we use the same 200 test samples from Khondaker et al. (2023) to ensure fair comparison. Similarly, we also use the exact same training samples from Khondaker et al. (2023) for k-shot settings. For all the models, we use free-form generation

0-Shot	k-Shot				
	You are a generative language model. Your task is to				
You are a generative language model. Your	do <i>task</i> . You will be given an input text and you				
task is to do <i>task</i> . You will be given an input	should generate the <i>task outcome</i> .				
text and you should generate the <i>task outcome</i> .	Here, we provide you k examples of the task:				
Now generate the output of the following	Example 1:				
input:	Example 2:				
Input: <i>test input</i>	Now predict the label of the following input:				
Output:	Input: test input				
	Output:				

Figure 1: Prompt templates for NLG tasks.

together with simple posthoc-processing, e.g., removing whitespace.

Models Comparisons. For benchmarking, we use LLaMA-3-70B¹ from *HuggingFace* (Wolf et al., 2020) library. We set the maximum length of the newly generated tokens to 256. We use 4 Nvidia A100 40GB GPUs to generate the responses. Additionally, we obtain the performance of SoTA finetuned AraT5² model (Nagoudi et al., 2022a; Elmadany et al., 2023) on the sampled 200 test set and the full test set from Khondaker et al. (2023). We also obtain the performance of ChatGPT (gpt-3.5-turbo)³ from Khondaker et al. (2023).

6 Evaluation Result

We present the results of our evaluation on natural language generation tasks in Table 1. We observe that ChatGPT *outperforms* LLaMA-3-70B *on most of the tasks. On the other hand, the finetuned* AraT5 *outperforms both* LLaMA-3-70B *and* ChatGPT *on the majority of tasks.* We now briefly describe our results for the different NLG tasks.

Text Rewriting. For text rewriting, LLaMA-3-70B achieves BLEU scores of 41.83 with 3-shots and outperforms ChatGPT. However, AraT5 outperforms all the models.

Paraphrase. For the paraphrase generation task, LLaMA-3-70B is outperformed by ChatGPT in all *k*-shot by large margins. Furthermore, both models are outperformed by AraT5.

Question Generation and Question Answering. For the question generation task, ChatGPT outperforms LLaMA-3-70B on all the respective zero- and

¹meta-llama/Meta-Llama-3-70B-Instruct

AraT5v2-base-1024

k-shots. However, both models are significantly outperformed by AraT5.

For *QA*, LLaMA-3-70B outperforms ChatGPT in all the respective few-shot settings. Specifically, LLaMA-3-70B achieves the best score of 70.53 with 3-shot learning, whereas ChatGPT achieves 54.14 at best with 5-shot learning. However, both models are outperformed by AraT5 (81.45).

Summarization. For *summarization*, LLaMA-3-70B with 0-shot (with 18.83 BLEU) beats ChatGPT in the corresponding shot. However, ChatGPT with 3- and 5-shots outperforms LLaMA-3-70B in the respective few-shots. Both models are outperformed by finetuned AraT5.

News Title Generation. For *news title generation*, LLaMA-3-70B with 0-shot achieves the best scores (4.99 BLEU) and outperforms ChatGPT. Again, AraT5 shows better performance than both (7.72).

Diacritization and Transliteration. For *diacritization*, LLaMA-3-70B is outperformed by ChatGPT with lower error rates, while AraT5 (CER=0.03) still outperforms both LLMs. For *transliteration*, LLaMA-3-70B (0.41) is outperformed by ChatGPT (0.23). Again, AraT5 (0.18) achieves the lowest CER scores.

Machine Translation. ChatGPT outperforms LLaMA-3-70B on all the X (English, Spanish, French, Russian) $\rightarrow Arabic$ MT tasks. Both models perform better when the source language is English, signifying the availability of English data the models have been pretrained. Nevertheless, AraT5 still exibits superior performance over all the models.

