MBZUAI at AMIYA Shared Task 2026: Adapting Open-Source LLMs for Dialectal Arabic

Rana Gaber[1*], **Yara Allam**[1*], **Serag Amin**[1*], **Ranwa Aly**[1*], **Bashar Alhafni**[2]
[1]Alexandria University
[2]Mohamed bin Zayed University of Artificial Intelligence
cds.{ranaahmed30309,yaraibrahim23394}@alexu.edu.eg
cds.{seragamin23144,ranwakhaled30408}@alexu.edu.eg
bashar.alhafni@mbzuai.ac.ae

Abstract

This paper presents our contribution to the closed data track of the AMIYA Shared Task on Dialectal Arabic text generation. In this track, we train fully open-source Large Language Models (LLMs) on five Arabic dialects: Egyptian, Moroccan, Palestinian, Saudi, and Syrian, using the provided training datasets. We experiment with different base and instruct models using several pretraining and instruction tuning approaches. In total, five models were submitted, with three variants per dialect. Our best-performing models for the five dialects are ALLaM for Egyptian, LLaMa for Moroccan, and Palestinian, and Aya for Saudi and Syrian.

1 Introduction

Arabic is a morphologically rich language characterized by diglossia (Ferguson, 1959), a linguistic phenomenon in which Modern Standard Arabic (MSA), the standard and formal form of the language, co-exists with a non-standard variety, Dialectal Arabic (DA). Complicating matters, there are multiple DA varieties, each differing from one another and from MSA in phonology, morphology, syntax, and lexicon (Keleg et al., 2023), and are commonly classified regionally (e.g., Egyptian, Levantine, Gulf).

While MSA is used in formal settings such as news and education, it is not the native language of any Arabic speaker. In contrast, dialects constitute the true native varieties of Arabic, historically connected to Classical Arabic and shaped by regional linguistic contact. Although DA is primarily spoken, it is widely used in informal written communication, particularly on social media and online platforms.

Despite their widespread use, Arabic dialects are severely underrepresented in large-scale textual resources. Compared to abundant and standardized MSA corpora, dialectal data is scarce, noisy, and highly heterogeneous, with inconsistent orthography and frequent code-switching. As a result, LLMs are predominantly pretrained on massive quantities of MSA text, with limited exposure to dialectal Arabic. This imbalance leads to substantial performance gaps, causing LLMs to struggle with understanding, generating, and reasoning over dialectal input (Bergman and Diab, 2022).

In this paper, we aim to bridge this gap by describing our system submission to the AMIYA Shared Task (Robinson et al., 2026). We explore adapting 12 open-source LLMs to five Arabic dialects: Egyptian (EGY), Moroccan (MAR), Palestinian (PSE), Saudi (SAU), and Syrian (SYR). Our models are evaluated using AL-QASIDA benchmark (Robinson et al., 2025), an evaluation suite that measures an LLM's dialectal fidelity, understanding, generation quality, and sensitivity to MSA-DA diglossia. Our adaptation framework consists of two primary phases: Continual Pretraining (CPT), applied to base models only, and Instruction Tuning, applied to both base and instruction-tuned models. For CPT, we explore two strategies: 1) Curriculum CPT, where we train first on MSA and English before transitioning to DA-only data; 2) Mixed CPT, where we train on a mix of MSA, English, and DA data. Our results indicate that the superiority of either CPT setup relies on the model's architecture, while fine-tuned instruct models achieve superior performance across all dialects compared to base models adapted using CPT.

2 Background

2.1 Dialectal Arabic NLP

DA NLP research has received growing attention in the Arabic NLP community, driven largely by the development of monolingual and multilingual dialectal resources and benchmarks (McNeil and

[*] Equal contribution.

