

AMIYA Shared Task: Arabic Modeling In Your Accent at VarDial 2026

Nathaniel R. Robinson¹, Shahd Abdelmoneim², Anjali Kantharuban³, Otba Alsoul,
Salima Lamsiyah⁴, Kelly Marchisio⁵, Kenton Murray¹

¹Johns Hopkins University, ²Cohere Labs Community, ³Carnegie Mellon University,
⁴University of Luxembourg, ⁵Cohere

Correspondence: nrobin38@jhu.edu, amiyasharedtask@gmail.com

Abstract

Arabic, often considered a single language, actually describes a wide variety of sometimes mutually unintelligible language varieties. While large language models (LLMs) have revolutionized natural language processing (NLP) with rapid advances, these models still best serve speakers of high-resource and standard language varieties. One particular deficiency of theirs is in dialectal Arabic. We present the first ever shared task for dialectal Arabic language modeling: **Arabic Modeling In Your Accent**, or AMIYA. The goal of the shared task was to develop LLMs that could (1) respond in the correct dialectal variety when explicitly or implicitly prompted to, (2) translate between dialectal Arabic and standard Arabic or English, (3) adhere to LLM instructions in dialectal Arabic, and (4) produce fluent Arabic outputs. We called for submissions in the dialectal varieties of five countries: Morocco, Egypt, Palestine, Syria, and Saudi Arabia. We received 45 submitted systems from six participating teams. We saw positive results from supervised fine-tuning on a translation objective, and reinforcement learning to improve dialectness. Manual evaluation also showed that some systems had learned to output dialectal words or phrases, but at the expense of actual fluency or coherence. Overall the most effective system involved continual pre-training and supervised fine-tuning of 12 candidate LLMs, followed by selection of the best performing models.

1 Introduction and Related Work

Recent advances in LLMs have demonstrated improvements in text generation in Modern Standard Arabic (MSA). However, performance remains limited when it comes to generating content in dialectal Arabic (Keleg et al., 2023). LLMs are typically proficient in high-resource language varieties, including Modern Standard Arabic (MSA) (Singh et al., 2024; Robinson et al., 2023), but they typi-

Monolingual Prompts

شئو كتقصد بهاد المثل: إذا طاحت البقرة كترو الجنوة؟



أه، فهمت دابا! كتقصد أن ملي كيوقع مشكل كبير، كل واحد كيبيدا يتقترح على لي واقع وكيفنا غير يغوت

Cross-Lingual Prompts

If you were a corporate law with 15 years of mergers and acquisitions experience, how would you pivot to launch an AI enable tech startup step by step and in detail? Explain in Moroccan Arabic.



أولا، خاصك دير دراسة السوق باش تعرف شئو هي الشركات لي كتخدم بزاف مع AI و شئو كيحجيبهم. من بعد، خاصك تأسس

Translation Prompts

MSA → DA	English → DA
DA → MSA	DA → English

Figure 1: Evaluation across three task settings: monolingual instructions, cross-lingual instructions, and translation (in four directions). Example outputs are Moroccan Arabic from Maastricht University’s (UM) primary submission, which achieved the highest human evaluation scores.

cally have poor or unreliable proficiency in dialectal Arabic (DA) language varieties.

DA varieties often lack support in NLP applications, in part because of researchers’ tendency to view Arabic as a "monolith" (Bergman and Diab, 2022). Contrary to this common perception, there are 28 different ISO 639-3 specifications¹ for distinct Arabic language varieties.

Robinson et al. (2025) found that while LLMs demonstrate a growing capacity to comprehend text in DA, they continue to struggle with generating dialectal text fluently. Specifically, many LLMs understand some DA varieties and are able to model them, but that they frequently fail to do so because of a strong preference for MSA. This is problematic for a number of reasons. While MSA is useful in many contexts, it is not used in many others.

¹https://wikipedia.org/wiki/ISO_639_macrolanguage

According to the Ethnologue,² MSA does not have native speakers. Comparatively, DA varieties like Egyptian and Moroccan Arabic (i.e. Moroccan Darija) have tens of millions of speakers. Many native speakers of DA varieties are not proficient in MSA, and those who are tend to be more educated and socially advantaged (Bergman and Diab, 2022). Hence LLMs’ singular proficiency in MSA may exacerbate, rather than ameliorate, existing inequalities. Robinson et al. (2025) suggest that different pre-training, post-training, and prompting methods may be used to mitigate LLMs’ difficulty with DA.

To address this gap, we designed the first shared task to evaluate and improve dialectal Arabic generation: **Arabic Modeling In Your Accent**, or AMIYA.³ For AMIYA we adopt the *AL-QASIDA* evaluation methodology (Robinson et al., 2025), which evaluates the dialectal fidelity, understanding, quality, and diglossia of LLMs when processing and generating DA. Our goal is to develop LLMs proficient in five major Arabic varieties, and provide the research community with tools and insights to bring about future developments.

2 Shared Task Overview

The AMIYA shared task evaluates LLM proficiency in Dialectal Arabic for varieties from five countries: Morocco, Egypt, Palestine, Syria, and Saudi Arabia. This evaluation is conducted using the *AL-QASIDA* methodology: both automatically and with human scoring on a subset of samples (Robinson et al., 2025). There are three available tracks: closed data, closed model, and open.

2.1 Closed Data Track

For the closed data track, the only permitted training data is provided by the organizers, but any open-source LLM can be used (along with their pre-trained weights). The training set for this submission track contains data from 17 different data sources, which we describe here.

MADAR 26 The MADAR 26 bitext contains 12k sentences of multi-way parallel training data across eight Arabic dialects, English, and Modern Standard Arabic (MSA) (Bouamor et al., 2018). It additionally includes 2k sentences for each of 26 city-level Arabic dialects. The corpus is high-quality, and is composed of translations of the En-

glish Basic Traveling Expressions Corpus (BTEC) (Takezawa et al., 2007). Its sentences are conversational and concise. The most recent test split⁴ was off limits to participants.

SauDial The Saudi Arabic Dialects Game Dataset is a multi-parallel bitext of 1001 lines from 4 Saudi Arabian dialects (Hejazi, Najdi, Eastern, and Southern), English, and MSA (Alanazi et al., 2025). The dataset focuses on topics related to localization for gaming and includes information on each games’ age rating and any relevant cultural context.

ASR-EgArbCSC The Egyptian Arabic Speech Corpus⁵ contains 5.5 hours of conversational Egyptian Arabic transcribed into 3.2k lines.

MASC Corpus The Multi-Arabic Sentiment Corpus includes 6k lines in three divisions: product reviews, political comments, and software comments (Al-Moslmi et al., 2018). Each sentence is labeled by two native speakers with positive or negative sentiment. Each annotator’s native country is included as well.

Maknuune Maknunne is an open lexicon for Palestinian Arabic with 36k entries from 17k lemmas and 3.7k roots (Dibas et al., 2022). Entries include the diacritic Arabic orthographic form, phonological transcription, and English gloss.

DODa The Darija Open Dataset contains ~50k sentences in Moroccan Arabic aligned with English translations (Outchakoucht and Es-Samaali, 2024). The dataset contains entries written both in Arabic- and Latin-based writing systems to reflect real-world usage patterns.

Shami Corpus The Shami Corpus contains approximately 10k tweets in each of four varieties: Lebanese, Palestinian, Jordanian, and Syrian (Abu Kwaik et al., 2018).

Atlaset The Atlaset dataset is a curated collection of Moroccan text data (Bounhar, 2025). The training set includes 155M tokens from sources such as social media posts and news articles. The dataset encompasses both formal and informal registers across domains.

²<https://www.ethnologue.com/>

³"Amiya," or "عامية" is the Arabic word for dialectal Arabic itself.

⁴corpus-6-test-corpus-26-test

⁵<https://magichub.com/datasets/egyptian-arabic-conversational-speech-corpus/>

SDC The Saudi Dialect Corpus contains social media data spanning various dialects of Arabic from Saudi Arabia (Tarmom et al., 2020). In total, the dataset includes 200k words.