Code-Switched Translation. For Jordanian Arabic (*Jo*) mixed with English \rightarrow English codeswitched translation task, LLaMA-3-70B (38.80 BLEU) with 5-shot outperforms ChatGPT by a small margin (0.25). Similarly, for Algerian

²https://huggingface.co/UBC-NLP/

³Snapshot of gpt-3.5-turbo from March 1st 2023

Task	Metric	LLaMA-3-70B (N-shot)		ChatGPT (N-shot)			AraT5 (Test No.)	AraT5 (Test No.)	
		0	3	5	0	3	5	200	Full
Text Rewriting	BLEU	8.14	41.82	41.13	41.59	58.75	53.34	99.64	91.19
Paraphrase	BLEU	6.56	3.45	6.74	7.89	8.92	9.19	14.40	18.69
Question-Gen	BLEU	10.21	18.91	19.13	14.48	19.86	20.08	35.17	33.64
QA	SQuAD F1	67.75	70.53	70.16	32.98	51.73	54.14	81.45	83.34
Summarization	RÒUGE-L	18.83	18.69	17.39	16.88	20.01	20.43	35.31	26.88
News Title-Gen	BLEU	4.99	4.67	3.15	3.24	4.72	4.62	7.72	9.64
Diacritization ↓	CER	0.20	0.08	0.08	0.11	0.06	0.05	0.03	0.01
Transliteration ↓	CER	0.41	0.47	0.46	0.27	0.24	0.24	0.18	0.18
MT (en \rightarrow ar)	BLEU	18.39	10.24	3.20	20.52	23.58	23.34	27.12	28.12
MT (es \rightarrow ar)	BLEU	15.03	14.48	14.37	16.47	18.11	17.45	21.16	21.74
MT $(fr \rightarrow ar)$	BLEU	13.58	12.43	11.34	15.12	15.44	15.57	18.48	20.51
MT (ru \rightarrow ar)	BLEU	13.33	13.74	12.95	15.83	17.52	17.46	19.32	18.29
CST (Jo-en→en)	BLEU	35.79	37.38	38.80	6.61	37.38	38.55	5.56	6.29
CST (Dz-fr→fr)	BLEU	35.76	39.71	38.29	34.61	35.40	36.45	17.49	16.16

Table 1: NLG Results. Higher is better unless otherwise specified by \downarrow . We evaluate LLaMA-3-70B and ChatGPT in 0-shot and in-context n-shot (where n = 3, 5) settings. AraT5 is our fully supervised model. The best scores are in **bold**. **QA** - Question Answering, **MT** - Machine Translation, **CST** - Code Switched Translation. We report the results of AraT5 on the same test set (200 samples) as ChatGPT and LLaMA-3-70B for a fair comparison.

Arabic (Dz) mixed with French \rightarrow French codeswitched translation task, LLaMA-3-70B (39.71) with 3-shot outperforms ChatGPT. Noticeably, both LLaMA-3-70B and ChatGPT exibit higher performance than the finetuned AraT5.

7 Performance on Dialectal Arabic

To determine the capability of open LLM for dialectal Arabic generation task, we compare the performance of LLaMA-3-70B and ChatGPT on dialectal machine translation task. Specifically, we use MSA and five Arabic dialects (*Egyptian, Jordanian, Palestinian, Syrian,* and *Tunisian*) \rightarrow English MT tasks from the Multi-dialectal Parallel Corpus (MDPC) Bouamor et al. (2014) and analyze the performance of the models for few-shot settings.



Figure 2: *K*-shot BLEU comparison between LLaMA-3-70B (indicated by *x*) and ChatGPT (indicated by *o*) on MSA and 5 dialects \rightarrow English MT.

As Figure 2 shows, ChatGPT outperforms LLaMA-3-70B on MSA and all the dialects (except

Palestinian and Syrian) across all the setups. Moreover, the performance gap tends to reduce as the number of shots increases. This indicates that with enough dialectal training data, open LLMs can improve their performance and close the gap with the closed LLM like ChatGPT.

Table 2 shows samples of generation outputs from ChatGPT and LLaMA-3-70B for translation tasks involving Jordanian-English code-switching translation, Egyptian dialect tranlation to English, as well as for diacritization and transliteration tasks. Following (Rust et al., 2020), we compute the fertility score to assess the capability of Llama-3 in handling Arabic languages, and compare it with the Arabic-based model. ⁴ We found that Llama-3 enhances its representation for Arabic compared to Llama-2, where Llama-3's fertility score is 2.24 and Llama-2's is 5.07. However, AraT5 outperforms both with a fertility score of 1.25.