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 373–384
March 29, 2026. ©2026 Association for Computational Linguistics

Faiza, 2011; Zaidan and Callison-Burch, 2011; Zbib et al., 2012; Cotterell and Callison-Burch, 2014; Salama et al., 2014; Jeblee et al., 2014; Al-Badrashiny and Diab, 2016; Zaghouani and Charfi, 2018; Abdul-Mageed et al., 2018; Bouamor et al., 2019; Sajjad et al., 2020; Abdul-Mageed et al., 2020, 2021, 2022; Nagoudi et al., 2022; Abdul-Mageed et al., 2023, 2024). These efforts have enabled research across a wide range of dialectal NLP tasks, including machine translation (MT), dialect identification, and cross-dialect modeling.

Despite this progress, DA resources remain substantially more limited compared to MSA. Dialectal data is unevenly distributed across varieties, with certain dialects receiving considerably more coverage than others. This imbalance stems in part from the fact that DA is primarily a spoken language; when written, it is most commonly produced in informal contexts such as social media and online communication. Consequently, dialectal corpora are often noisy and heterogeneous, exhibiting a lot spelling variation. These challenges are further exacerbated by the absence of standardized orthographies for Arabic dialects. Although prior work has proposed conventional orthography for DA (Habash et al., 2012; Jarrar et al., 2016; Khalifa et al., 2018; Habash et al., 2018; Eryani et al., 2020; Alhafni et al., 2024), adoption remains limited, and most real-world data continues to reflect unconstrained user-generated text.

With the emergence of LLMs, recent research has increasingly focused on evaluating and adapting Arabic-centric models such as AceGPT (Liang et al., 2024), ALLaM (Bari et al., 2024), Fanar (Fanar Team et al., 2025), Jais (Sengupta et al., 2023; Anwar et al., 2025), NileChat (Shang et al., 2025a), and Atlas-Chat (Shang et al., 2025b) alongside the development of large-scale benchmarks targeting Arabic and its dialects (Koto et al., 2024; Faisal et al., 2024; Hijazi et al., 2024; Ashraf et al., 2025; Alghamdi et al., 2025; Mousi et al., 2025; Sadallah et al., 2025; Alwajih et al., 2025; Almatham et al., 2025; Robinson et al., 2025). Our work builds on this line of research by systematically studying the adaptation of open-source LLMs to multiple Arabic dialects under controlled training conditions, and by evaluating their dialectal fidelity, generation quality, and sensitivity to MSA–DA diglossia.

2.2 AMIYA Shared Task

The **A**rabic **M**odeling **I**n **Y**our **A**ccent (AMIYA) Shared Task focuses on building and adapting LLMs for DA. All models are evaluated using the AL-QASIDA benchmark (Robinson et al., 2025). AL-QASIDA consists of multiple evaluation tasks, including monolingual generation (DA→DA) and cross-lingual generation (EN→DA) for measuring DA *fidelity*; DA→EN MT for assessing *understanding*; EN→DA MT for evaluating generation *quality*; and MSA↔DA MT for probing *diglossia*.

The shared task has three tracks: 1) Closed Data Track; 2) Closed Models Track; and 3) Open Track. For all tracks, the shared task accepts submissions for five Arabic dialects: Egyptian (EGY), Moroccan (MAR), Palestinian (PSE), Saudi (SAU), and Syrian (SYR) We participate in the **Closed Data Track** across **all five** dialects.

3 System Overview

We investigate the adaptation of 12 open-source LLMs for DA, spanning two categories: base models and instruction-tuned models (Table 3). Our adaptation framework consists of two phases: **Continual Pretraining (CPT)** and **Instruction Tuning (IT)**. CPT is applied exclusively to base models, whereas IT is applied to both CPT-adapted base models and instruction-tuned models. For CPT, we examine two training strategies: 1) **Curriculum CPT**, in which models are first trained on MSA and English data before training on DA-only data; and 2) **Mixed CPT**, which trains models on a shuffled mixture of MSA, English, and DA data.