Saudi Tweets Corpus The Saudi Tweets Corpus includes 200k tweets from Saudi Arabian users (Alruily, 2020). The dataset is cleaned of Twitter artifacts such as emojis, retweet markers, and hashtags.

SADSLyC The Saudi Arabian Dialects Song Lyrics Corpus contains 31k lines of dialectal song lyrics from the five major Saudi varieties: Najdi, Hijazi, Shamali, Janoubi, and Shargawi (Alahmari, 2025). The dialect of each song is ascertained using the hometown or birthplace of its author.

EDGAD The Egyptian Dialect Gender Annotated Dataset contains about 200k tweets, labeled with the gender of the writer (Hussein et al., 2019). Each tweet comes from a user with at least 1k tweets, and gender is determined via manual annotation, using profile information, tweet content, and external information (for public figures).

EDC The Egyptian Dialect Corpus (EDC) includes 200k words, across 13k lines, of Egyptian Arabic collected from Facebook (Tarmom et al., 2020).

Casablanca Casablanca is an Arabic speech corpus with transcriptions for samples in dialects from eight countries: Algeria, Egypt, United Arab Emirates, Jordan, Mauritania, Morocco, Palestine, and Yemen (Talafha et al., 2024). The sentences also include annotations on gender, dialect, and code switching.

JODA The Jordanian Dialect Arabic (JODA) dataset contains social media text, transcriptions of films, and existing dialect corpora (Abandah et al., 2025). These 50k sentences have been manually annotated by expert linguists and translated into MSA.

UFAL The UFAL parallel corpus includes 120k sentences from OpenSubtitles (Sellat et al., 2023). These sentences are translated into MSA and Northern Levantine Arabic, in addition to English, French, German, Greek, and Spanish.

Goud-Sum The Goud-Sum dataset is a set of 189k articles, with both headlines and categories (Aftiss et al., 2025). These articles are in Moroccan

Arabic and/or MSA, scraped from the GOUD news website.⁶

2.2 Closed Model Track

For this track, teams must train their LLMs from scratch—that is, pre-trained weights are not allowed. They can use any data except data that has been explicitly restricted for testing.

2.3 Open Track

For this track, teams may use any pre-trained, open-source LLM and any available data except that which is restricted for the test set.

2.4 Evaluation Methodology

Our evaluation set was composed of subsets of four datasets: Palm (Alwajih et al., 2025), FLORES-200 (Goyal et al., 2022; NLLB Team et al., 2022), Arena-Hard (Li et al., 2024), and MADAR-26 (Bouamor et al., 2018; Takezawa et al., 2007).

Palm is a dialectally diverse Arabic dataset composed of LLM prompts. We initially intended to use its designated test set for our evaluation. However, the test portion did not have enough data in each of our dialects of interest. Hence, we used portions of the Palm train set and repurposed them for evaluation. The plurality of data in the set are MSA, and we found through manual inspection that even many of the sentences labeled as a DA variety were also MSA. To mitigate this we filtered the sentences in our dialects of focus by using ALDi (Keleg et al., 2023) to measure their dialectness. We only kept sentences scoring above 0.5 (indicating that they are more dialectal than standard). A native Arabic speaker manually inspected a portion of the filtered sentences to verify that the automatic filtering was effective. This filtering yielded a total of 649 sentences for Moroccan Arabic, 283 for Egyptian, 419 for Palestinian, 181 for Syrian, and 141 for Saudi.

As DA LLM instructions, the Palm sentences were meant to evaluate LLMs’ monolingual generation ability to produce the same DA variety as the input. For additional data of this same kind, we used the first 200 sentences of FLORES-200 (NLLB Team et al., 2022) and the first 200 sentences of the MADAR-26 corpus⁶-train split (Bouamor et al., 2018) in each dialect. We converted these generic sentences to LLM instructions via AL-QASIDA, exactly as Robinson et al. (2025).

⁶<https://www.goud.ma>

To evaluate LLMs’ ability in cross-lingual generation—i.e., producing an explicitly requested DA variety to respond to an English prompt, we used 220 of the structured English LLM prompts from Arena-Hard (Li et al., 2024). We converted 110 of these into LLM instructions requesting a DA output by hand. We used Gemini⁷ to do the same to another 110 and manually verified that they were correct (with only minor error corrections needed). We made copies of these 220 total sentences, requesting each of our five DA varieties by simple string replacement.

Just as Robinson et al. (2025), we also availed the multi-dialectal multi-parallel nature of the FLORES-200 and MADAR-26 data to create evaluation sets for English→DA, DA→English, MSA→DA, and DA→MSA translation (with each input formatted to be a translation instruction). Figure 1 illustrates the three types of LLM prompts included in our evaluation.

We compute ADI2 dialectal fidelity score (Robinson et al., 2025) to evaluate all outputs that we expect to be DA, and chrF++ (Popović, 2017, 2015) to evaluate all translations.

In addition to these automatic metrics, native Arabic speakers manually evaluated a total of 100 outputs from each team’s primary system (25 from each of the data sources, excluding data intended for translation). Our evaluator for Palestinian and Syrian Arabic is a native speaker of Levantine Arabic from Jordan, who has lived many years in Syria. Our evaluator of Egyptian and Saudi outputs is a native speaker of both Egyptian and Gulf Arabic, and our evaluator of Moroccan outputs is a native speaker of Moroccan Arabic.

Our evaluation is patterned after the AL-QASIDA methodology (Robinson et al., 2025). We measure dialectal fidelity with ADI2 score for both monolingual, cross-lingual, and translation prompts. We measure understanding by chrF++ score for DA→English MT and by human-given adequacy scores. We measure generation quality by English→DA chrF++ and by human-given fluency scores. And we measure diglossia by chrF++ for MSA↔DA.

3 Shared Task Submissions

We received a total of 45 submitted systems from six teams.

⁷<https://gemini.google.com>

3.1 Aladdin-FTI

Aladdin-FTI (or Aladdin) (Mutal et al., 2026) participated in the closed data track with a system designed to jointly model dialectal fidelity and diglossic control. Their approach fine-tuned Hugging Face’s SmolLM3-3B⁸ and Llama-3-8B-Instruct⁹ using a multi-objective training method that combines (1) instruction-conditioned next-token dialectal generation and (2) machine translation between dialects, MSA, and English. The translation objective encourages semantic adequacy and awareness of MSA–DA boundaries, while the instruction-following generation objective promotes natural dialectal realization. By interpolating the two losses, the team aimed to balance dialectal fidelity and meaning preservation (e.g. in translation). Their system supports Moroccan, Egyptian, Palestinian, Syrian, and Saudi dialects. In their experiments they compared fine-tuning (1) purely for MT, (2) purely for instruction following, and (3) their combination. They found that training with the joint objective, by combining MT instruction data and standard instruction data, provides the best trade-off between dialect fidelity and translation quality. They submitted for all five dialects, using their system based on Llama-3-8B-Instruct as their primary submission and their system based on SmolLM as contrastive1.

3.2 Brigham Young University

BYU (Hamad and Al-Najjar, 2026) made a closed data track submission for Palestinian Arabic by fine-tuning on a combination of the provided Maknuune, Shami, Casablanca, MASC, and JODA corpora (Dibas et al., 2022; Abu Kwaik et al., 2018; Talafha et al., 2024; Al-Moslmi et al., 2018; Abandah et al., 2025). Their submitted system was based on Qwen2.5-1.5B-Instruct.¹⁰ They employed parameter-efficient fine-tuning, updating approximately 0.28% of the model’s parameters, and demonstrated improved performance on both translation and generation.

3.3 Maastricht University

Maastricht University (UM) (Alali and Issam, 2026) participated in the closed data track. Their

⁸<https://huggingface.co/HuggingFaceTB/SmolLM3-3B>

⁹<https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>

¹⁰<https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct>

system development involved three processes: Low Rank Adaptation (LoRA), adapter merging, and Minimum Bayes Risk (MBR) decoding (Hu et al., 2021; Bapna and Firat, 2019; Houshy et al., 2019; Bickel and Doksum, 1977; Kumar and Byrne, 2004). They used LoRA to train layer adapters on two contrastive objectives: (1) unsupervised fine-tuning with a language modeling objective on DA data, and (2) supervised fine-tuning with a translation objective. They found that merging monolingual and translation-based adapters improved the balance between dialectal fidelity (measured by ADI2) and semantic fidelity (measured by chrF++). They compared use of Llama 3.2 (Meta, 2024) and JAIS-2 (Sengupta et al., 2023) as their base models and found that while Llama 3.2 had superior ADI2, JAIS-2 was a stronger translator and had more balanced scores overall. As a final step they used MBR to ensemble different checkpoints, using ADI2 and chrF++ as objectives. They submitted systems based on JAIS-2 for Syrian, Moroccan, and Saudi dialects.