8 Conclusion and Future Work

In this work, we conduct a systematic evaluation of LLaMA-3-70B on a wide range of Arabic NLG tasks. We benchmark LLaMA-3-70B and comparing against closed LLM like ChatGPT, we show that open LLMs like LLaMA-3-70B noticeably lags behind its counterpart. Additionally, we compare both closed and open-sourced LLMs with smaller Arabic-focused finetuned model to demonstrate the superiority of such smaller but finetuned models.

⁴The fertility score measures the average number of subwords produced per tokenized word. A minimum fertility of 1 indicates that the tokenizer's vocabulary encompasses every single word in the text.

Task	CST (Jo-en→en)			
Prompt	احساسي طلع بمحله'very nice'لسا مبيرح حكيت الامتحان حيكون			
Gold Prediction	just yesterday, i told the exam, it will be 'very nice', my feeling turns to be true			
LLAMA3-70B	i still haven't taken the exam, it will be'very nice', my feeling is that it will come out in its place			
ChatGPT	just yesterday, i told the exam, it will be 'very nice', my feeling turns to be true			
Task	Dialect Translation (EG \rightarrow En)			
Prompt	جيت عشان اقول هو فين ده لاقيت المسج بيقول في الاخر ده مضش حقيقه			
Gold Prediction	i came to say where it is i found a massage saying in the last this is not truth			
LLAMA3-70B	i came to say where is this? i didn't find the message, it says in the end this is not true at all			
ChatGPT	i came to say where is he, but i found the message saying in the end it was not true.			
Task	Diacritization			
Prompt	كأن أتلفت ممسكها البالغ الحر ، وإلا فعلى من أمرهما .			
Gold Prediction	كَان أتلفت ممسكها البالغ الحر ، وإلا فعلى من أمرهما . كَأَنْ أَتْلَفَتْ مُمْسِكَهَا الْبَالِغَ الْحُرَّ ، وَإِلَّا فَعَلَى مَنْ أَمَرَهُمَا .			
LLAMA3-70B	كَأَنْ أَتْلِفْتُ ثُمْسِكَهَا الْبَالِغُ الْحُرُّ ، وَإِلَّا فَعَلَى مَنْ أَمْرَهُمَا.			
ChatGPT	كَأَنَّ أَتْلَفَتْ مُمْسِكَهَا الْبَالِغَ الْحُرِّ ، وَإِلَّا فَعَلَى مَنْ أَمْرَهُمَا .			
Task	Transliteration			
Prompt	توبمين			
Gold Prediction	topmen			
LLAMA3-70B	tubmin			
ChatGPT	tubmin			

Table 2: A randomly picked examples from models generations

Finally, we extend our analysis on dialectal Arabic to reveal the inferiority of LLaMA-3-70B against ChatGPT. We believe our effort will encourage researchers to develop and improve Arabic-focused LLMs. In future, we aim to scale our benchmark to other open LLMs as well as Arabic language understanding tasks.

9 Limitations

Model Variation. We conduct our evaluation on the most recent and most prominent open LLM, LLaMA-3-70B. However, we do not include other open LLMs such as AceGPT-7B (Huang et al., 2023b), jais-13B (Sengupta et al., 2023), gemma-7B (Gemini et al., 2023). We acknowledge that incorporating a large number of models can facilitate the comparative analysis and we intend to include them in our future work.

Experimental Setup. In this work, we use the performance of ChatGPT benchmarked by Khondaker et al. (2023). However, ChatGPT's version can be updated, as a consequence, the performance of ChatGPT can vary. Therefore, the results and the corresponding analyses reported in this work should be treated accordingly, since the model's responses can change over time (Chen et al., 2023). **Evaluation.** We suspect that the inferior performance of LLaMA-3-70B can potentially be attributed to insufficient Arabic data during the pretraining stage as well as the lack of Arabic instruction data for finetuning. Although a different sets of prompt may improve the performance, we attempt to keep the prompt template same across all the models for a fair comparison.

Results. We observe that LLaMA-3-70B's performance does not necessarily increase as with the increased number of shots. This is because the improvement with few-shot learning is model- and task-dependent, and often sensitive to the order of the shots Wei et al. (2021); Brown et al. (2020).

10 Ethics Statement

Data Collection. We collect the NLG evaluation datasets from Dolphin (Nagoudi et al., 2023). To ensure proper credit assignment, we refer users to the original publication (Section 3).

