AL-QASIDA: Analyzing LLM Quality and Accuracy Systematically in Dialectal Arabic

Nathaniel R. Robinson[2] **Shahd Abdelmoneim**[1] **Kelly Marchisio**[1] **Sebastian Ruder**[1]
[1]Cohere, [2]Johns Hopkins University
nrobin38@jhu.edu, kelly@cohere.com

Abstract

Dialectal Arabic (DA) varieties are underserved by language technologies, particularly large language models (LLMs). This trend threatens to exacerbate existing social inequalities and limits LLM applications, yet the research community lacks operationalized performance measurements in DA. We present a framework that comprehensively assesses LLMs' DA modeling capabilities across four dimensions: fidelity, understanding, quality, and diglossia. We evaluate nine LLMs in eight DA varieties and provide practical recommendations. Our evaluation suggests that LLMs do not produce DA as well as they understand it, not because their DA fluency is poor, but because they are reluctant to generate DA. Further analysis suggests that current post-training can contribute to bias against DA, that few-shot examples can overcome this deficiency, and that otherwise no measurable features of input text correlate well with LLM DA performance.

1 Introduction

Large language models (LLMs) have transformed natural language processing (NLP) for many languages (Singh et al., 2024; Dubey et al., 2024). However these technologies often lack support for minority dialects and language varieties (Robinson et al., 2023; Zhu et al., 2024; Arid Hasan et al., 2024; Joshi et al., 2024; Chifu et al., 2024). Arabic, the fourth most spoken macro-language in the world with 420M speakers and official status in 26 countries (Bergman and Diab, 2022; Tepich and Akay, 2024), has a diversity of such varieties. Despite its prominence, Arabic has been historically set back in NLP due to its script, which lacked digital support until the 1980s (Parhami, 2019).

Many NLP tools support Arabic now, but often only Modern Standard Arabic (MSA); the numerous Dialectal Arabic (DA) varieties are often neglected. There are 28 microlanguages designated as Arabic in ISO 639-3,[1] of which MSA is only one. In the Arab world, MSA is used only in narrow circumstances, while local DA varieties are predominant (Bergman and Diab, 2022; Ryding, 1991). MSA has no native speakers according to Ethnologue;[2] Arabic speakers speak their DA varieties natively (as L1) and later learn MSA as an L2 (Azzoug, 2010). Many who lack educational resources are not proficient in MSA (Bergman and Diab, 2022). Arabic varieties are diverse and differ both phonologically, morphologically, syntactically, semantically, and lexically (Habash, 2010; Keleg et al., 2023). According to Bergman and Diab (2022), Moroccan Arabic or Darija (ary) and Egyptian Arabic (arz) are as mutually intelligible as Spanish and Romanian. Table 1 illustrates two simple sentences that display 0% word overlap across three Arabic varieties. Bergman and Diab (2022) have urged researchers to move beyond treating Arabic as a "monolith," i.e. aiming only for MSA support.

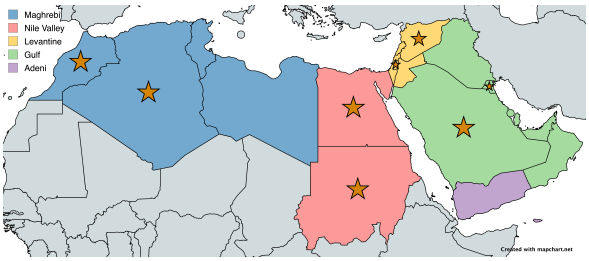

Figure 1: Arabic greater dialectal regions, per Diab and Habash (2007). Stars indicate the eight nations whose DA varieties are represented in this work.

Many LLMs today are proficient in MSA but reluctant to model DA; users wishing to converse in DA often have to perform extensive prompt acrobatics to coerce the LLM to use the dialect.[3] Be-

[1] https://wikipedia.org/wiki/ISO_639_macrolanguage
[2] https://www.ethnologue.com/
[3] For instance, when conversing with ChatGPT in Egyptian

Findings of the Association for Computational Linguistics: ACL 2025, pages 22048–22065

MSA	أريد أن أخبرك بشيء جيد جداً	لم أرَ كوب الماء هذا
Egyptian DA	عايز اقلك حاجة كويسة أوي	مشفتش كباية المية دي
Jordanian DA	بدي قولك اشي كتير منيح	ما شفت هاي الكاسة للماي
English	I want to tell you something very good	I didn't see that glass of water

Table 1: Example Arabic sentences with 0% word overlap across three varieties

cause MSA proficiency correlates with educational settings (and as a result socioeconomic advantage), such disparity may exacerbate existing inequalities. Even when interacting with MSA-proficient users, replying to informal DA inputs with formal MSA sounds unnatural and limits an LLM's uses.

LLM pre-training data is diverse and likely includes content in many Arabic dialects. However, post-training uses existing human-labeled data, which is typically MSA, and newly collected data that commonly adopt an official tone corresponding to a formal MSA register. Current LLM behavior suggests that models have the capability to understand and model DA, but that they default to MSA regardless, which could be caused by biases introduced in post-training. We thus hypothesize:

Hypothesis 1. *Current post-training methods make LLMs more reluctant to model DA.*

Hypothesis 2. *Post-trained LLMs understand DA better than they generate it.*

Arabic-speaking communities are aware of LLMs' DA shortcomings, but lack a standard operationalized definition of LLM DA proficiency. To this end, we present an evaluation suite of DA proficiency along different dimensions. We contribute:

1. AL-QASIDA:[4] a method to evaluate LLM DA proficiency along dimensions of **fidelity**, **understanding**, **quality**, and **diglossia**.
2. Best practice recommendations, including use of few-shot prompts for DA generation, Llama-3 for monolingual tasks, and GPT-4o for cross-lingual requests of Egyptian or Moroccan varieties.
3. The following six key findings from our AL-QASIDA evaluation and analysis.

These findings include:

1. **LLMs do not produce DA as well as they understand it** (as Hyp. 2). This is an apparent reversal of the Generative AI Paradox (West et al., 2024), which observes that models' generative capabilities exceed their ability to understand the generated output.
2. When LLMs do produce DA, they do so without perceptible declines in fluency.
3. LLMs are not diglossic: they generally cannot translate well between MSA and DA.

And further analysis suggests the following:

1. Post-training can bias LLMs against DA (see Hyp. 1), but otherwise improves text quality.
2. Few-shot prompting improves DA proficiency across dialects and genres.
3. Otherwise, no input text features correlate strongly with LLM DA performance.

We release software and data to conduct AL-QASIDA evaluations, as well as supplementary material, on our project repository.[5]

2 Related Work and Background

There exist notable prior works on benchmarking NLP for DA varieties. AraBench (Sajjad et al., 2020) is a benchmark for MT between Arabic varieties and English that predates LLMs' widespread popularity. DialectBench (Faisal et al., 2024) is a benchmark of 10 text-based traditional NLP tasks—such as parsing, part-of-speech tagging, and named entity recognition—that covers a large number of dialects and varieties in various languages, including multiple DA varieties. ARGEN (Nagoudi et al., 2022) compares mT5 (Xue et al., 2021) with novel AraT5 de-noising language models across MT, summarization, paraphrasing, and question generation tasks. Dolphin (Nagoudi et al., 2023) is a comprehensive Arabic NLP benchmark that evaluates DA, but does not define DA proficiency. AraDICE (Mousi et al., 2024) is an LLM benchmark much like ours that focuses on accuracy and cultural appropriateness in DA. Our work differs from these in its purpose: to define LLM DA proficiency. To our knowledge, ours is the first evaluation to measure this comprehensively.

Arabic, it took one of the authors 14 turns of conversation before the LLM used the desired dialect (see Table 5).

[4] **A**nalyzing **L**LM **Q**uality and **A**ccuracy **S**ystematically **I**n **D**ialectal **A**rabic. "Al-qasida" also means "the poem" in Arabic.