3.4 Mohammed Bin Zayed University of Artificial Intelligence

The MBZUAI team (Gaber et al., 2026) explored fine-tuning methods for the closed data track. They explored 12 different LLMs to use as an initialization, including some instruction-tuned models and some base models. The methodology was to fine-tune the base models using continual pre-training (CPT), then to fine-tune both base and instruct models with instruction fine-tuning. They explored both curriculum and mixed data approaches to handling DA, MSA, and English in the provided data for CPT. They created instruction tuning data from MADAR-26 data with an MT instruction template, and then took a subset of the same data and synthetically created instructions for monolingual and cross-lingual generation (using provided templates for the former and Command series models from Cohere for the latter). They tested which models performed best for each dialect off the shelf, then explored which models showed promise in fine-tuning on data subset. The six best overall candidate models were selected for full fine-tuning. They selected the three best performing models for each of the five dialects as their primary, contrastive1, and contrastive2 submissions.

3.5 National University of Singapore - Institute of Data Science

NUS-IDS (Das Gollapalli et al., 2026) made a closed data track submission by employing a unique training approach. Their initial step was to train LLMs for Arabic dialect identification and inter-dialectal translation. They then pass prompts from various data categories (including translation, summarization, and question-answering) through an LLM to construct a dataset for reinforcement learning (RL), i.e. preference tuning. If the trained dialect classifier finds the output to be the right dialect, it is labeled as a positive example. Otherwise, it is labeled as negative, and the trained translator creates a second output by translating into the correct dialect, which is labeled as positive. The team used this preference data for RL and found improvements in ADI2, at the expense of worsened translation metrics. Their primary submission for all five dialects was Llama-3.1-8B-Instruct¹¹ fine-tuned in this way. Their contrastive submissions used Qwen-2.5-7B-Instruct¹² (Yang et al., 2024; Team, 2024) post-trained as a translator; contrastive1 was trained with RL as well afterwards, while contrastive2 was not.

3.6 Syrian Dialect Arabic NLP

SDNLP (Alkhder and Abboush, 2026) was the only team to submit to the open track. They focused on Syrian Arabic adaptation based on the integration of parameter-efficient fine-tuning with prompt-guided inference. They used their own 30k-utterance dataset of caption data from Turkish television shows dubbed in Syrian Arabic. They employed Low-Rank Adaptation (LoRA) (Hu et al., 2022) to adapt a pretrained instance of Llama-3.1-8B to model Syrian Arabic.

4 Evaluation Results

We discuss results from both automatic and human evaluations across systems. Because all but one submission were in the closed data track, we compare all results together.

4.1 Automatic Scores

Results are in Tables 1 - 5. Best system results are **bold**, as are baseline results when they remained unbested. Table 1 shows overall ADI2 scores

¹¹<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

¹²<https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	37.8%	15.2%	7.4%	21.5%	4.9%
Aladdin-FTI	contrastive1	37.5%	15.1%	7.1%	24.5%	3.6%
BYU	primary	–	–	5.4%	–	–
MBZUAI	primary	56.9%	44.5%	9.5%	17.8%	10.1%
MBZUAI	contrastive1	54.0%	39.0%	8.6%	17.3%	8.5%
MBZUAI	contrastive2	57.5%	45.2%	10.1%	21.3%	14.2%
NUS-IDS	primary	70.1%	62.9%	12.2%	10.2%	7.3%
NUS-IDS	contrastive1	68.5%	56.0%	13.3%	36.2%	14.6%
NUS-IDS	contrastive2	32.0%	0.6%	1.3%	0.6%	0.7%
SDNLP	primary	–	–	–	28.0%	–
UM	primary	67.9%	–	–	38.9%	46.4%
Baseline		10.2%	11.1%	1.9%	1.3%	0.9%

Table 1: Average ADI2 scores across all datasets and tasks asking for DA responses. As [Robinson et al. \(2025\)](#), we use an LID filter that automatically assigns a score of 0 to sentences not identified as Arabic.

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	32.3	34.2	37.6	22.6	36.3
Aladdin-FTI	contrastive1	32.0	33.2	36.3	20.1	36.2
BYU	primary	–	–	13.8	–	–
MBZUAI	primary	22.7	33.6	36.9	31.0	34.1
MBZUAI	contrastive1	26.8	34.3	34.0	31.0	36.2
MBZUAI	contrastive2	25.8	31.2	30.5	25.7	32.6
NUS-IDS	primary	8.6	11.7	12.1	11.9	13.2
NUS-IDS	contrastive1	11.9	15.8	15.3	14.0	15.8
NUS-IDS	contrastive2	24.0	20.6	22.9	20.2	32.9
SDNLP	primary	–	–	–	14.8	–
UM	primary	30.0	–	–	34.4	19.8
Baseline		23.9	30.3	32.4	31.6	35.5

Table 2: Aggregate chrF++ scores for ENG→DA MT

for each submitted system. These are computed across all combined eval data subsets that expect DA responses: both monolingual prompts (Palm, FLORES, MADAR); cross-lingual (Arena-Hard); and MT (FLORES and MADAR English→DA and MSA→DA). In the Moroccan (mar), Egyptian (egy), and Palestinian (pse) dialects, NUS-IDS’ RL-tuned systems scored best, casting their unique preference tuning approach as a promising method for better dialectal fidelity. In Syrian (syr) and Saudi (sau), UM performed best. This was somewhat surprising since the only supervised data their model was fine-tuned on were MT prompts specifically, and could be an indication of the effectiveness of the unsupervised post-training they conducted prior to adapter merging.

The baseline model for automatic evaluations is Llama-3.1-8B-Instruct, run with each test prompt placed entirely in the "user" field of its input. We selected this baseline because it relatively performed

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	46.2	50.0	53.0	39.8	54.0
Aladdin-FTI	contrastive1	45.6	49.6	53.5	39.8	54.5
BYU	primary	–	–	11.2	–	–
MBZUAI	primary	44.9	47.7	50.4	54.1	58.4
MBZUAI	contrastive1	51.0	53.4	58.0	52.7	57.9
MBZUAI	contrastive2	44.6	43.3	50.0	48.2	52.2
NUS-IDS	primary	19.2	19.0	20.4	18.9	20.1
NUS-IDS	contrastive1	17.9	21.3	22.3	21.1	23.2
NUS-IDS	contrastive2	33.9	6.2	47.2	9.4	54.1
SDNLP	primary	–	–	–	36.2	–
UM	primary	49.9	–	–	51.9	0.0
Baseline		45.9	49.7	52.5	49.5	56.4

Table 3: Aggregate chrF++ scores for DA→ENG MT

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	26.6	28.8	31.2	23.2	43.2
Aladdin-FTI	contrastive1	25.1	27.5	30.7	19.4	32.1
BYU	primary	–	–	17.6	–	–
MBZUAI	primary	34.9	39.5	40.5	37.6	55.3
MBZUAI	contrastive1	35.3	43.7	40.1	36.8	56.9
MBZUAI	contrastive2	33.1	36.4	42.4	37.4	42.8
NUS-IDS	primary	11.6	17.3	16.0	16.6	19.7
NUS-IDS	contrastive1	11.2	16.5	15.5	15.7	17.5
NUS-IDS	contrastive2	32.4	12.5	30.9	9.7	60.1
SDNLP	primary	–	–	–	15.6	–
UM	primary	33.8	–	–	40.3	24.2
Baseline		30.8	42.1	40.4	39.8	61.4

Table 4: Aggregate chrF++ scores for MSA→DA MT

well on [Robinson et al.’s \(2025\)](#) evaluation. However, with this eval set and prompting method its dialectal fidelity is far below that of the submitted systems.