Intended Use. We believe our findings will encourage further research on studying open LLMs on Arabic NLG benchmarks, as we show that LLaMA-3-70B still lags behind compared to the

smaller finetuned model. Therefore, our work can spur the interest among the researchers to develop Arabic language dedicated LLMs that can match or outperform the SOTA finetuned models.

Potential Misuse and Bias. LLMs can produce potentially harmful and biased contents (Laskar et al., 2023). Therefore, we recommend that these models not be used in applications without careful prior consideration of potential misuse and bias.

References

- Harika Abburi, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. 2023. Generative ai text classification using ensemble Ilm approaches. Computation and Language (cs.CL); Artificial Intelligence (cs.AI).
- Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Yousseif Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024. Larabench: Benchmarking arabic ai with large language models. *arXiv preprint arXiv:2305.14982v2*.
- Khalid Ahmed et al. 2024. Impact of dialect-specific training on text classification performance in arabic. *Language Resources and Evaluation*, 58(1):45–67.
- Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023.
 Mega: Multilingual evaluation of generative ai. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card.

- Abdulmohsen Al-Thubaity, Sakhar Alkhereyf, Hanan Murayshid, Nouf Alshalawi, Maha Bin Omirah, Raghad Alateeq, Rawabi Almutairi, Razan Alsuwailem, Manal Alhassoun, and Imaan Alkhanen. 2023. Evaluating chatgpt and bard ai on arabic sentiment analysis. In *Proceedings of the First Arabic Natural Language Processing Conference (Arabic-NLP 2023)*, pages 335–349. Association for Computational Linguistics.
- Firoj Alam, Shammur Absar Chowdhury, Sabri Boughorbel, and Maram Hasanain. 2024. Llms for low resource languages in multilingual, multimodal and dialectal settings. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33. Association for Computational Linguistics.

- Hessa Abdulrahman Alawwad, Areej Alhothali, Usman Naseem, Ali Alkhathlan, and Amani Jamal. 2024. Enhancing textbook question answering task with large language models and retrieval augmented generation. arXiv:2402.05128 [cs.CL].
- Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022. User-centric gender rewriting. In *Proceedings* of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 618–631, Seattle, United States. Association for Computational Linguistics.
- Marwah Alian, Arafat Awajan, Ahmad Al-Hasan, and Raeda Akuzhia. 2019. Towards building arabic paraphrasing benchmark. In *Proceedings of the Second International conference on Data Science E-learning and Information Systems (DATA' 2019)*, pages 1–5.
- M. Alkaoud. 2024. A bilingual benchmark for evaluating large language models. *PeerJ Computer Science*, 10:e1893.
- Zaid Alyafeai and Moataz Ahmed. 2023. Investigating zero-shot cross-lingual language understanding for arabic. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, pages 324–334, Dhahran, Saudi Arabia. Association for Computational Linguistics.
- Mohamed Seghir Hadj Ameur, Farid Meziane, and Ahmed Guessoum. 2019. Anetac: Arabic named entity transliteration and classification dataset. *arXiv preprint arXiv:1907.03110*.

Anthropic. 2023. Introducing claude.

- Md Adnan Arefeen, Biplob Debnath, and Srimat Chakradhar. 2024. Leancontext: Cost-efficient domain-specific question answering using llms. *Journal of Natural Language Processing*, 2024(100065).
- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023a. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *Preprint*, arXiv:2302.04023.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023b. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. arXiv preprint arXiv:2302.04023.

- Abhik Bhattacharjee et al. 2021. Crosssum: Beyond english-centric crosslingual abstractive text summarization for 1500+ language pairs. *CoRR*, abs/2112.08804.
- Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of arabic. In *LREC*, pages 1240–1245.
- Houda Bouamor et al. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).*
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.
- Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. Llms are few-shot in-context low-resource language learners.
- Daniel Cer, Mona Diab, Eneko Agirre, Inigo LopezGazpio, and Lucia Specia. 2017. Semeval-2017 task
 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.
- Youngjin Chae and Thomas Davidson. 2023. Large language models for text classification: From zeroshot learning to fine-tuning. *Rutgers University*.
- Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is chatgpt's behavior changing over time? *arXiv* preprint arXiv:2307.09009.
- Amina Chouigui, Oussama Ben Khiroun, and Bilel Elayeb. 2021. An arabic multi-source news corpus: Experimenting on single-document extractive summarization. Arabian Journal for Science and Engineering, 46:3925–3938.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1, pages 4171–4186.
- Gourab Dey, Adithya V Ganesan, Yash Kumar Lal, Manal Shah, Shreyashee Sinha, Matthew Matero, Salvatore Giorgi, Vivek Kulkarni, and H. Andrew Schwartz. 2024. Socialite-llama: An instructiontuned model for social scientific tasks. *Computation and Language*.