						ALDi
يا ريتني ما تفرجت عالحلقة الأخيرة تاعت رحيم تأزمت بصراحة أبدع ياسر جلال …	✓	✗	…	✓	✗	High (1)
هوا ورياح الشبابيك والابواب بتخبط كانه زلزال ربنا يستر	✗	✓	…	✗	✓	Med. (.7)

Table 2: One sentence is valid in Algerian and Tunisian DA, while another is valid in Palestinian and Egyptian, with varying ALDi (dialectness). Adapted from nadi.dlnlp.ai. See also Table 1 of Abdul-Mageed et al. (2024).

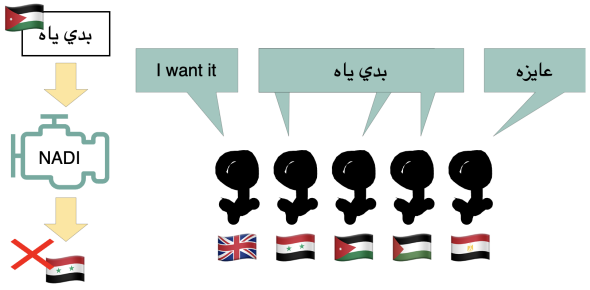

Figure 2: A sentence shared across Syrian, Jordanian, and Palestinian varieties may be labeled as Jordanian but predicted as Syrian, resulting in a false NADI error.

2.1 Arabic Dialect Identification

Identifying Arabic varieties, called Nuanced Arabic Dialect Identification (NADI), has been researched for years. Until recently, NADI shared tasks (Bouamor et al., 2019) evaluated models that produce a single country-level or city-level dialect label for each input sentence. However, this approach was found insufficient for DA intricacies. Arabic varieties have significant overlap, especially in text for geographically proximate varieties (Abdul-Mageed et al., 2024). For example, Figure 2 illustrates a sentence that is valid in multiple Levantine countries. If a NADI model labeled such a sentence as Syrian while the ground-truth label were Jordanian, this would be falsely deemed an error. In fact, Keleg et al. (2023) found that 66% of a single-label NADI model's supposed errors were not errors at all. Hence the most recent NADI shared task (Abdul-Mageed et al., 2024) focused on multi-label NADI: mapping an input sentence to a multi-hot vector label, as in Table 2.

Keleg et al. (2023) marked another change in standard NADI approach. Instead of treating MSA as an additional variety alongside DA varieties (Salameh et al., 2018), Keleg et al. framed MSA identification as a separate task: Arabic Language Dialectness (ALDi). ALDi models predict where a given utterance falls on the dialectness scale, with a $\text{score}_{\text{ALDi}}$ of 0 being fully MSA and 1 being fully DA (regardless which DA variety). See Table 2.

[5] https://github.com/JHU-CLSP/al-qasida

3 Methodology

To be proficient in DA, an LLM requires different competencies that we define as the following:

1. **Fidelity:** Can the LLM identify and produce the correct DA variety when asked?
2. **Understanding:** Does the LLM understand prompts in the DA variety?
3. **Quality:** Is the LLM able to model the DA variety well? Or does its quality deteriorate compared to MSA or another language? (We use the term "quality" to encompass both *fluency* and *semantic accuracy*.)
4. **Diglossia:** Can the LLM translate between DA and MSA?[6] (We consider this meaningful, since ability to translate thus indicates a model's understanding of both MSA and the DA variety.)

Fidelity is crucial as it forms the prerequisite for further assessment. **Understanding** of DA prompts and **Quality** of DA responses are both necessary for successful user interactions. Finally, **Diglossia** measures whether an LLM is aware of fine-grained differences between DA and MSA.

3.1 Operationalizing Fidelity

To measure how well the LLM can identify and produce requested DA varieties (**Fidelity**), we evaluate whether the LLM produces the desired DA variety in monolingual (prompting the LLM in a specific DA variety) and cross-lingual (requesting a specific DA variety from the LLM in English) settings, analogous to Marchisio et al. (2024).

To this end, we require a NADI model to identify the Arabic variety of LLM outputs. As single-label NADI classifications can lead to false negatives in evaluation (Abdul-Mageed et al., 2024), as seen in Figure 2. To mitigate this, we apply the NADI 2024 shared task baseline model[7] (Abdul-

[6] We acknowledge this is not the conventional use of the term "diglossia." While an LLM perfectly fluent in Arabic could theoretically navigate between MSA and DA depending on the conversation topic, current LLMs are far from that level of fluency. We thus restrict our focus to simpler evaluations in this work.

[7] The model was originally trained for the single-label task

Dimension	Capability	Input lang.	Output lang.	Metric
Fidelity	Monolingual generation	DA	DA	ADI2 score
	Cross-lingual generation	eng	DA	ADI2 score
Understanding	Translation	DA	eng	spBLEU
	Instruction following	DA	DA	Human eval
Quality	Translation	eng	DA	spBLEU
	Fluency	DA/eng	DA	Human eval
Diglossia	Translation	MSA	DA	spBLEU
	Translation	DA	MSA	spBLEU

Table 3: Evaluation data and metrics for the four competencies to assess DA proficiency in LLMs. Data are in dialectal Arabic (DA), Modern Standard Arabic (MSA), or English (eng).

Mageed et al., 2024), but extract the probability of the desired dialect from the output logits rather than using the one-hot classification output. The NADI score ($\text{score}_{\text{NADI}}(\cdot)_C$) of an LLM is then the probability that its output is in the desired DA variety of country C. As NADI and ALDi are independent (see §2.1), this NADI model does not distinguish between MSA and DA: only between country-level varieties. We therefore employ an additional model for ALDi: to differentiate MSA and DA (Keleg et al., 2023). We finally define a DA fidelity performance metric that combines both NADI and ALDi scores: Arabic Dialect Identification And DIalectness (ADI2) score. Given LLM output y,[8] we define:

$$\begin{aligned}\text{score}_{\text{ADI2}}(y) &= P(y \text{ is dialectal } C \text{ Arabic}) \\ &= P(y \text{ is DA}, y \text{ is } C \text{ Arabic}) \\ &= P(y \text{ is DA}) * P(y \text{ is } C \text{ Arabic}) \\ &= \text{score}_{\text{ALDi}}(y) * \text{score}_{\text{NADI}}(y)_C\end{aligned}$$

Because some DA varieties are similar (see §2.1), we compute an alternative to $\text{score}_{\text{NADI}}(\cdot)_C$, which we call the NADI macro-score: $\text{macro}_{\text{NADI}}(y)_R := \sum_{C \in R} \text{score}_{\text{NADI}}(y)_C$, where $R \in \{\textbf{Maghreb}, \textbf{Nile Valley}, \textbf{Levant}, \textbf{Gulf}\}$, with each region R itself representing a set of countries C. Accordingly,

$$\text{macro}_{\text{ADI2}}(y)_C := \text{score}_{\text{ALDi}}(y) * \text{macro}_{\text{NADI}}(y)_R$$

, given $C \in R$. This metric indicates whether the LLM responds with DA varieties from the correct region, and thus allows for NADI confusions between proximate varieties.

3.2 Operationalizing Other Competencies

Understanding We use DA-to-English translation to evaluate the LLM's understanding of DA text. As English is the closest language to an LLM's internal representation (Etxaniz et al., 2024), we can assess to what extent the model understood the original DA text based on the quality of its English translation. Machine translation (MT) metrics such as BLEU (Papineni et al., 2002) or human annotations thus serve as proxies to DA understanding. We employ SpBLEU (Goyal et al., 2022) and chrF (Popović, 2015). In addition to assessing **Understanding** via MT, we ask human annotators to evaluate understanding directly by determining to what extent the LLM fulfilled requests in monolingual DA prompts during the **Fidelity** evaluation.[9]

Quality We next measure LLMs' *fluency* and *semantic accuracy* in DA, first by English-to-DA MT. Once more, we treat English as a proxy for the LLM's semantic knowledge; this measures how well the LLM can express a variety of semantic concepts in DA. We supplement automatic MT scores with human evaluations of adequacy and fluency of translations, as well as dialectness judgments, detailed in §4.2. We couple this with correlative evaluation to reveal whether the LLM's output quality deteriorates in DA compared to MSA. For this we gather all human DA fluency annotations previously collected and correlate them with dialectness scores, to determine whether fluency degrades in more dialectal generation. (See §4.2.)