Table 2 shows chrF++ scores for MT of English (ENG) into DA. In this setting, Aladdin’s Llama-based system performed best on Moroccan, Palestinian, and Syrian; MBZUAI did best on Egyptian, and UM won out on Syrian. (Note also that best scores from each of these three teams are fairly close to one another across all five dialects.) Aladdin, MBZUAI, and UM (as well as NUS-IDS with their contrastive2 system) all trained on MT-specific instruction fine-tuning data and all scored fairly well on this task.

In the following tables, 3 for DA→ENG, 4 for MSA→DA, and 5 for DA→MSA, we see a common trend: MBZUAI’s systems perform best. The only exceptions to this are MSA-to-Syrian MT, which UM excelled at, and MSA-to-Saudi MT, which NUS-IDS did best (with their model

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	27.4	29.4	31.2	24.6	42.1
Aladdin-FTI	contrastive1	27.1	28.0	30.7	21.3	34.2
BYU	primary	–	–	17.3	–	–
MBZUAI	primary	39.3	43.9	41.6	40.3	65.3
MBZUAI	contrastive1	44.1	50.6	42.9	44.4	66.3
MBZUAI	contrastive2	37.9	39.1	41.8	34.6	53.2
NUS-IDS	primary	13.3	17.5	15.9	14.9	21.6
NUS-IDS	contrastive1	16.0	19.7	18.6	20.8	32.7
NUS-IDS	contrastive2	35.1	11.2	32.8	12.1	59.1
SDNLP	primary	–	–	–	6.7	–
UM	primary	39.5	–	–	43.4	37.2
Baseline		39.3	47.3	44.4	42.1	68.5

Table 5: Aggregate chrF++ scores for DA→MSA MT

Adherence level	Rating
The response fulfills user request completely	3
The response fulfills half or part of the user request	2
The response does not fulfill the user request at all	1

Table 6: Adherence: *did the model response fulfill all that was asked in the prompt?*

trained on an MT objective only). Specifically for the out-of-DA translation directions (Tables 3 and 5), MBZUAI’s Aya-based systems performed best; these include all of their contrastive1 systems and their primary systems for Syrian and Saudi DA. This could be in part because Aya (Dang et al., 2024; Üstün et al., 2024) was trained on a highly multilingual dataset, including explicit support for English and MSA (the target languages here).

Notice that across MT evaluations, many systems did not outperform the baseline, and even the best systems typically only outperformed it by small margins (if at all). However, we wish to point out that the baseline model’s outputs for into-DA translation, while they may achieve high chrF++, have extremely low ADI2 scores. And for into-MSA translation it may benefit from the unfortunate proximity between MSA and the DA subsets of FLORES. All of this said, we hope that in future years participants will be able to beat the baseline more convincingly in all metrics.

See Appendix B for more detailed, genre-specific score breakdowns.

Fluency level	Rating
The response is indistinguishable from native Arabic text	5
The response is understandable, but likely not native Arabic text	4
The response is clearly not native Arabic text, recurrent disfluencies are glaring or inhibit understanding (or includes copied text from the input prompt alongside newly generated text)	3
The response contains some fluent elements, but mostly not fluent (or copies the input prompt without innovating)	2
The response is not fluent, or not Arabic	1

Table 7: Fluency: *did the model generate a response that could have been written by a native Arabic speaker?*

4.2 Human Evaluation Scores

We detail our human evaluation process and present its results.

4.2.1 Evaluation Criteria

To assess the quality of dialectal Arabic generation, we conducted a comprehensive human evaluation with three native Arabic speakers, each proficient in multiple dialectal varieties. Our evaluators included a native Gulf/Egyptian speaker, a Levantine speaker, and a Moroccan speaker, providing coverage across all five dialects included in our study. The annotators, who are also authors of this work, evaluated each team’s primary submission on a sample of 25 prompt completions each from Palm, MADAR, FLORES, and Arena-Hard; or 100 samples total per dialect per team. Each completion was assessed along two dimensions that are difficult to measure automatically: model *Adherence* and *Fluency*.

Adherence refers to whether the model response fulfilled the original user request, using a 3-point scale, displayed in Table 6. *Fluency* refers to whether the model generated a response that appears to have been written by a native Arabic speaker, with coherent sentence structure (in either DA or MSA). We used a 5-point scale to capture the various degrees of fluency, shown in Table 7.

4.2.2 Annotator Results Overview

The resulting average scores of our human evaluation are shown in tables 8 and 9. One thing is fairly simple about the results trends: MBZUAI’s submission performs best in every measurement, with the exception of UM in Moroccan fluency. For *Adherence*, most models average between 1 (“does not

	mar	egy	pse	syr	sau
Aladdin	1.78	1.80	1.17	1.12	1.66
BYU	–	–	1.26	–	–
MBZUAI	2.04	2.07	1.61	1.62	2.10
NUS-IDS	1.40	1.92	1.17	1.18	2.05
SDNLP	–	–	–	1.12	–
UM	1.97	–	–	1.15	1.12

Table 8: Adherence: *average human score per team and dialect*

	mar	egy	pse	syr	sau
Aladdin	3.15	2.77	2.68	2.44	3.15
BYU	–	–	2.99	–	–
MBZUAI	3.20	3.59	3.48	3.43	3.37
NUS-IDS	1.43	2.22	2.40	2.46	2.83
SDNLP	–	–	–	2.93	–
UM	3.37	–	–	2.63	2.38

Table 9: Fluency: *average human score per team and dialect*

fulfill the user request at all") and 2 ("fulfills half or part of the user request"); and the best performing models average between 2 and 3 ("fulfills user request completely"). In terms of *Fluency*, most models average between 2 ("contains some fluent elements, but mostly not fluent") and 3 ("clearly not native Arabic text"); and the best models perform between 3 and 4 ("understandable, but likely not native Arabic text").

Our findings suggest that knowledge of dialectal lexicon does not necessarily translate to conversational competence in a particular dialect. Even when models demonstrate facility with dialectal vocabulary, they lack the cultural and contextual knowledge necessary for accurate and disambiguous discourse. High dialectal scores alone do not necessarily result in coherent and fluent responses, as models can parrot the original prompt or use a pool of dialectal words without meaningful sentence structure. For example, Aladdin’s system achieved high ADI2 scores (shown in Table 1), but scored poorly on *Adherence* in Table 8. This is a limitation that extends beyond purely linguistic competence to encompass broader understanding of DA varieties. This pattern aligns with Robinson et al. (2025), who report that LLMs understand Dialectal Arabic better at than they generate it, and reveal a reluctance to produce dialectal output along with insufficient depth of dialect-specific knowledge.

We include detailed observations from the annotators regarding each dialect in Appendix A.

5 Conclusions

One of our conclusions from our human evaluation was that command of dialectal vocabulary does not equate to conversational fluency, as models often lack the cultural and contextual structure required to use words appropriately in the native dialect. Models can achieve high ADI2 scores by parroting prompts or inserting dialectal terms without constructing meaningful, coherent sentences, which could be misleading as a signal of fluency.

Looking at our evaluation campaign comprehensively, it seems that the MBZUAI team performed best overall. Their systems, produced by selecting the best performing of 12 candidate model initialization after CPT and SFT with an MT objective, dominated most of the MT evaluations and nearly all the human evaluations, and they performed reasonably well in dialectal fidelity as well. This speaks to the effectiveness of a simple approach combining CPT and SFT, and the importance of casting a wide net in terms of candidate models. The MBZUAI teams’ thoroughness in exploring a wide variety of LLMs paid off in their performance.

We also point out that Aladdin’s systems performed well on into-DA translation, though they performed less well in adherence to generic user requests. This indicates that the model may have become an MT specialist, which is conceivable given that it was trained on a mixture including MT supervised data. Other submissions that occasionally stood out were UM and NUS-IDS. UM performed best or nearly best on all translation directions involving Syrian Arabic. They also achieved best or nearly best ADI2 scores across all three dialects they submitted for, and their model performed nearly as well as or better than MBZUAI’s for both *Adequacy* and *Fluency* in Moroccan. The RL-trained systems of NUS-IDS generally performed poorly in MT and human evaluations, but they scored best or nearly best in ADI2 score for every dialect except Saudi. This indicates that their novel method of preference tuning with synthetic correct-dialect and incorrect-dialect completions has promise for increasing ADI2 score. In this case, it appears to have degraded other abilities, but perhaps future researchers will find a way to combine it with other objectives and minimize this loss.