- Nadir Durrani, Yaser Al-Onaizan, and Abraham Ittycheriah. 2014. Improving egyptian-to-english smt by mapping egyptian into msa. In *Computational Linguistics and Intelligent Text Processing - 15th International Conference, CICLing 2014*, volume 8404 of *Lecture Notes in Computer Science*, pages 271– 282. Springer.
- Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA).
- AbdelRahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Octopus: A multitask model and toolkit for Arabic natural language generation. In *Proceedings of ArabicNLP 2023*, pages 232–243, Singapore (Hybrid). Association for Computational Linguistics.
- Ali Fadel, Ibraheem Tuffaha, Bara' Al-Jawarneh, and Mahmoud Al-Ayyoub. 2019. Arabic text diacritization using deep neural networks.
- Zhaopeng Feng, Yan Zhang, Hao Li, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Improving llm-based machine translation with systematic self-correction. *Journal of Computation and Language*, 58(1):101–123.
- Kamel Gaanoun et al. 2022. Automatic Text Summarization for Moroccan Arabic Dialect Using an Artificial Intelligence Approach, pages 158–177.
- Sebastian Gehrmann et al. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. In *Proceedings*, pages 96–120.
- Team Gemini, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for textannotation tasks. *Preprint*, arXiv:2303.15056.
- Ojas Gramopadhye, Saeel Sandeep Nachane, Prateek Chanda, Ganesh Ramakrishnan, Kshitij Sharad Jadhav, Yatin Nandwani, Dinesh Raghu, and Sachindra Joshi. 2024. Few shot chain-of-thought driven reasoning to prompt llms for open ended medical question answering. *arXiv:2403.04890*.
- Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.
- Yuting Guo, Anthony Ovadje, Mohammed Ali Al-Garadi, and Abeed Sarker. 2024. Evaluating large language models for health-related text classification tasks with public social media data.

- Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. Exams: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444. Association for Computational Linguistics.
- Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023a. Not all languages are created equal in llms: Improving multilingual capability by cross-lingualthought prompting. *Preprint*, arXiv:2305.07004.
- Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2023b. Acegpt, localizing large language models in arabic. *Preprint*, arXiv:2309.12053.
- Walaa Ismail and Masun Nabhan Homsi. 2018. Dawqas: A dataset for arabic why question answering system. *Procedia Computer Science*, 142:123–131.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. arXiv preprint arXiv:2310.06825.
- Feihu Jiang, Chuan Qin, Kaichun Yao, Chuyu Fang, Fuzhen Zhuang, Hengshu Zhu, and Hui Xiong. 2024. Enhancing question answering for enterprise knowledge bases using large language models. *arXiv preprint arXiv:2404.08695*.
- Karima Kadaoui, Samar Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. TARJAMAT: Evaluation of bard and ChatGPT on machine translation of ten Arabic varieties. In *Proceedings of ArabicNLP 2023*, pages 52–75, Singapore (Hybrid). Association for Computational Linguistics.
- Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. *Preprint*, arXiv:2303.18027.
- Sara Kazemi and Jian Li. 2024. Hybrid approaches to machine translation for arabic dialects. *Journal of Machine Translation*, 38(2):112–130.

- Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 220–247, Singapore. Association for Computational Linguistics.
- Jaehyung Kim, Jaehyun Nam, Sangwoo Mo, Jongjin Park, Sang-Woo Lee, Minjoon Seo, Jung-Woo Ha, and Jinwoo Shin. 2024. Sure: Summarizing retrievals using answer candidates for open-domain qa of llms. arXiv preprint arXiv:2404.13081.
- Anis Koubaa, Adel Ammar, Lahouari Ghouti, Omar Najar, and Serry Sibaee. 2024. Arabiangpt: Native arabic gpt-based large language model. arXiv:2402.15313 [cs.CL].
- Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP* 2023, pages 13171–13189, Singapore. Association for Computational Linguistics.
- Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq R. Joty, and J. Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. *ArXiv*, abs/2305.18486.
- Teven Le Scao et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv* preprint arXiv:2211.05100.
- Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.
- Bryan Li, Tamer Alkhouli, Daniele Bonadiman, Nikolaos Pappas, and Saab Mansour. 2024. Eliciting better multilingual structured reasoning from llms through code.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.
- Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual language models.
- Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? a study on solving multilingual tasks with large language models.

- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
- Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul McVay, Oleksandr Maksymets, Sergio Arnaud, et al. 2024. Openeqa: Embodied question answering in the era of foundation models. MoMa WS 2024 Oral Presentation.
- Yuval Merhav and Stephen Ash. 2018. Design challenges in named entity transliteration. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 630–640, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Md Saef Ullah Miah, Md Mohsin Kabir, Talha Bin Sarwar, Mejdl Safran, Sultan Alfarhood, and M. F. Mridha. 2024. A multimodal approach to crosslingual sentiment analysis with ensemble of transformer and llm. *Scientific Reports*, 14:9603.
- Isabel Moreno and Wei Zhang. 2024. Evaluating multilingual models on nlp tasks in arabic. *Computational Linguistics*, 50(3):425–445.
- Yasmin Moslem. 2024. Language modelling approaches to adaptive machine translation.
- Hussein Mozannar, Karl El Hajal, Elie Maamary, and Hazem Hajj. 2019. Neural arabic question answering. *arXiv preprint arXiv:1906.05394*.
- Hamdy Mubarak. 2018. Dial2msa: A tweets corpus for converting dialectal arabic to modern standard arabic. In OSACT 3: The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning.
- El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022a. Arat5: Textto-text transformers for arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.
- El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022b. Turjuman: A public toolkit for neural arabic machine translation. In Proceedings of the 5th Workshop on Open Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection, pages 1–11, Marseille, France. European Language Resources Association.

- El Moatez Billah Nagoudi, AbdelRahim Elmadany, Ahmed Oumar El-Shangiti, and Muhammad Abdul-Mageed. 2023. Dolphin: A challenging and diverse benchmark for arabic nlg. In *Findings of the Association for Computational Linguistics: EMNLP* 2023, pages 1404–1422, The University of British Columbia, MBZUAI. Association for Computational Linguistics.
- El Moatez Billah Nagoudi et al. 2022c. Arabic ntg dataset for news title generation.

OpenAI. 2022. Introducing chatgpt.

- OpenAI. 2023. Gpt-4 technical report. ArXiv, abs/2303.08774.
- Aissam Outchakoucht and Hamza Es-Samaali. 2021. Moroccan dialect -darija- open dataset.
- Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *Preprint*, arXiv:2302.06476.
- Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. Lareqa: Language-agnostic answer retrieval from a multilingual pool. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5919–5930. Association for Computational Linguistics.
- Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2020. How good is your tokenizer? on the monolingual performance of multilingual language models. *arXiv preprint arXiv:2012.15613*.
- Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. Arabench: Benchmarking dialectal arabic-english machine translation. In Proceedings of the 28th International Conference on Computational Linguistics, pages 5094–5107.
- Hassan Sajjad, Fahim Dalvi, Nadir Durrani, Ahmed Abdelali, Yonatan Belinkov, and Stephan Vogel. 2017.
 Challenging language-dependent segmentation for arabic: An application to machine translation and part-of-speech tagging. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 601–607. Association for Computational Linguistics.
- M. Sallam and D. Mousa. 2024. Evaluating chatgpt performance in arabic dialects: A comparative study showing defects in responding to jordanian and tunisian general health prompts. *Mesopotamian Journal of Artificial Intelligence in Healthcare*, pages 1–7.
- Yves Scherrer. 2020. Tapaco: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France. European Language Resources Association.

- Djamé Seddah et al. 2020. Building a user-generated content north-african arabizi treebank: Tackling hell. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1139– 1150. Association for Computational Linguistics.
- Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jaischat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.
- Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of llms in multilingual contexts.
- Amit Singh and Dev Patel. 2024. Integrating llms with traditional nlp techniques for improving arabic qa systems. In *Proceedings of the 2024 International Conference on Language Resources and Evaluation*, pages 738–752.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- Bashar Talafha, Analle Abuammar, and Mahmoud Al-Ayyoub. 2021. Atar: Attention-based lstm for arabizi transliteration. *International Journal of Electrical and Computer Engineering*, 11:2327–2334.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, et al. 2022. No language left behind: Scaling humancentered machine translation.
- Atnafu Lambebo Tonja, Israel Abebe Azime, Tadesse Destaw Belay, Mesay Gemeda Yigezu, Moges Ahmed Mehamed, Abinew Ali Ayele, Ebrahim Chekol Jibril, Michael Melese Woldeyohannis, Olga Kolesnikova, Philipp Slusallek, Dietrich Klakow, Shengwu Xiong, and Seid Muhie Yimam. 2024. Ethiollm: Multilingual large language models for ethiopian languages with task evaluation. *arXiv:2403.13737 [cs.CL]*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. ArXiv, abs/2302.13971.

- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Daniel Varab and Natalie Schluter. 2021. Massivesumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2023a. Evaluating open-qa evaluation. In Advances in Neural Information Processing Systems, volume 36 of Datasets and Benchmarks Track. NeurIPS.
- Hanyin Wang, Chufan Gao, Christopher Dantona, Bryan Hull, and Jimeng Sun. 2023b. Drg-llama: Tuning llama model to predict diagnosis-related group for hospitalized patients. *Artificial Intelligence (cs.AI); Computation and Language (cs.CL)*.
- Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.
- Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. ArXiv, abs/2302.11382.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.
- Niraj Yagnik, Jay Jhaveri, Vivek Sharma, and Gabriel Pila. 2024. Medlm: Exploring language models for medical question answering systems. *arXiv:2401.11389 [cs.CL]*.
- Changtong Zan, Liang Ding, Li Shen, Yibing Zhen, Weifeng Liu, and Dacheng Tao. 2024. Building accurate translation-tailored llms with language aware instruction tuning. *Journal of Computation and Language*, 42(3):78–97.

- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
- Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages.
- Jiawei Zheng, Hanghai Hong, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Finetuning large language models for domain-specific machine translation. *Journal of Computation and Language*, 52(4):356–367.
- Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. Why does chatgpt fall short in answering questions faithfully?
- Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *Computation and Language*.
- Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis.
- Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus. *LREC*, pages 3534–3539.

Appendices

A Literature Review

A.1 Machine Translation

The challenges in machine translation (MT) of Arabic stem from its complex morphology and the diversity of dialects, which deviate significantly from Modern Standard Arabic (MSA), influencing phonetics, lexicon, and grammar. Durrani et al. (2014) addressed these challenges by proposing a method to map Egyptian Arabic to MSA, aiming to standardize input data to enhance the effectiveness of translation models. This foundational work paved the way for further explorations into language-specific issues, such as segmentation challenges, which Sajjad et al. (2017) investigate, emphasizing the importance of tailored segmentation strategies in improving both MT and partof-speech tagging. Recognizing the need for specialized resources to support the development of robust MT systems, Sajjad et al. (2020) introduce AraBench. This extensive dataset is designed to benchmark the performance of dialectal Arabic-English MT systems, encompassing a wide range of dialect categories and covering diverse genres. Such comprehensive data resources are crucial for training MT systems to handle the spectrum of dialectal variations encountered in real-world applications.

Recent advancements also highlight the role of hybrid models that combine proprietary and opensource technologies, which have significantly improved translating dialectal Arabic (Kazemi and Li, 2024). In parallel, studies like those by Zhu et al. (2023) evaluate the capabilities of large language models (LLMs), such as ChatGPT, XGLM (Lin et al., 2022), OPT (Zhang et al., 2022), and BLOOMZ (Muennighoff et al., 2022). These evaluations demonstrate that while ChatGPT excels in zero-shot and in-context few-shot scenarios, it still trails behind fully supervised models such as those in the 'No Language Left Behind' initiative (Team et al., 2022).

Further expanding the resources for Arabic MT, Abdelali et al. (2024) introduce LAraBench, a comprehensive benchmark for Arabic AI that evaluates various LLMs, including GPT-3.5 and GPT-4, along with speech models like Whisper and USM. Their extensive testing across 61 datasets for 33 tasks reveals that while state-of-the-art (SOTA)

models typically outperform LLMs in zero-shot learning, the gap narrows considerably with applying few-shot techniques. This suggests a promising potential for LLMs in Arabic language applications, particularly when leveraged with appropriate learning strategies.