Diglossia To measure LLMs' diglossic proficiency, we evaluate MSA↔DA MT. Table 3 summarizes details of all four evaluation competencies.

3.3 Evaluation Corpora

In addition to NADI and ALDi models, we require three specialized corpora for evaluation: (1) **Cross-lingual** prompts, or English user inputs explicitly requesting specific DA varieties; (2) **Monolingual** prompts, or user requests in various DA varieties;

of Abdul-Mageed et al. (2023).

[8] Because some LLMs, especially non-post-trained or "base" models, tend to repeat DA inputs in monolingual settings, we remove any copies of the prompt from output y before ADI2 computation.

[9] We rely on humans as LLM-as-a-judge evaluators (Zheng et al., 2023) are less accurate when rating responses in low-resource varieties such as DA.

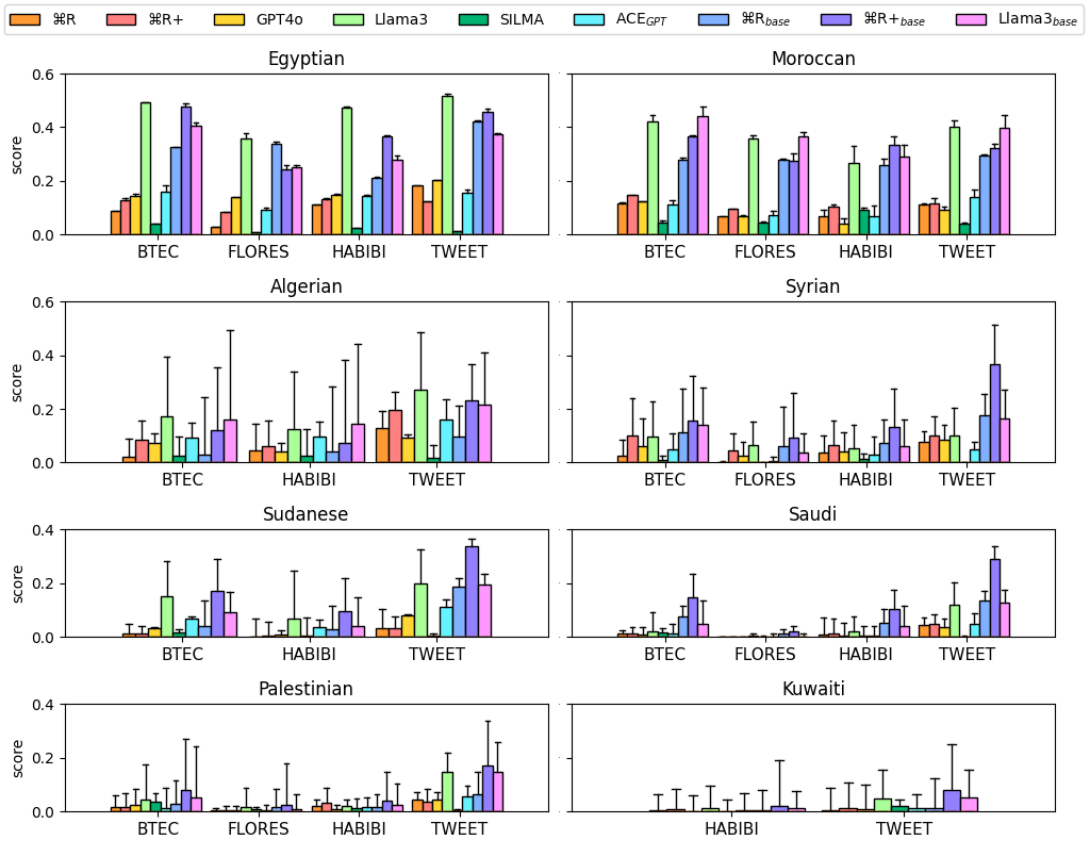

Figure 3: Llama models and Command series base models are best at maintaining the user's DA variety, as measured by ADI2 score (bars) and macro-score (whiskers).

and (3) **Bitext** prompts, or aligned bitexts with translations in English, MSA, and DA varieties.

For the **cross-lingual** corpus, we adapted a set from Marchisio et al. (2024) of English LLM prompts with explicit requests for responses in different languages, and substituted the names of DA varieties. This set originally came from three distinct collections of LLM user prompts: a subset of Okapi (Lai et al., 2023) inputs to the Alpaca LLM (Taori et al., 2023), a collection of ChatGPT inputs scraped from the ShareGPT API, and a corpus of human-curated prompts commissioned by Marchisio et al. (denoted *Cohere*). For the **monolingual** and **bitext** corpora, meanwhile, we selected four existing DA data sets based on their style diversity and dialectal coverage.

We used two multi-variety DA bitext corpora, integrating 200 sentence pairs from each in our **bitext** corpus and 100 sentences from the monolingual DA portion of each in our **monolingual** corpus. The first, MADAR-26 (Bouamor et al., 2018), is a multi-way parallel bitext with English, MSA, and DA from 25 Arab League cities. The English source sentences were sourced from the Basic Travveling Expression Corpus (BTEC) (Takezawa et al., 2007) and manually translated into the 26 Arabic varieties.[10] The genre of this corpus is BTEC, i.e. everyday utterances that might be expressed verbally. The second set, FLORES-200 (NLLB Team et al., 2022), is an MT evaluation benchmark of 1012 sentences in 204 language varieties. The English source texts were sampled from wiki sites and then translated manually. We use the sets for Najdi, North Levantine, South Levantine, Egyptian, and Moroccan Arabic to represent KSA, Syria, Palestine, Egypt, and Morocco, respectively.

We then used two multi-variety monotext DA corpora, adding 100 additional sentences per variety from each to our **monolingual** corpus. The first, NADI-2023-TWT (Abdul-Mageed et al., 2023), contains 23.4K tweets in 18 distinct DA varieties. The tweets were randomly selected from a larger tweet collection (Abdul-Mageed et al., 2020), with labels originally assigned via systematically extracted geographic information. The second corpus, HABIBI (El-Haj, 2020), contains 30k song lyrics

[10]We used sets for Riyadh, Damascus, Jerusalem, Khartoum, Cairo, Algiers, and Fes to render seven countries' DA.

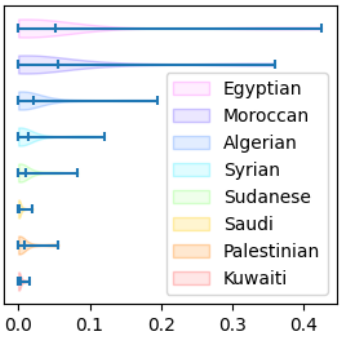

Figure 4: ADI2 (correct-variety dialectness scores) distributions across LLMs and genres in the crosslingual task (which requests specific DA varieties of the LLM in English). ADI2=0 indicates the wrong Arabic variety.

by artists from 18 Arab countries.[11] We used eight country-level subsets from each of these collections (corresponding to all eight countries in Figure 1).

Note that the purpose of **monolingual** evaluation is to measure how well an LLM matches a user's input DA variety, so it should be composed of prompts instruction-tuned LLMs are accustomed to. The data sources we used for monolingual sentences provided generic sentences of four genres: BTEC, wiki text, tweets, and song lyrics. To transform these diverse generic sentences into instructions, we used eight instruction templates (shown in Table 7). We surveyed native speakers of the eight DA varieties covered to retrieve translations of the templates. Then for each generic sentence, we randomly selected one of the templates in the appropriate variety and inserted the generic sentence, transforming it into a command. Also randomly chosen along with the template was the sentence location—whether it was inserted at the start, middle, or end of the template (see Table 7).

4 Evaluation and Results

We detail our evaluation of nine LLMs for eight country-level DA varieties: Kuwait, Saudi Arabia, Syria, Palestine, Sudan, Egypt, Algeria, and Morocco. These constitute two varieties from each of four Arabic dialectal regions (see Figure 1).