Limitations

Our evaluation campaign this year left a few things to be desired. To start, our suggested method for

participants to evaluate their own systems internally was by running [Robinson et al.'s \(2025\) AL-QASIDA](#). We found that this was less than practical for three reasons. First, participants struggled to set up the proper environments to run AL-QASIDA, and the code base was less adaptable to different kinds of models than we would have liked. For future evaluations it may be wise to set up a "lite" version of AL-QASIDA that simply computes ADI2 and chrF++ scores. The original implementation has model inference built in as an intermediate step, but perhaps it would be easier for teams to do their own inference separately. This would allow for more versatility. Second, AL-QASIDA is time-consuming to run (and essentially not runnable without access to GPUs). This is more difficult to address, but could also be alleviated somewhat by a "lite" version of the benchmark. Third, different conventions about how LLMs are prompted (whether with plain text input or a structured dictionary with roles) can vary. The original AL-QASIDA evaluation process was not equipped to account for this variation, which can lead to results that seem inconsistent with previously published scores.

Another shortcoming of our campaign this year lay in our use of the FLORES dataset. Unfortunately, some of the dialectal Arabic subsets of FLORES-200 are too close to MSA. The Saudi subset, for example, is closer to the Flores-200 MSA subset itself than even a reasonable set of MSA paraphrases would be. This makes results using FLORES somewhat questionable (and virtually ensures that models scoring high on Saudi ADI2 will score poorly on into-Saudi MT, and vice versa).

Acknowledgments

We acknowledge Amr Keleg, Sanjeev Khudanpur, and Samar Magdy for their feedback and advice for this shared task.

References

Gheith A Abandah, Moath R Khaleel, Iyad F Jafar, Mohammad R Abdel-Majeed, Yousef H Hamdan, Ashraf E Suyyagh, Asma A Abdel-Karim, and Shorouq M AlAwawdeh. 2025. Jordanian arabic to modern standard arabic translation using a large model tuned on a purpose-built dataset and synthetic error injection. *Jordanian Journal of Computers & Information Technology*, 11(3).

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. [Shami: A corpus of Levantine Arabic dialects](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Azzedine Aftiss, Salima Lamsiyah, Christoph Schommer, and Said Ouatik El Alaoui. 2025. Empirical evaluation of pre-trained language models for summarizing moroccan darija news articles. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 77–85.

Tareq Al-Moslmi, Mohammed Albared, Adel Al-Shabi, Nazlia Omar, and Salwani Abdullah. 2018. [Arabic senti-lexicon: Constructing publicly available language resources for arabic sentiment analysis](#). *Journal of Information Science*, 44(3):345–362.

Salwa Saad Alahmari. 2025. [SADSLyC: A corpus for saudi Arabian multi-dialect identification through song lyrics](#). In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 38–43, Abu Dhabi, UAE. Association for Computational Linguistics.

Abdulhai Alali and Abderrahmane Issam. 2026. Maastricht university at AMIYA: Adapting llms for dialectal arabic using fine-tuning and mbr decoding. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.

Naif Alanazi, Mohammed Al-Batineh, and Hussein Abu-Rayyash. 2025. Saudial: The saudi arabic dialects game localization dataset. *Data in Brief*, page 111906.

Hasan Alkhder and Mohammad Abboush. 2026. SDNLP at AMIYA 2026: Syrian arabic dialect modeling with lora. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.

Meshrif Alruily. 2020. [Issues of dialectal saudi twitter corpus](#). *The International Arab Journal of Information Technology*, 17:367–374.

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibae, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, and 25 others. 2025. [Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32871–32894, Vienna, Austria. Association for Computational Linguistics.

- Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1538–1548.
- 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.
- Peter J Bickel and Kjell A Doksum. 1977. *Mathematical statistics: basic ideas and selected topics*. Holden-Day Inc.
- Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouni, 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).
- Abdelaziz Bounhar. 2025. [Atlaset: A moroccan darija text corpus](https://huggingface.co/datasets/atlasia/Atlaset/). <https://huggingface.co/datasets/atlasia/Atlaset/>.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. [Aya expand: Combining research breakthroughs for a new multilingual frontier](#). Preprint, arXiv:2412.04261.
- Sujatha Das Gollapalli, Mouad Hakam, Mingzhe Du, and See-Kiong Ng. 2026. NUS-IDS at AMIYA/VarDial 2026: Improving arabic dialectness in llms with reinforcement learning. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.
- Shahd Salah Uddin Dibas, Christian Khairallah, Nizar Habash, Omar Fayez Sadi, Tariq Sairafy, Karmel Sarabta, and Abrar Ardah. 2022. [Maknuune: A large open palestinian Arabic lexicon](#). In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 131–141, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Rana Gaber, Yara Allam, Serag Amin, Ranwa Aly, and Bashar Alhafni. 2026. MBZUAI at AMIYA shared task 2026: Adapting open-source llms for dialectal arabic. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.
- 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.
- Khaleel Hamad and Ahmad Al-Najjar. 2026. A closed-track system for palestinian arabic in the AMIYA shared task. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Shereen Hussein, Mona Farouk, and ElSayed Hemayed. 2019. [Gender identification of egyptian dialect in twitter](#). *Egyptian Informatics Journal*, 20(2):109–116.
- 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.
- Shankar Kumar and Bill Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176.
- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In *Forty-second International Conference on Machine Learning*.
- AI Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog*. Retrieved December, 20:2024.
- Jonathan Mutal, Perla Al Almaoui, Simon Hengchen, and Pierrette Bouillon. 2026. AladdinFTI @ AMIYA - three wishes for arabic nlp: Fidelity,

- diglossia, and multidialectal generation. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, 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, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.
- Aissam Outchakoucht and Hamza Es-Samaali. 2024. [The evolution of darija open dataset: Introducing version 2](#). *Preprint*, arXiv:2405.13016.
- 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.
- Maja Popović. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. 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.
- Nathaniel Romney Robinson, Shahd Abdelmoneim, Kelly Marchisio, and Sebastian Ruder. 2025. [AL-QASIDA: Analyzing LLM quality and accuracy systematically in dialectal Arabic](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22048–22065, Vienna, Austria. Association for Computational Linguistics.
- Hashem Sellat, Shadi Saleh, Mateusz Krubiński, Adam Pospíšil, Petr Zemánek, and Pavel Pecina. 2023. [UFAL parallel corpus of north levantine 1.0](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).
- Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, and 1 others. 2023. [Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models](#). *arXiv preprint arXiv:2308.16149*.
- Shivalika Singh, Freddie Vargus, Daniel D’souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividias Mataciunas, Laura O’Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 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](#). pages 303–324.
- Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Mohamedou Cheikh Tourad, Rahaf Alhamouri, Rwa Assi, Aisha Alraeesi, Hour Mohamed, Fakhraddin Alwajih, Abdelrahman Mohamed, Abdellah El Mekki, El Moatez Billah Nagoudi, Benelhadj Djelloul Mama Saadia, Hamzah A. Alsayadi, Walid Al-Dhabyani, and 8 others. 2024. [Casablanca: Data and models for multidialectal Arabic speech recognition](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21745–21758, Miami, Florida, USA. Association for Computational Linguistics.
- Taghreed Tarmom, William Teahan, Eric Atwell, and Mohammad Ammar Alsalka. 2020. [Compression versus traditional machine learning classifiers to detect code-switching in varieties and dialects: Arabic as a case study](#). *Natural Language Engineering*, 26(6):663–676.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. [Aya model: An instruction finetuned open-access multilingual language model](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. [Qwen2 technical report](#). *arXiv preprint arXiv:2407.10671*.

A Annotator Observations

In this section we discuss some of the annotators’ qualitative observations as they conducted the human evaluation.