A.2 Text Classification

Text classification has witnessed significant advancements with the use of Large Language Models (LLMs) like ChatGPT, which excels in zeroshot and in-context few-shot settings. (Zhong et al., 2023) highlight that ChatGPT performs comparably to fully supervised models on the GLUE NLU benchmark, outperforming BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) on specific tasks such as MNLI, SST2, and RTE, although it lags in other areas. The importance of localized training datasets is also emphasized, as demonstrated by Ahmed et al. (2024), who find that models trained on dialect-specific data significantly surpass those trained on generalized Arabic datasets, underscoring the value of tailored data in text classification.

Further developments include instruction-tuned models like SOCIALITE-LLAMA, which have shown improved performance across specialized NLP tasks (Dey et al., 2024). Another innovative approach is seen in the DRG-LLaMA model, which uses fine-tuning on clinical notes to enhance DRG assignment accuracy and efficiency, substantially outperforming established models like ClinicalBERT and CAML (Wang et al., 2023b). Despite these advancements, challenges remain, particularly in handling Arabic dialects. For instance, even sophisticated models like GPT-4 encounter difficulties generating precise dialectal sentiment text, highlighting ongoing issues within specialized text classification contexts (Al-Thubaity et al., 2023).

A.3 Question Answering (QA)

Advancements in QA have seamlessly integrated traditional NLP techniques with Large Language Models (LLMs), enhancing the handling of idiomatic expressions in Arabic as demonstrated by Singh and Patel (2024). Further exploring crosslingual capabilities, Alyafeai and Ahmed (2023) investigate the effectiveness of mBERT (Devlin et al., 2019) in zero-shot language transfer to Arabic. Their study focuses on tasks such as natural language inference and question answering, emphasizing the impact of morphological similarities between languages like Russian and Arabic on model performance. Additionally, integrating retrievalaugmented generation with models like Llama-2 (Touvron et al., 2023b) has shown substantial improvements in complex, domain-specific QA settings (Alawwad et al., 2024).

Domain-specificity also plays a crucial role in enhancing QA systems, as evidenced by Yagnik et al. (2024), who evaluate the performance of both general and medical-specific distilled models in medical QA. Their work highlights the significant benefits of domain-specific fine-tuning in improving model accuracy. Complementing these domainspecific approaches, Wang et al. (2023a) introduce the QA-Eval task and the EVOUNA dataset, designed to advance the evaluation of Open-QA systems. Their research underscores the ongoing challenges by comparing AI-generated answers with standard responses and highlights potential improvements in automatic evaluation methods. In the broader scope of open-domain question answering, Zheng et al. critically assess ChatGPT's performance, revealing its capabilities in handling complex user queries where no context is provided. Their analysis demonstrates the model's impressive results and proposes methods to enhance the faithfulness of its answers, addressing key areas where ChatGPT may falter.

A.4 Multilinguality

Research in multilingualism has highlighted significant variations in the performance of Large Language Models (LLMs) across different languages, particularly focusing on Arabic. Moreno and Zhang (2024) examine the efficacy of multilingual BERT variants, which show enhanced capabilities across various Arabic dialects. Further studies, such as those by Sallam and Mousa (2024), assess ChatGPT's handling of Tunisian and Jordanian Arabic, revealing gaps in dialect understanding where both ChatGPT-3.5 and ChatGPT-4 perform variably, underscoring the necessity for dialectspecific model enhancements. Alkaoud (2024) contributes to this discourse by introducing a bilingual benchmark leveraging the General Aptitude Test (GAT), which illustrates substantial improvements in GPT-4's Arabic capabilities compared to its predecessors, offering insights into bilingual language processing. Furthermore, Lai et al. (2023) evaluate ChatGPT across seven tasks in 37 languages from low, medium, and high-resource families. Their findings suggest that ChatGPT matches or even

exceeds fully supervised state-of-the-art models, particularly in high-resource languages. Interestingly, they note that providing task descriptions in a high-resource language enhances performance for low-resource languages.

Additionally, Koubaa et al. (2024) develop ArabianGPT, a native Arabic LLM designed specifically to tackle the complex linguistic features of Arabic, demonstrating improvements in tasks such as sentiment analysis and summarization. Moreover, expanding the scope of multilingual studies, (Tonja et al., 2024) introduce EthioLLM and Ethiobenchmark—a multilingual suite and corresponding dataset for Ethiopian languages and English. This initiative significantly enhances resource availability and task performance, marking considerable progress in multilingual NLP for lowresource languages.