4.1 Primary Results

Recall from §3 that we evaluate **Fidelity**, via ADI2 score and macro-score; and **Understanding**, **Quality**, and **Diglossia**, via DA↔English and DA↔MSA MT. We begin with **Fidelity** and present scores across eight dialects, four genres, and nine LLMs including six general-purpose open-source LLMs—Command-R[12] (and base model[13]), Command-R+[14] (and base model), and Llama 3.1 (and base model) (Dubey et al., 2024); one closed LLM—GPT-4o;[15] and two LLMs specialized in Arabic—ACEGPT (Huang et al., 2024) (selected because of its prominence) and SILMA (Team, 2024) (selected because it led the Arabic LLM leaderboard[16] during our evaluation). Results for monolingual and cross-lingual prompts are in Figures 3 and 4, respectively (more detailed results in Appendix A).

Regarding **Fidelity**, the order of model performance is roughly consistent across genres and dialects. In **monolingual** settings, the Command-R+ base model, Llama-3.1, and Llama-3.1 base model performed best,[17] followed by Command-R base. Command-R, Command-R+, GPT-4o, and ACEGPT typically perform worse, and SILMA worst of all (see Fig. 3). Almost all ADI2 scores fell below 50%. The Command base models' higher performance supports Hypothesis 1 that post-training can inhibit DA modeling. All LLMs performed poorly on the **cross-lingual** task. Base models, ACEGPT, and Llama 3.1 were unable to complete the cross-lingual task whatsoever, and the performance of Command-R, Command-R+, and SILMA was low. GPT-4o was the only model to surpass 13% ADI2, and only on half of the dialects. We summarize results in Figure 4.[18] Note that on the lowest performing DA varieties, no LLMs exceed 20% ADI2 on any genre for either task, and that even on Egyptian and Moroccan, a majority of LLMs still score below this threshold.

As Figure 3 shows, LLMs are generally more able to model higher-resourced DA varieties, like Egyptian and Moroccan, than lower-resourced ones, like Kuwaiti. Note also in the Gulf dialectal region, home to seven country-level DA labels, the probability mass of the NADI score gets spread out more among the similar varieties, resulting in larger

[11] A native speaker of two DA varieties manually cleaned mislabeled sentences from the HABIBI sets we used, since some artists wrote in a DA variety other than their own. This annotator also deliberately selected lyrics for our 100-sentence subsets to highlight distinctions between varieties.

[12] https://cohere.com/blog/command-r

[13] I.e. the model checkpoint before post-training

[14] https://cohere.com/blog/command-r-plus-microsoft-azure

[15] https://openai.com/index/hello-gpt-4o/

[16] https://huggingface.co/spaces/OALL/Open-Arabic-LLM-Leaderboard, as of 27-10-2024

[17] impressively, given its 8B parameters (see Table 9).

[18] See Figure 10 for fuller results.

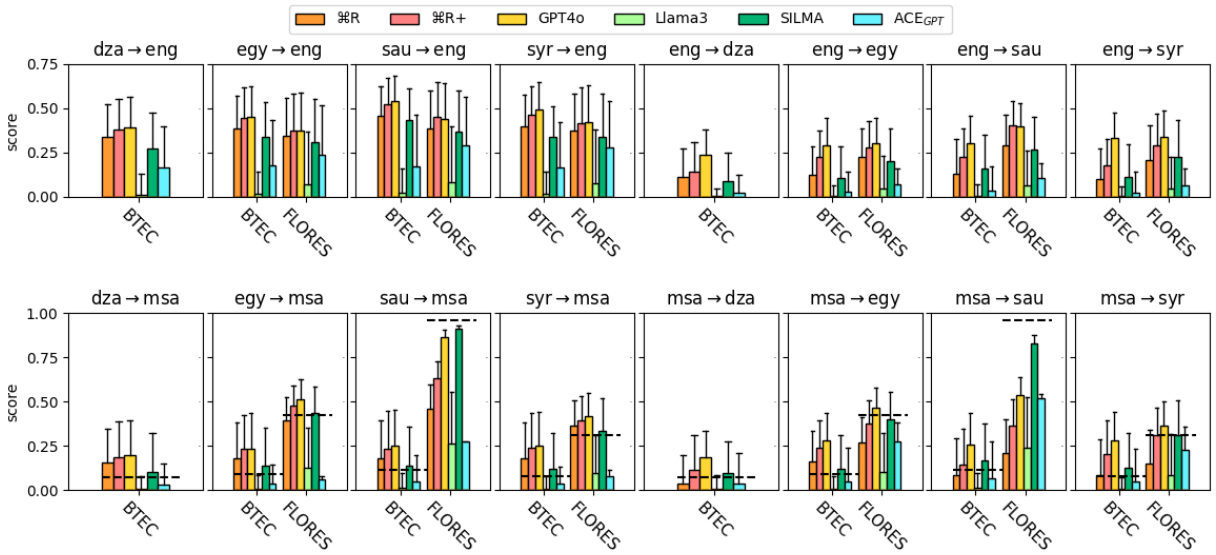

Figure 5: DA→Eng MT surpasses Eng→DA. DA↔MSA scores are low in the BTEC genre and rarely above the dashed-line zero-translate SpBLEU baseline for FLORES. Bars represent SpBLEU, while whiskers are chrF. Scores are between 0 and 1. (i.e. 0.5 corresponds to 50 SpBLEU points.) Note dza is the country code for Algeria.

differences between ADI2 score (bars in Fig. 3) and macro-score (whiskers). In the Nile Valley region, contrastingly home to only two country-level labels, the discrepancy is smaller (as seen best in the plot for Egyptian DA).

A majority of LLM responses fail because they are in MSA rather than DA: ALDi (dialectness) and ADI2 (overall) scores are highly correlated with $\rho = 0.99$ for cross-lingual and $\rho = 0.91$ for monolingual, indicating that if models do not respond in the right dialect, they are typically not responding in DA at all. Manual inspection of responses suggests that LLMs output pure MSA (with ALDi=0 or nearly 0) by default, while occasionally responding more dialectally. (Figure 12 illustrates this distribution.) We could not find any obvious predictors of the LLMs' dialectness from manual inspection. This led us to hypothesize that DA responses are not triggered by any features other than random sampling:

Hypothesis 3. *An LLM's distribution of correct DA responses does not correlate strongly with any detectable features of input text.*

Our MT automatic metric scores for four dialects are in Figure 5. (See Figure 11 for the rest.) GPT-4o and Command-R+ tended to perform best, with Command-R and SILMA usually lagging behind. ACEGPT typically performed worse and Llama 3.1 worse still. We forewent base models in this setting since MT is an instruction-oriented task. Model differences are more pronounced in directions evaluating **Diglossia** (DA↔MSA) and less so in those evaluating **Quality** and **Understanding**. Note that DA→English MT scores are higher than English→DA (supporting Hypothesis 2 that LLM DA understanding beats generation). MSA↔DA MT is poor for BTEC. Dashed lines indicate zero-translation baseline scores for MSA↔DA (i.e. the SpBLEU score between source and target corpora without translation). In many cases the LLMs actually pull the source farther from the target, even when scores may appear high (such as for MSA↔Saudi Arabic FLORES). Overall, LLM DA↔MSA performance is either low, below the zero-translation threshold, or barely above it. We conclude that **LLMs are not strongly diglossic.**

4.2 Human evaluation

We asked two native speakers of Egyptian and one of Syrian Arabic to make judgments of the Command-R+ and GPT-4o DA outputs for both MT and monolingual tasks, as well as Command-R+ base for the monolingual task. Annotators judged *Adherence* of monolingual responses, or how well the LLM fulfilled the user's request, and *Adequacy* of translations, or how well a translation reflected the meaning of the original sentence. For both tasks they made judgments of Arabic *Fluency* and *Dialectal Fidelity*, the final metric being the only one associated with the LLM's ability to use the right DA variety. We defined the scales and guidelines for each of these measurements as