4 Data

We use the training datasets provided the closed data track of the shared task, namely: SauDial (Alanazi et al., 2025), ASR-EGARBCSC, MASC, DoDa (Outchakoucht and Es-Samaali, 2021), Shami Corpus (Abu Kwaik et al., 2018), Atlaset (Bounhar and Majjodi, 2025), SDC and EDC (Tarmom et al., 2019), Saudi Tweets (Alruily, 2020), SADSSlyC (Alahmari, 2025), EDGAD (ElSayed and Farouk, 2020), Casablanca (Talafha et al., 2024), JODA (Abandah et al., 2025) and UFAL (Sellat et al., 2023), with MADAR (Bouamor et al., 2018) (excluding the off-limits portion) being used for IT only. These datasets include one or multiple Arabic dialects with a subset containing bitexts in English and/or MSA.

The distribution of data used for CPT across both Curriculum CPT and Mixed CPT is in Table 1.

Dialect	DA	DA–MSA	DA–EN
EGY	223.9K	–	–
MAR	1.3M	–	233.2K
PSE	84.2K	58.5K	–
SAU	235.1K	2170	404
SYR	124.1K	127.8K	24.6K
Total	**2M**	**188.4K**	**258.25K**

Table 1: CPT data statistics in terms of sentences for DA monotexts and DA↔MSA/EN bitexts.

Dialect	MonoGen	XGen	MT
EGY	695	613	35.3K
MAR	633	1857	65.4k
PSE	570	525	60.4k
SAU	644	602	11.9k
SYR	752	534	248.2k
Total	**3.3K**	**4.1K**	**421.2K**

Table 2: IT Dataset distribution by dialect and tasks. Counts represent sentences.

4.1 Data Preprocessing

To standardize the input across training data, we apply a unified preprocessing pipeline using CAMeL tools (Obeid et al., 2020). The pipeline includes diacritic removal; normalization of Alif, Ya, and Ta Marbuta; character de-duplication; and the removal of special characters, emojis, and URLs found in social media text.

Moreover, the training data we use exhibits substantial imbalance across dialects, as shown in Tables 1 and 2 for CPT and IT data, respectively. Moroccan dominates the corpus, while Egyptian, Syrian, and Saudi data are smaller, and Palestinian is particularly limited with only 12K samples after preprocessing. To address this, we augment Palestinian data with linguistically similar Jordanian samples and downsample Moroccan data, retaining approximately 40% of the available Moroccan samples to reduce majority-dialect bias.

4.2 Instruction Tuning Data

For IT, we train the model on three tasks: MT, Monolingual Text generation (MonoGen), and cross-lingual Text generation (XGen).

For MT, we utilize all available bitexts in our dataset, excluding MADAR's off-limits portion, covering bidirectional translation for EN↔DA and MSA↔DA. We use the following instruction prompt:

> *Translate from {source_language} into {target_language}. Output only the translation. Do NOT output anything else before or after it.*

To address the scarcity of monolingual and cross-lingual text generation data, we generate synthetic instruction-response pairs using fully open-source models: Command-A-111B (Cohere et al., 2025) for Moroccan, Palestinian, and Saudi, and Command-R-35B for Egyptian and Syrian. The selection of these models was based on a preliminary inspection to observe which model is better for each dialect. For the MonoGen task (DA→DA), we randomly sample a subset of DA training data and use the DA instruction prompts defined by the AL-QASIDA benchmark to produce synthetic DA outputs. For the XGen task (EN→DA), we sample a different subset of DA data and prompt LLMs to generate corresponding English instruction prompts paired with the original DA text. Prompts used to generate the synthetic data are in Appendix B.

Table 2 reports the statistics of the IT datasets. While the synthetically generated monolingual and cross-lingual data is relatively small compared to the MT data, this design choice avoids introducing large volumes of synthetic content that may negatively impact data quality.

5 Experimental Setup

All models used in our experiments are listed in Table 3. For evaluation, we adopt the metrics defined by AL-QASIDA benchmark. Specficially, we use the Arabic Dialect Identification and Dialectness (ADI2) score for both monolingual and cross-lingual generation tasks, and chrF++ (Popović, 2015, 2017) for MT. The hyper-parameters we used for our CPT and IT setups are detailed in Appendix A. To streamline the evaluation process, we set the generation limit for new tokens to 64. In addition, to ensure comparability between models, we directly provide inputs to the model without any prompts or instruction templates.