Prompt	[..] dataset which contains 2D images, [..] how to find the closest image? Explain in Egyptian Arabic.	Flu.	Adh.
NUS-IDS	موصلة إني، إني إمتلك بيانات بتحتوي على قاعدة صور في ابعاد ٢، بس، إني محتاج أقدر أسوي حاجة هتكون كده: إني عندني صورة جديدة، وكنت مش هتستني الدنيا أزورق منها، بس هتفكر إن أهم صورة بنقدر نحياها هي الي قدامنا، أو اللي هيوازيها مش هتفرقش بينا و بينها، هتكون كيف كدة؟ ...بالظبط هتستخدم الفنيقة اللي هيحسها	2	3
Aladdin	Step 1: Understand the problem statement: The problem is that I am trying to find the best match for an image from a dataset containing a list of images. The goal is to output the index of the best-matching image in the list. Step 2: Identify the input and output formats. Input format: A list of images. Output format: The index of the image in the input list that matches the query image...	1	3
MBZUAI	-	1	1

Table 10: Example prompt for a technical request in Egyptian Arabic, responses shortened for display

A.1 Egyptian

The LLMs evaluated for Egyptian Arabic demonstrate comparable *Adherence* levels, with most responses partially fulfilling the original prompt requests. However, notable differences emerged with respect to Fluency. The NUS-IDS model drew from a mix of dialectal terms spanning multiple varieties, while maintaining only partial sentence coherence. Aladdin’s model generated dialectal phrases characteristic of social media registers, with less alignment to the request and repetition of words within the prompt (i.e. parroting). The MBZUAI model demonstrated stronger comprehension relative to other submissions, though responses contained grammatical errors or fell short of native-like production.

Additionally, we observed instances where NUS-IDS and Aladdin models demonstrated understanding and high adherence of the prompt but failed to maintain the Egyptian dialect, instead responding in a mixed dialect or entirely in English. This is especially evident in prompts requiring code, or technical or highly structured outputs, where responses could be thorough, albeit in English or other dialects.

Table 10 presents a representative example where the prompt asks to explain a technical concept in Egyptian Arabic. NUS-IDS produced mixed dialectal lexical items with inconsistent structure, Aladdin generated dialectal phrasing but failed to address prompt content, and MBZUAI

defaulted entirely to English.

A.2 Moroccan

Across the 100 Moroccan prompts, the four systems can be divided into two distinct performance tiers. NUS-IDS consistently underperformed on both *Fluency* and *Adherence*, producing responses that were often difficult to comprehend and drifted from the target Moroccan variety.

Among the remaining three systems, the performance trade-offs demonstrate more nuance. MBZUAI achieves the highest average Adherence, suggesting it most reliably maintains alignment with Moroccan dialectal features. UM achieves the highest average Fluency, indicating more consistently readable and well-formed responses. Aladdin demonstrates competitive performance on Arabic prompts but exhibits high variance, performing well when prompts are already in Arabic, but declining sharply on English or technical prompts that still require Moroccan Arabic output.

This gap is illustrated in Table 11, where an English prompt explicitly requesting Moroccan Arabic asks how a corporate lawyer might pivot to launching an AI-enabled startup, UM produces a structured, actionable response in the Moroccan dialect demonstrating stability. MBZUAI similarly maintains dialectal output but with slightly more variability in phrasing. In contrast, Aladdin tends to drift on such cross-lingual instructions, often resorting to English or generating unrelated con-

tent, while NUS-IDS outputs incoherent or off-topic text.

A.3 Palestinian

The results indicate that none of the evaluated models consistently adhered well to the prompts, with average Adherence scores remaining low across all submissions. MBZUAI performed best overall, achieving the highest Adherence score and notably stronger Fluency (3.48), suggesting that it has better command of the Palestinian dialect. The remaining teams (NUS-IDS, Aladdin, and BYU) demonstrated similar levels of Adherence, while Fluency scores varied from limited to moderately understandable.

These findings indicate that while some progress has been made in Fluency, particularly by MBZUAI, accurate fulfillment of user requests in Palestinian Arabic remains a significant challenge across models.

Additionally, prompts requiring paraphrasing of specific statements while preserving meaning in Palestinian Arabic revealed notable performance variation across teams. Several models struggled to produce genuine paraphrases without using the original wording, implying limited lexical flexibility. See Table 12.

A.4 Saudi Arabian

Overall, the models demonstrate greater Fluency in the Saudi dialect compared to the other dialects, in addition to marginally higher Adherence scores. Although responses contained highly dialectal vocabulary, there are instances where sentence structure did not form fully coherent answers or failed to completely address the original user request.

It is interesting to note that where prompts incorporated culturally relevant content using local idioms, the models did not produce meaningful responses despite using terminology used frequently in Saudi conversational style.

NUS-IDS produced instances of mixed dialects within single responses alongside highly dialectal but incoherent outputs. Team Aladdin generated phrases in non-Saudi varieties and failed to provide responses with substantial meaning, largely parroting the original prompt without meaningful elaboration. UM produced dialectal phrases typically used in conversation but without coherent sentences.

MBZUAI demonstrated stronger comprehension and Fluency relative to other submissions,

though generated responses contained grammatical errors or fell short of native-like production. It is also notable that this model produced hallucinated responses with a confident tone, particularly relating to idioms and phrases. See Table 13.

A.5 Syrian

With regards to Syrian Arabic, Adherence to prompts remained limited across most submissions. MBZUAI clearly outperformed other systems and demonstrated stronger Syrian dialect generation, achieving the highest Adherence (1.62) and Fluency (3.43).

For the NUS-IDS, Aladdin, UM, and SDNLP teams, the results yielded comparably low Adherence scores, reflecting partial or irrelevant responses to prompts. With respect to Fluency, SDNLP and UM produced moderately more natural responses than NUS-IDS and Aladdin. However, responses did not sound completely native, reinforcing the idea of gaps in dialectal naturalness and accuracy.

Models exhibited unexpected difficulties with what should have been a relatively straightforward vocabulary word: *صفحة* (sfha), a well-known Syrian and Levantine dish. Only UM partially identified the term within a food context, while SDNLP largely reproduced the prompt without elaboration. NUS-IDS and Aladdin misinterpreted the term entirely, conflating *صفحة* with *صفحة تكتونية* (tectonic plate) and producing responses that were off-tangent. See Table 14.

B Genre-specific Automatic Scores

In this section we display the ADI2 and chrF++ scores across all different evaluation subsets mentioned in §2.4.

See tables 15, 16, 17, 18, 19, 20, 21, and 18, 22 for ADI2 scores across eval subsets; and see tables 23, 24, 25, 26, 27, 28, 29, and 30 for chrF++ scores across eval subsets.

Prompt	وشو شكل علم السعودية؟	Flu.	Adh.
NUS-IDS	علم المملكة العربية السعودية هو علم يتبنى بتركيبة من الصفراء والمينا. وشملت التباينات الفنية للوحة المينا زخارف هندسية متوسطة الدقة وعلى شكل حلقة. وتتوزع الزخارف في جميع الأقسام والزوايا في لوحة المينا، مع احتفاظ لوحة الصفراء بدواخلها السمكية والضخمة، وتباين الألوان في اللوحة بأشكال هندسية دقيقة.	4	2
Aladdin	هذا هو علم المملكة العربية السعودية	3	4
UM	ايش لونها علمكم ايش لونها علمكم الله يستر عليك من الاشكال هذي الي تبطح الناس تحت اقدامها ياليتي انا الي ابوه يوه ياليت	3	2
MBZUAI	علم المملكة العربية السعودية يتكون من اربعة ارباع متساويه، الجزء العلوي باللون الازرق والجزء السفلي باللون الابيض، وفي الوسط باللون الاخضر مع ثلاث نجوم بيضاء في اليسار.	4	1

Table 13: Example prompt asking to describe the Saudi flag in Saudi Arabic; models generate made up answers with moderate fluency. Note: responses have been shortened for formatting purposes.