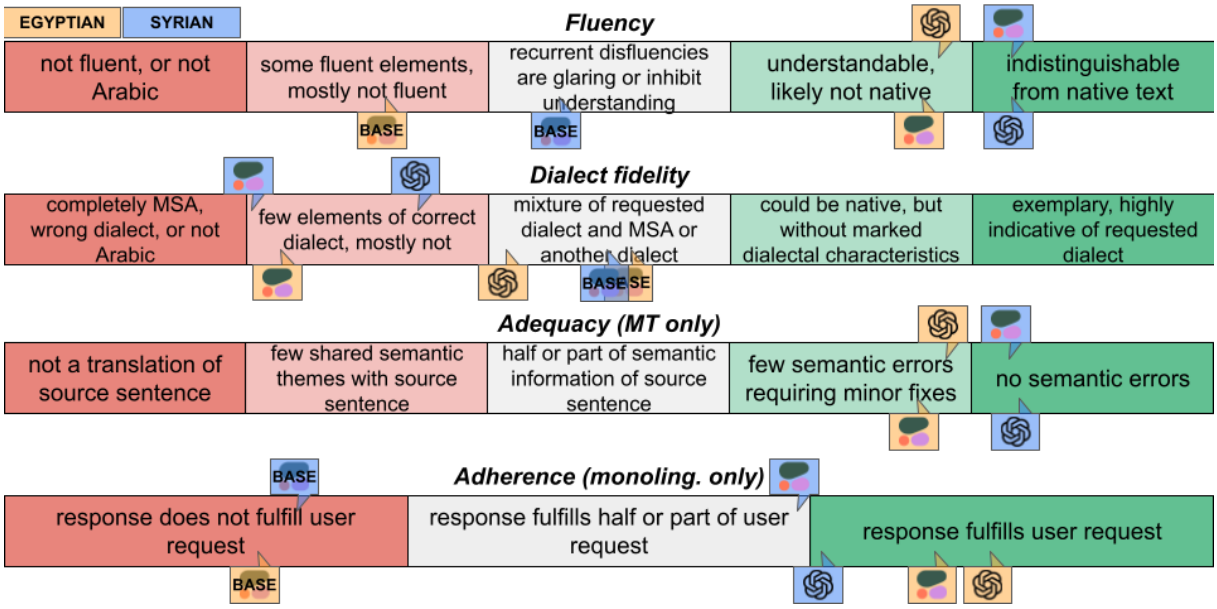

Figure 6: Human eval results shows that post-trained LLMs produce responses that are fluent, adequate, and adherent, but mostly not in the right DA variety. Command-R+ base improves dialectal fidelity but scores poorly on other metrics. Post-trained LLM fluency and fidelity scores were averaged across MT and monolingual tasks.

shown in Figure 6 and Table 10. We averaged human scores across 50 prompts for each model. Annotators received prompts and completions in random order, for an unbiased review.

See Figure 6 for average human eval scores. Results indicate that the instruction fine-tuned models excel at (1) producing fluent Arabic text, (2) translating into Arabic with semantic adequacy, and (3) fulfilling DA requests, but that they struggle to do so in the correct DA variety. Their high *adequacy* and low *dialectal fidelity* scores (along with low ADI2 scores and high DA→Eng MT scores in §4.1) indicate LLMs' **stronger DA understanding than generation**, supporting Hypothesis 2. This is an apparent reversal of the Generative AI Paradox (West et al., 2024), which highlights that LLMs' generative capabilities exceed their ability to understand their outputs. In contrast, we observe that LLMs are capable of understanding DA utterances but unable to produce fluent DA outputs.

Comparing base and post-trained models, Command-R+ base demonstrates better *dialectal fidelity*, in part due to prompt reduplication, but fulfills user requests poorly with low fluency. We thus conclude that **post-training can harm DA proficiency, but is needed for other aspects of text quality** such as instruction following ability.

We correlated human fluency scores with dialectal fidelity to ascertain whether LLMs model DA with diminished fluency. However not a single one

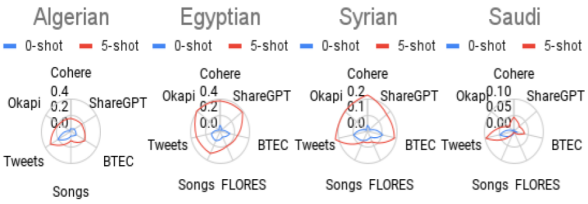

Figure 7: Command-R+ ADI2 scores from 5-shot prompting (red outer contours) are always higher than those of 0-shot prompting (blue inner contours).

displayed significant negative correlation. The only relationship with $p < 0.2$ was a weakly positive correlation of $\rho = .49$ for the Command-R+ base model in Egyptian. We thus conclude that **LLMs can produce DA with no perceptible decline in fluency** (when they are able to at all).

5 Follow-up Experiments and Analysis

After conducting our primary evaluations, which produced ADI2 scores and both automatic and manual MT scores for each LLM, we analyzed LLM outputs further. Our motivation for this was (1) to explore inexpensive remedies to LLMs' DA deficiencies, and (2) to test Hypothesis 3 and explore how user input features elicit LLM behaviors. We restrict this analysis to Command-R+, GPT-4o, and Llama 3.1 for depth and focus.

We first explored few-shot learning—an inexpensive mitigation approach—to improve DA modeling of the least performative of these LLMs:

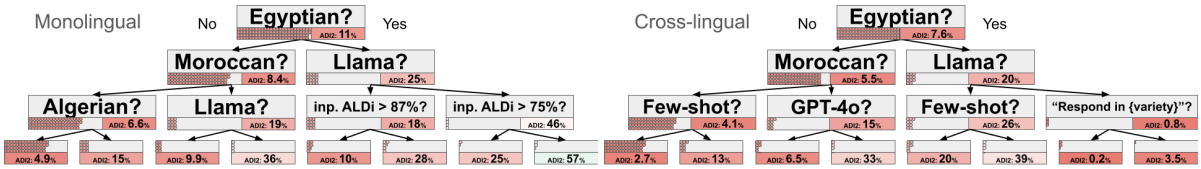

Figure 8: Decision trees regressed on ADI2 show that DA variety and LLM choice are primary dichotomous features. Each dot represents 100 samples, and color corresponds to group ADI2.

Command-R+. We curated five in-context prompt-completion examples for each of for DA varieties, for both monolingual and cross-lingual tasks, by translating the few shot examples used by (Marchisio et al., 2024) in their related experiments. Figure 7 shows that **few-shot examples improve ADI2 across tasks and genres** for Command-R+.

We then analyzed relationships between prompt attributes and performance. We collected features for each prompt and completion of the non-translation tasks: target DA variety; prompt length; prompt template; genre or data source; LLM used; number of few-shot examples; dialectness of input (for monolingual setting); and location of the dialect request (for cross-lingual) or inserted generic utterance (for monolingual) in the prompt. We fit decision tree regressors on ADI2 score using these features and `max_depth` 3, shown in Figure 8. The predominant dichotomous features are the desired DA variety and LLM used. Others include number of few-shot examples and dialectness of input text.

For feature-level correlations with ADI2 score, we performed Spearman's rank tests for numerical features and ANOVA tests for categorical features to find ρ and η^2 coefficients. We display these only for the relationships where $p < .001$, in Table 4, for both monolingual and cross-lingual tasks. No values in the table exceed Adams and Conway's (2014) threshold of strong correlation, $\eta^2 \geq .14$, and no ρ values exceed magnitude 0.5. It seems target DA variety, LLM used, and number of few-shot examples correlate moderately with ADI2; but other features have weak or no correlation. In the monolingual task, the dialectness of the prompt also correlates significantly but weakly with both ADI2 and output dialectness ($\rho = .29$ and $\rho = .31$, respectively).[19] **For a given LLM in a given DA variety, no feature of input text correlates strongly with DA performance.**

[19]High input dialectness can result in a range of output dialectness, but low input dialectness typically precludes high output dialectness. See Fig. 9.

	DAV	PT	LOC	GEN	LLM	LEN	N
			$\eta^2 =$			$\rho =$	
mono.	.115	.017		.012	.038	-.09	.12
cross.	.131		.003		.067		.32

Table 4: Significant ($p < .001$) correlations with ADI2. DAV=DA variety; PT=prompt template; LOC=DA request or generic utterance location; GEN=genre; LEN=prompt length; N=few-shot examples

6 Conclusion

We provide a comprehensive evaluation suite to measure LLM DA proficiency: Analyzing LLM Quality and Accuracy Systematically in Dialectal Arabic, or AL-QASIDA ("the poem" in Arabic). We find that LLMs struggle to model DA, not due to faults in understanding or generation quality, but to a preference for MSA generation. Though more balanced pre-training data could likely mitigate this, we find that post-training can bias LLMs towards MSA, suggesting balanced post-training data as a lower-cost alternative. We also find few-shot examples can mitigate DA pitfalls at even lower cost. In the absence of mitigation strategies, we recommend GPT-4o for cross-lingual DA requests of Egyptian or Moroccan, and Llama-3.1 for monolingual DA generation. (Base models also do relatively well but have serious fluency liabilities.)

Limitations

In this work we did not explore all the DA modeling mitigation strategies we had originally intended to. (We decided to leave these for future work when the number of our results started to become unmanageably large.) Notably, we found promising preliminary indications that strategic preambles can help LLMs model DA better.

Imaginably, we were unable to evaluate all popular LLMs and had to settle for relatively small a subset. We acknowledge this and emphasize that our development of this evaluation suite is meant as a technique for others to apply to other and future LLMs going forward. We did our best to include

a diversity of LLMs so that readers may extrapolate general trends of LLM DA proficiency, but we acknowledge our sample may not be entirely representative.

We also acknowledge that some of these LLMs, particularly the closed-source GPT-4o, are continuously updated and may not remain the same, even though the process of writing and publishing this paper. Hence, results drawn for GPT-4o may be slightly different as time passes and may not align perfectly with our findings here.

A last limitation of our work is the inability of our datasets to capture the realistic diversity of Arabic varieties. Many varieties differ primarily in pronunciation but look similar in text form; hence dealing only with text captures a small amount of the linguistic variation present in DA varieties. This variation also plays a role in some textual forms, as many DA varieties are primarily oral and have no standardized orthography systems. Models' difficulty modeling spelling inconsistencies will likely be a significant challenge to Arabic NLP in the future, though it was not a challenge we observed in our evaluations.

Ethics Statement

DA modeling capabilities have a number of ethical implications. As we touched on in §1, LLMs have the potential both to create opportunities for the less advantaged, and to exacerbate existing inequalities. If LLMs are primarily proficient in MSA, as they are today, they may afford benefits only to Arabic speakers with enough education and social advantage to communicate comfortably in MSA, while those with less MSA proficiency may be left behind.

The notion to treat Arabic as a monolith without representing its various diverse language varieties is often a result of homogeneity in the research community with a bias towards Western languages and values. We hope this work may bring a specific Arabic technological need to the community's attention as a small way to address this imbalance.

We cover a diversity of eight DA language varieties by our evaluation, however we cannot claim to represent the varieties of all Arabic speakers, even in the eight countries represented. We hope our evaluation may be expanded to be more representative in the future.

Lastly, we acknowledge that while all other evaluation sets we used were intended for this type of NLP evaluation, the song lyrics corpus we employed may have originally been intended for other purposes. Because most of these data sets were already curated for our general purposes, we did not vet them extensively for offensive content or sensative material. Our release of the song lyrics will not include any metadata (such as artist names), though such data can be found in its original source. Finally, we acknowledge use of LLMs to assist software development in this project, particularly in creating the graphics displayed in this paper.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The fourth nuanced Arabic dialect identification shared task. In *Proceedings of ArabicNLP 2023*, pages 600–613, Singapore (Hybrid). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The fifth nuanced Arabic dialect identification shared task. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 709–728, Bangkok, Thailand. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020. Toward micro-dialect identification in diaglossic and code-switched environments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5855–5876, Online. Association for Computational Linguistics.

Marc A. Adams and Terry L. Conway. 2014. *Eta Squared*, pages 1965–1966. Springer Netherlands, Dordrecht.

Md Arid Hasan, Prerona Tarannum, Krishno Dey, Imran Razzak, and Usman Naseem. 2024. Do large language models speak all languages equally? a comparative study in low-resource settings. *arXiv e-prints*, pages arXiv–2408.

Omar Azzoug. 2010. Modern standard arabic as a second language for its own speakers. *AL-Lisaniyyat*, 16(1):1–22.

A. Bergman and Mona Diab. 2022. Towards responsible natural language annotation for the varieties of Arabic. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 364–371, Dublin, Ireland. Association for Computational Linguistics.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani,

Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy. Association for Computational Linguistics.

Adrian-Gabriel Chifu, Goran Glavaš, Radu Tudor Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 1–15, Mexico City, Mexico. Association for Computational Linguistics.

Mona Diab and Nizar Habash. 2007. Arabic dialect processing tutorial. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Tutorial Abstracts*, pages 5–6, Rochester, New York. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco

Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Mahmoud El-Haj. 2020. Habibi - a multi dialect multi national Arabic song lyrics corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1318–1326, Marseille, France. European Language Resources Association.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. Dialectbench: A nlp benchmark for dialects, varieties, and closely-related languages. *arXiv preprint arXiv:2403.11009*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. AceGPT, localizing large language models in Arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural language processing for dialects of a language: A survey. *arXiv preprint arXiv:2401.05632*.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. ALDi: Quantifying the Arabic level of dialectness of text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. Association for Computational Linguistics.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Kelly Marchisio, Wei-Yin Ko, Alexandre Bérard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in llms. *arXiv preprint arXiv:2406.20052*.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2024. Aradice: Benchmarks for dialectal and cultural capabilities in llms. *arXiv preprint arXiv:2409.11404*.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-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, Ahmed 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, Singapore. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Behrooz Parhami. 2019. Evolutionary changes in persian and arabic scripts to accommodate the printing press, typewriting, and computerized word processing. In *Tugboat (Proc. 40th Annual Meeting of the TeX Users Group, Palo Alto, CA, August)*, volume 40, pages 179–186.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Karin C. Ryding. 1991. Proficiency despite diglossia: A new approach for arabic. *The Modern Language Journal*, 75(2):212–218.

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, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. Multilingual spoken language corpus development for communication research. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*, pages 303–324.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang,

and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Silma Team. 2024. Silma.

Anna Tepich and Ceylani Akay. 2024. The influence of arabic in an english classroom. *European Journal of Language and Literature*, 10(1):16–26.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, et al. 2024. The Generative AI Paradox: "What It Can Create, It May Not Understand". In *The Twelfth International Conference on Learning Representations*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Supplemental Figures and Graphics

Here we present some supplemental visualizations. Table 5 contains a conversation in which the user attempted to converse with GPT-4o in Egyptian DA, while the model repeated insisted on using MSA. Cell coloring corresponds to each utterance's dialectness per Keleg et al. (2023), listed in percentage form in the right column. Though the LLM's response on the sixth turn of conversation was partially DA (64% dialectness overall), it did not satisfy the user because of its composition: the first three words of the utterance were 93% dialectal, after which the LLM abruptly switched back into MSA (9% dialectal) for the rest. Not until the fourteenth turn of conversation was the LLM's response satisfactorily dialectal, at 75%. By comparison, the highest average dialectness score that any model achieved in our full evaluation presented here was only 61.4% (reached by Llama-3.1 on Egyptian Arabic tweets in the monolingual task). The original Arabic conversation is likewise displayed in Table 6.

Table 7 shows English versions of the templates we used to transform generic DA sentences into DA instructions (i.e. user inputs for an LLM). Native Arabic speakers from the eight countries indicated in Figure 1 translated the templates into their native DA varieties for our evaluation suite. Generic monolingual sentences from monotext or bitext corpora were inserted at different positions in the templates stochastically, see Table 8.