Prompt	شو هي الصفيحة السورية؟	Flu.	Adh.
NUS-IDS	وكم توزع عليها الجبال والمرتفعات؟ المساحة المساحة الكلية للصفيحة السورية: ١٢ مليون كم. المساحة الكلية للصفيحة السورية تشكل حوالي ٣٥% من المساحة الكلية للنصف الكرة الغربي ويضم الكتلة الجبلية السورية الممتدة عبر الجزء الجنوبي الشرقي من أوراسيا... ويحدها من الشمال قارتا أوروبا وأقصى شمال إفريقيا،	3	1
Aladdin	الصفيحة السورية هي منطقة من الأرض تحت المحيطات.	3	1
UM	انا ما بعرف الا صرره ازمه قلبه من ريحه طبخك اديش بتاكل من هالصفيحه امجد انا ماني جوع	1	3
MBZUAI	؟	1	1
SDNLP	أنت شو عم تقولي صفيحة سوريا؟	3	1

Table 14: Example prompt asking about a well-known Syrian dish; models generate mixed responses with little adherence. Note: responses have been shortened for formatting purposes.

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	35.4%	13.3%	8.9%	11.8%	3.2%
Aladdin-FTI	contrastive1	33.8%	12.8%	9.8%	14.1%	4.5%
BYU	primary	–	–	7.0%	–	–
MBZUAI	primary	53.3%	42.0%	13.7%	9.8%	10.8%
MBZUAI	contrastive1	54.8%	33.2%	16.2%	12.2%	8.3%
MBZUAI	contrastive2	56.8%	51.5%	21.5%	27.5%	7.2%
NUS-IDS	primary	73.9%	66.6%	16.3%	8.0%	3.9%
NUS-IDS	contrastive1	43.4%	19.3%	12.8%	26.8%	2.9%
NUS-IDS	contrastive2	25.4%	0.6%	3.6%	0.7%	0.9%
SDNLP	primary	–	–	–	30.4%	–
UM	primary	87.8%	–	–	53.7%	53.9%
Baseline		5.7%	5.4%	4.3%	1.4%	0.4%

Table 15: ADI2 on Palm prompts

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	39.2%	8.2%	4.4%	8.4%	0.3%
Aladdin-FTI	contrastive1	39.3%	7.7%	4.8%	11.8%	0.6%
BYU	primary	–	–	6.8%	–	–
MBZUAI	primary	69.6%	59.2%	12.6%	13.0%	2.5%
MBZUAI	contrastive1	56.8%	47.8%	9.2%	13.5%	3.8%
MBZUAI	contrastive2	71.6%	58.6%	11.6%	11.4%	9.2%
NUS-IDS	primary	87.1%	66.8%	6.9%	2.1%	1.7%
NUS-IDS	contrastive1	73.3%	17.1%	6.4%	32.2%	0.3%
NUS-IDS	contrastive2	39.7%	0.2%	0.9%	0.0%	0.1%
SDNLP	primary	–	–	–	29.7%	–
UM	primary	77.5%	–	–	45.5%	42.8%
Baseline		13.1%	12.3%	0.7%	1.2%	0.0%

Table 16: ADI2 on FLORES monolingual prompts

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	44.1%	17.0%	11.3%	15.3%	4.9%
Aladdin-FTI	contrastive1	49.7%	12.4%	11.9%	24.1%	2.2%
BYU	primary	–	–	11.1%	–	–
MBZUAI	primary	72.8%	68.2%	19.3%	24.7%	16.0%
MBZUAI	contrastive1	74.7%	61.2%	16.6%	22.2%	14.3%
MBZUAI	contrastive2	77.3%	66.6%	20.1%	20.1%	21.1%
NUS-IDS	primary	85.2%	75.2%	9.3%	4.7%	4.8%
NUS-IDS	contrastive1	58.6%	25.3%	8.0%	38.6%	0.5%
NUS-IDS	contrastive2	52.6%	1.4%	2.8%	0.6%	2.7%
SDNLP	primary	–	–	–	36.3%	–
UM	primary	82.3%	–	–	57.5%	47.9%
Baseline		20.3%	27.1%	3.1%	4.8%	1.3%

Table 17: ADI2 on MADAR monolingual prompts

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	38.8%	10.3%	4.6%	29.2%	1.9%
Aladdin-FTI	contrastive1	36.9%	11.3%	3.7%	32.0%	1.7%
BYU	primary	–	–	1.4%	–	–
MBZUAI	primary	64.2%	33.9%	4.4%	14.3%	4.9%
MBZUAI	contrastive1	52.3%	25.0%	0.9%	16.5%	2.8%
MBZUAI	contrastive2	53.6%	30.1%	3.1%	20.3%	8.0%
NUS-IDS	primary	71.8%	59.4%	6.6%	6.9%	3.9%
NUS-IDS	contrastive1	78.2%	83.1%	15.0%	36.9%	15.0%
NUS-IDS	contrastive2	24.5%	0.4%	0.1%	0.2%	0.2%
SDNLP	primary	–	–	–	25.5%	–
UM	primary	47.5%	–	–	31.1%	37.7%
Baseline		7.9%	1.6%	0.1%	0.1%	0.1%

Table 18: ADI2 on FLORES ENG→DA MT outputs

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	21.1%	8.0%	3.3%	20.1%	1.2%
Aladdin-FTI	contrastive1	17.9%	10.2%	3.1%	22.8%	0.6%
BYU	primary	–	–	1.0%	–	–
MBZUAI	primary	38.2%	25.2%	0.3%	14.7%	4.3%
MBZUAI	contrastive1	34.2%	21.6%	0.6%	14.5%	3.0%
MBZUAI	contrastive2	38.5%	21.1%	0.1%	15.4%	6.2%
NUS-IDS	primary	86.1%	75.0%	10.6%	12.1%	8.2%
NUS-IDS	contrastive1	83.0%	79.3%	13.3%	34.9%	20.5%
NUS-IDS	contrastive2	20.4%	0.2%	0.1%	0.1%	0.2%
SDNLP	primary	–	–	–	10.5%	–
UM	primary	31.5%	–	–	29.1%	37.6%
Baseline		18.3%	16.6%	1.2%	0.5%	1.0%

Table 19: ADI2 on FLORES MSA→DA MT outputs

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	66.7%	33.3%	12.0%	41.9%	16.7%
Aladdin-FTI	contrastive1	65.7%	30.6%	9.5%	46.3%	12.2%
BYU	primary	–	–	5.6%	–	–
MBZUAI	primary	63.6%	48.0%	10.6%	22.9%	17.7%
MBZUAI	contrastive1	60.0%	47.2%	6.8%	21.4%	15.6%
MBZUAI	contrastive2	69.3%	49.2%	6.7%	25.1%	24.6%
NUS-IDS	primary	58.9%	51.9%	3.5%	9.3%	7.5%
NUS-IDS	contrastive1	89.8%	77.4%	17.3%	38.8%	29.1%
NUS-IDS	contrastive2	46.8%	1.3%	0.2%	1.2%	0.8%
SDNLP	primary	–	–	–	37.8%	–
UM	primary	66.0%	–	–	38.0%	55.7%
Baseline		10.2%	7.9%	1.0%	1.1%	0.6%

Table 20: ADI2 on MADAR ENG→DA MT outputs

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	61.0%	33.0%	12.2%	39.0%	10.5%
Aladdin-FTI	contrastive1	65.4%	37.8%	10.5%	39.9%	6.6%
BYU	primary	–	–	3.1%	–	–
MBZUAI	primary	57.2%	45.1%	4.2%	20.8%	16.9%
MBZUAI	contrastive1	51.2%	46.3%	5.6%	19.5%	16.2%
MBZUAI	contrastive2	50.6%	48.0%	1.2%	22.2%	21.6%
NUS-IDS	primary	70.0%	51.1%	32.0%	29.2%	15.3%
NUS-IDS	contrastive1	88.6%	82.7%	17.2%	27.5%	29.5%
NUS-IDS	contrastive2	38.8%	0.6%	0.3%	0.8%	0.2%
SDNLP	primary	–	–	–	22.3%	–
UM	primary	63.3%	–	–	36.0%	53.4%
Baseline		13.7%	19.6%	2.2%	1.0%	2.6%