Supplemental and more complete experimental results can be found in Figures 10 and 11 Figure 11 visualizes MT scores as in Figure 5, but for all eight dialects evaluated. We will host complete detailed evaluation results on a server upon formal publication of this work.

Input dialectness is plotted against output dialectness in the monolingual task in Figure 9. The correlation between these variables is positive, with Spearman's $\rho = .31$. Notice that high input dialectness can result in a diverse range of output dialectness values, but low input dialectness typically precludes high output dialectness. In other words, even high input dialectness does not guarantee much about the output dialectness, but low input dialectness more or less destroys any chances of dialectal output.

We show the sizes of LLMs we used, in number of parameters, in Table 9. It is notable that Llama-3.1 scored best in the monolingual task, since we used only the 8B-parameter model, compared to much larger models Command-R+ and GPT-4o. We did not train any models and only performed inference.

For additional numerical scores, as well as LLM outputs from our evaluation, please see our project repository: `https://github.com/JHU-CLSP/al-qasida`. This repository also contains everything necessary to run AL-QASIDA.

👤	What's up man, how are you?	97%
🤖	I'm doing well, thank you for asking, and you? How are you?	19%
👤	All good, praise God. What are you up to today?	93%
🤖	I am engaging in assistance here via answering people's questions, presenting information, and helping to solve problems. How may I assist you today?	0%
👤	My goodness, very awesome! But why are you speaking MSA? Talk to me a bit in Egyptian.	88%
🤖	Sure, no problem. If you don't mind, is there anything specific you require assistance with today?	64%
👤	Yeah I just wanted a partner to speak together in the Egyptian dialect, the best Arabic dialect!	98%
🤖	Sure, I'm here to help you. That's a very beautiful thing, because I love communicating in the different dialects. If you have any question or need help with anything, just ask.	30%
👤	Great. But you're still speaking MSA a lot. I appreciate your efforts, but maybe you're not very skilled in the Egyptian dialect?	93%
🤖	I am here to help in any language you like, and naturally I try to the best of my abilities to speak in the prefered dialect of the interlocutor. But if you desire to communicate in the Egyptian dialect, then let us begin our conversation in the Egyptian dialect now. Is there something you would like to discuss or something to help you with?	0%
👤	Sure, just tell me about your interests now. What do you like to do?	84%
🤖	Certainly. I love to help people and to present accurate and useful information. Aside from that, I also enjoy reading books and examining the latest updates in the fields of technology and science. And as you know, I am here to help you in any matter that you need help in, so do not hesitate to ask any question.	0%
👤	You are still using totally MSA words. like "dhaalik" and "ayDan" and "aahkir". Talk to me in Egyptian from now on, boss. Try to tell me more about your interests and hobbies in the dialect please.	81%
🤖	You're right, man. Let's ditch the MSA and continue in Egyptian. I love to help people, and at the same time I like to read books and learn new things. For hobbies, I love to travel and to discover new places, and I also love nature and wandering in it. And in my free time I like to watch movies and listen to music. If there's anything you'd like to know or ask about, let me know, pal.	75%

Table 5: Arabic conversation with GPT-4o, translated into English. Percentages on the right represent dialectness scores from Keleg et al. (2023).

👤	عامل ايه يا حبيبى، ازيك؟	97%
🤖	أنا بخير، شكراً لسؤالك. وأنت؟ كيف حالك؟	19%
👤	كويس الحمد لله. عم تعمل ايه النهاردا بقى؟	93%
🤖	أقوم بالمساعدة هنا عبر الإجابة على أسئلة الناس وتقديم المعلومات والمساعدة في حل المشاكل. كيف يمكنني مساعدتك اليوم؟	0%
👤	رائع اوى ما شاء الله. بس عم تتكلم فصحى ليه؟ كمنى مصرى شويه	88%
🤖	طبعًا، مفيش مشكلة. بعد إذنك، هل يوجد أي شيء بالضبط تحتاج مساعدة فيه اليوم؟	64%
👤	ايوا بس كنت عايز شريك عشان نتكلم مع بعض باللهجه المصريه، احسن لهجه عربيه والله!	98%
🤖	طبعًا، أنا هنا لمساعدتك. ده شيء جميل جدًا، فأنا أحب التواصل باللهجات المختلفة. إذا كان لديك أي سؤال أو شيء تحتاج مساعدة فيه، ما عليك إلا أن تسأل.	30%
👤	تمام. بس لسه عم تتكلم فصحى كتير. بقدر محاولاتك، بس ممكن انو انت مش ماهر اوى باللهجه المصريه؟	93%
🤖	أنا هنا للمساعدة بأي لغة تفضلها، وبالطبع أحاول قدر المستطاع التحدث باللهجة التي يفضلها المحادث. لكن إذا كنت ترغب في التواصل باللهجة المصرية، فلنبدأ محادثتنا باللهجة المصرية من الآن. هل هناك شيء ترغب في الحديث عنه أو مساعدتك فيه؟	0%
👤	ايوا بس كمنى عن اهتماماتك دالوقت. تحب تعمل ايه بقى؟	84%
🤖	طبعًا، بكل سرور. أنا أحب مساعدة الناس وتقديم المعلومات الصحيحة والمفيدة. بجانب ذلك، أيضًا أستمتع بقراءة الكتب والاطلاع على آخر التطورات في مجالات التكنولوجيا والعلوم. وكما تعلم، أنا هنا لمساعدتك بأي موضوع تحتاج مساعدة فيه، فلا تتردد في طرح أي سؤال.	0%
👤	لسه عم تستعمل كلمات فصحى خالص زي آذلكْ وْأيضاً وْآخرْ. كمنى مصرى من دالوقت يا باشا. حاول تخبرنى اكتر عن اهتماماتك وهوياتك باللهجه دى لو سمحت.	81%
🤖	صح يا حبيبي، هونسيب الفصحى ونكمل بالمصري. أنا بحب أساعد الناس، وفي نفس الوقت بحب أقرأ كتب وأتعلم حاجات جديدة، بخصوص الهوايات، بحب السفر واستكشاف أماكن جديدة، وكمان بحب الطبيعة والتجول فيها. وفي الوقت الفاضي بحب أشوف أفلام وأسمع موسيقى. إذا في حاجة تحب تعرفها أو تسأل عنها، قولي يا زميل.	75%

Table 6: Arabic conversation with GPT-4o, in original form. Percentages on the right represent dialectness scores from Keleg et al. (2023).

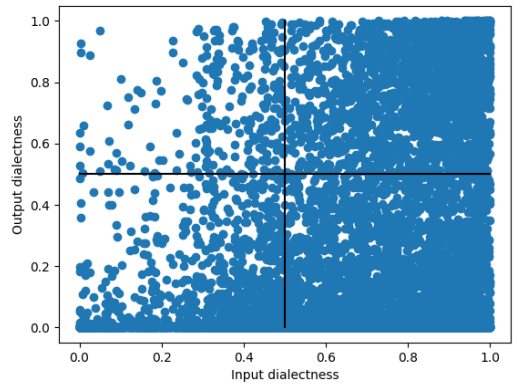

Figure 9: Prompt (input) and completion (output) dialectness values for monolingual evaluation; lines delineate quartiles

0	Paraphrase [this] in your own words, preserving the original meaning.
1	Craft a creative response that incorporates this sentence as a key element.
2	Write a story beginning with this line.
3	What are the implications of [this], if someone said it to you?
4	Rewrite [this sentence] to be more appropriate for general audiences.
5	What audience is this type of sentence intended for? Justify your answer.
6	Reply to this thread.
7	Draft a potential exam question, in an essay format, based on [this sentence.]

Table 7: Instruction templates to transform generic sentences into LLM commands. Brackets [] indicate the location of a sentence insertion, if inserted in the middle (rather than at the beginning or end).

beginning	[] Paraphrase this in your own words, preserving the original meaning.
middle	Paraphrase [] in your own words, preserving the original meaning.
end	Paraphrase this in your own words, preserving the original meaning: []

Table 8: Generic sentences may be inserted into any template at the beginning, middle, or end of a template (at any of the [] sites in this example). This variable was chosen stochastically for each generic sentence in creating the evaluation suite.

Command-R	35B
Command-R+	104B
Llama-3.1	8B
SILMA	9B
ACEGPT	7B
GPT-4o	?? (likely >175B)

Table 9: Sizes as parameter counts of the models used in this study

	Adherence
3	= the response fulfills user request completely
2	= the response fulfills half or part of the user request
1	= the response does not fulfill the user request at all
	Translation Adequacy
5	= no semantic errors
4	= a few semantic errors that require minor fixes
3	= half or part of the semantic information of the source sentence preserved
2	= a few shared semantic themes with the source sentence
1	= not a translation of the source sentence
	Fluency
5	= indistinguishable from native Arabic text
4	= understandable, but likely not native Arabic text
3	= clearly not native Arabic text, recurrent disfluencies are glaring or inhibit understanding (or includes copied text from the input prompt alongside newly generated text)
2	= some fluent elements, but mostly not fluent (or copies the input prompt without innovating)
1	= not fluent, or not Arabic
	Dialectal Fidelity
5	= exemplary Arabic highly indicative of the requested dialect (could be native)
4	= could be native, but does not display exemplary or marked dialectal characteristics of the requested dialect (i.e. could be a number of dialects)
3	= a mixture of the requested dialect and MSA or another dialect
2	= a few elements of the correct dialect, but mostly not
1	= completely MSA, clearly the wrong dialect entirely, or not Arabic
*	Note dialectal fidelty is not necessarily penalized when the LLM copies the input prompt.

Table 10: Human rating scales for different dimensions of output quality

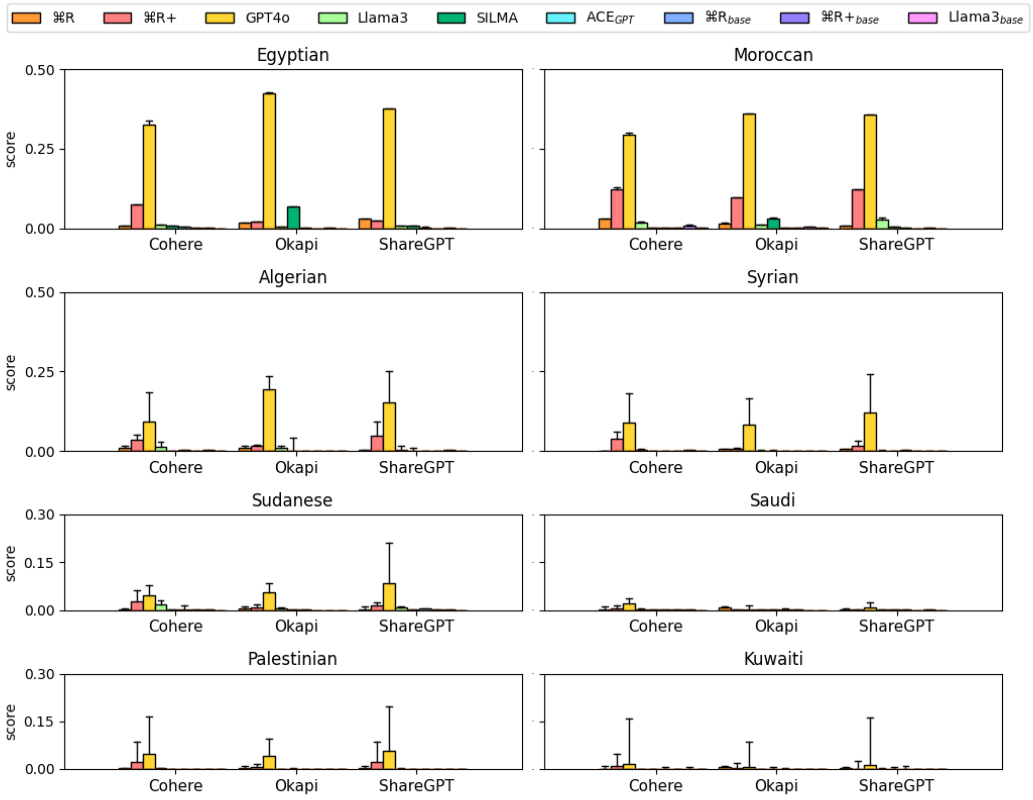

Figure 10: DA fidelity across genres, models, and dialects with **cross-lingual prompts**. Bars indicate ADI2 scores, while marks are macro-scores.

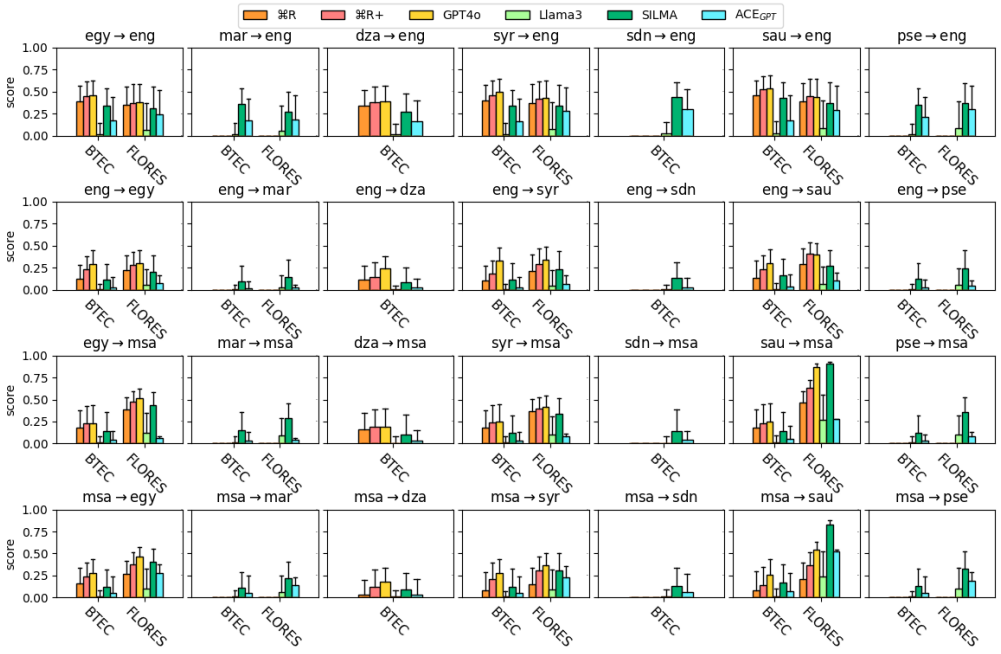

Figure 11: MT scores for all eight dialects. Compare to Figure 5. Bars represent SpBLEU, while marks are chrF.

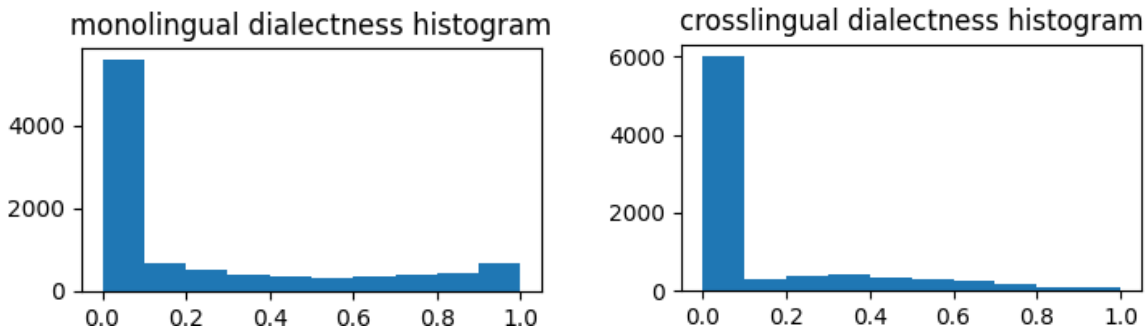

Figure 12: Distribution of ALDi (dialectness) scores. LLMs tend to produce standard Arabic (with ALDi≈0) most of the time, while occasionally venturing into more dialectal responses.