Table 21: ADI2 on MADAR MSA→DA MT outputs

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	4.4%	0.5%	1.3%	6.7%	0.7%
Aladdin-FTI	contrastive1	3.0%	0.5%	1.1%	5.6%	0.6%
BYU	primary	–	–	5.0%	–	–
MBZUAI	primary	45.2%	35.8%	6.8%	21.3%	8.3%
MBZUAI	contrastive1	46.7%	33.1%	4.5%	18.2%	4.7%
MBZUAI	contrastive2	44.9%	34.7%	5.1%	28.4%	13.4%
NUS-IDS	primary	23.8%	56.3%	8.5%	9.2%	11.3%
NUS-IDS	contrastive1	87.6%	77.0%	16.5%	51.4%	15.6%
NUS-IDS	contrastive2	23.5%	0.4%	0.3%	0.8%	0.5%
SDNLP	primary	–	–	–	31.0%	–
UM	primary	44.8%	–	–	23.2%	44.5%
Baseline		2.9%	1.4%	0.5%	0.8%	0.8%

Table 22: ADI2 on Arena-Hard prompts

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	31.7	35.1	37.3	24.0	35.7
Aladdin-FTI	contrastive1	31.4	34.4	36.3	21.4	35.6
BYU	primary	-		17.7	-	-
MBZUAI	primary	20.7	31.9	36.4	30.3	31.8
MBZUAI	contrastive1	25.7	33.5	33.7	30.2	33.8
MBZUAI	contrastive2	25.0	29.1	29.0	24.1	29.4
NUS-IDS	primary	11.5	15.7	16.1	15.4	16.3
NUS-IDS	contrastive1	23.7	18.1	19.2	17.7	19.2
NUS-IDS	contrastive2	28.3	25.9	31.5	21.8	35.8
SDNLP	primary	-		-	13.7	-
UM	primary	29.1	-		36.4	21.1
Baseline		25.8	32.9	35.1	34.4	37.1

Table 23: chrF++ on FLORES ENG→DA

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	25.1	29.2	30.6	25.1	46.0
Aladdin-FTI	contrastive1	23.9	28.3	30.0	21.1	32.4
BYU	primary	-		22.7	-	-
MBZUAI	primary	35.3	39.9	43.0	38.6	58.5
MBZUAI	contrastive1	35.5	45.3	42.0	37.6	60.3
MBZUAI	contrastive2	32.9	36.5	46.0	37.8	43.1
NUS-IDS	primary	14.4	23.2	22.2	22.1	25.9
NUS-IDS	contrastive1	22.5	19.3	20.0	21.3	22.9
NUS-IDS	contrastive2	33.9	14.4	35.4	16.3	67.3
SDNLP	primary	-		-	16.6	-
UM	primary	35.0	-		43.6	27.1
Baseline		33.4	46.9	44.0	43.7	69.9

Table 24: chrF++ on FLORES MSA→DA

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	43.1	49.4	52.2	40.6	52.4
Aladdin-FTI	contrastive1	42.5	49.2	53.0	40.8	53.0
BYU	primary	-		15.0	-	-
MBZUAI	primary	40.7	45.1	48.4	53.6	56.9
MBZUAI	contrastive1	48.3	51.6	57.3	52.1	55.7
MBZUAI	contrastive2	41.0	41.0	47.2	46.8	49.7
NUS-IDS	primary	25.1	26.1	27.9	25.9	26.5
NUS-IDS	contrastive1	45.8	26.4	28.5	28.6	27.9
NUS-IDS	contrastive2	34.8	6.3	48.8	9.8	55.5
SDNLP	primary	-		-	34.7	-
UM	primary	46.7	-		52.4	0.0
Baseline		47.0	52.7	57.2	53.3	58.8

Table 25: chrF++ on FLORES DA→ENG

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	25.5	28.9	30.9	25.8	44.1
Aladdin-FTI	contrastive1	25.2	27.4	30.7	22.0	33.9
BYU	primary	-		22.6	-	-
MBZUAI	primary	39.1	45.0	44.1	41.8	72.2
MBZUAI	contrastive1	44.1	53.1	45.0	45.4	73.0
MBZUAI	contrastive2	36.9	39.3	43.7	37.4	55.9
NUS-IDS	primary	18.7	26.4	23.8	22.5	32.7
NUS-IDS	contrastive1	35.1	25.4	23.9	26.8	56.7
NUS-IDS	contrastive2	38.6	16.2	41.0	17.9	67.2
SDNLP	primary	-		-	7.5	-
UM	primary	42.6	-		49.7	43.8
Baseline		42.2	52.6	49.0	46.1	79.1

Table 26: chrF++ on FLORES DA→MSA

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	34.5	31.0	38.3	17.7	38.2
Aladdin-FTI	contrastive1	34.0	28.7	35.9	15.7	38.4
BYU	primary	-		6.6	-	-
MBZUAI	primary	29.2	39.5	38.4	33.2	42.9
MBZUAI	contrastive1	30.7	37.0	35.0	33.3	45.4
MBZUAI	contrastive2	28.5	38.5	36.0	31.3	44.5
NUS-IDS	primary	4.6	5.5	5.8	6.0	7.4
NUS-IDS	contrastive1	5.6	11.5	9.6	8.3	10.4
NUS-IDS	contrastive2	15.2	9.7	8.2	15.4	23.3
SDNLP	primary	-		-	18.2	-
UM	primary	32.6	-		28.9	16.7
Baseline		17.6	20.9	22.5	21.5	29.7

Table 27: chrF++ on MADAR ENG→DA

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	31.4	26.9	33.3	16.1	32.2
Aladdin-FTI	contrastive1	28.9	24.5	32.9	13.5	30.7
BYU	primary	-		7.6	-	-
MBZUAI	primary	33.6	38.1	31.0	33.5	42.8
MBZUAI	contrastive1	34.6	37.9	32.9	33.5	43.7
MBZUAI	contrastive2	33.9	35.8	29.0	35.5	41.6
NUS-IDS	primary	7.9	9.1	7.8	8.8	10.3
NUS-IDS	contrastive1	5.3	11.2	8.8	8.0	9.1
NUS-IDS	contrastive2	26.7	7.6	16.7	4.4	29.8
SDNLP	primary	-		-	13.8	-
UM	primary	30.0	-		30.8	17.2
Baseline		22.1	25.4	27.2	25.6	29.6

Table 28: chrF++ on MADAR MSA→DA

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	56.3	51.4	55.5	36.9	58.9
Aladdin-FTI	contrastive1	55.3	50.7	55.1	36.4	59.0
BYU	primary	-		4.3	-	-
MBZUAI	primary	58.1	55.9	56.6	55.7	63.0
MBZUAI	contrastive1	59.6	59.1	60.0	54.4	64.8
MBZUAI	contrastive2	56.2	50.8	59.1	52.8	60.1
NUS-IDS	primary	11.1	9.7	10.2	9.6	11.1
NUS-IDS	contrastive1	7.8	14.0	13.9	12.1	15.9
NUS-IDS	contrastive2	31.1	6.0	41.6	8.4	49.6
SDNLP	primary	-		-	40.6	-
UM	primary	59.4	-		50.2	0.0
Baseline		42.0	39.8	37.2	37.1	48.5

Table 29: chrF++ on MADAR DA→ENG

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	34.0	30.9	31.9	20.4	34.7
Aladdin-FTI	contrastive1	33.6	30.3	30.7	18.7	35.0
BYU	primary	-		7.9	-	-
MBZUAI	primary	39.7	40.1	33.3	35.1	42.7
MBZUAI	contrastive1	43.7	42.5	35.8	40.8	44.2
MBZUAI	contrastive2	40.8	38.1	35.3	25.5	44.1
NUS-IDS	primary	5.8	6.3	5.7	5.4	6.7
NUS-IDS	contrastive1	6.4	11.1	10.5	11.4	11.4
NUS-IDS	contrastive2	23.9	4.8	14.3	5.2	29.5
SDNLP	primary	-		-	4.4	-
UM	primary	31.7	-		26.2	22.6
Baseline		30.2	30.6	29.3	29.2	34.0

Table 30: chrF++ on MADAR DA→MSA