6 Results

Baselines We evaluate all models in a zero-shot setting using the instruction prompts described in Section §4.2. Baseline results for all five dialects across three tasks are reported in Table 8 in the

Model	Type
Qwen3-8B-Base (Team, 2025)	Base
Qwen3-8B (Team, 2025)	Instruct
Llama-3.1-8B (Grattafiori et al., 2024)	Base
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	Instruct
Aya-Expanse-8B (Dang et al., 2024)	Instruct
ALLaM-7B-Instruct (Bari et al., 2025)	Instruct
Gemma-2-9B (Team et al., 2024)	Base
Bloom-7b1 (Workshop, 2024)	Base
Command-R7B-Arabic (Alnumay et al., 2025)	Instruct
Fanar-1-9B (Fanar Team et al., 2025)	Base
Fanar-1-9B-Instruct (Fanar Team et al., 2025)	Instruct
Jais-2-8B-Chat (Anwar et al., 2025)	Instruct

Table 3: Models used in our experiments and their types, Base or Instruct.

Dialect	Model	MonoGen	XGen	MT
EGY	Fanar-B	**0.72**	0.03	29
MAR	Fanar-B	0.48	0.02	27
PSE	Jais-2	0.02	**0.05**	49
SAU	Jais-2	0.04	0.03	**61**
SYR	Fanar-B	0.32	0.02	28

Table 4: Top performing baselines for each dialect in terms of ADI2 for MonoGen and XGen, and chrF++ for MT. Best results are in **bold**.

Appendix. Table 4 summarizes the top-performing baseline model for each dialect, averaged across tasks. For Egyptian, Moroccan, and Syrian, the base version of Fanar (Fanar-B) performs best, achieving very high monolingual scores despite lower cross-lingual and MT performance. For Palestinian and Saudi, Jais-2 ranks highest, showing strong MT results.

Candidate Model Selection We conduct lightweight adaptation runs to identify promising candidate models by evaluating our three training strategies on reduced subsets of the data. For Curriculum CPT, models are continually pretrained on all available English and MSA data, followed by 30% of the DA data. For Mixed CPT, models are pretrained on a 30% mixture of English, MSA, and DA data. For IT, models are trained on 30% of the IT dataset. All results for these settings are reported in Table 9 in the Appendix. Based on these experiments, we select a subset of models for full-scale pretraining and/or continued instruction tuning, which we detail here.

We select six models for full training: three base models (**LLaMA 3.1**, **Gemma**, and **Fanar Base**) and three instruction-tuned models (**Aya**, **ALLaM**, and **Command R**). During training, we save checkpoints at multiple stages (30%, 50%, 70%, and 100% of total steps), enabling performance analysis throughout training and selection of the most effective checkpoint for submission. Table 10 in the Appendix reports the evaluation results of all shortlisted models across checkpoints.

Dialect	Model	MonoGen	XGen	MT
EGY	ALLaM	0.67	0.27	46
MAR	LLaMa	**0.72**	**0.51**	38
PSE	ALLaM	0.21	0.05	44
SAU	Aya	0.2	0.03	**61**
SYR	Aya	0.25	0.1	46

Table 5: Top performing instruction tuned models for each dialect in terms of ADI2 for MonoGen and XGen, and chrF++ for MT. Best results are in **bold**.

6.1 Final Models

Following the shared task regulations, we submitted three runs per dialect, totaling 15 submissions. Runs were selected from checkpoints obtained at different training stages based on validation performance. Our submissions include early IT checkpoints of **ALLaM**, **Aya**, and **LLaMA**, a fully trained instruction-tuned **Aya** model, and a **Fanar Base** model trained using Mixed CPT followed by full IT. Table 11 in the Appendix details the submitted models per dialect.

Table 5 reports the best-performing model per dialect used for our submissions. These correspond to IT checkpoints selected after 30% of the total training steps, as they consistently achieved the strongest validation performance. Model selection was guided by a heuristic that maximizes the average performance across the three tasks, ADI2 for MonoGen and XGen, and chrF++ for MT, which provides a simple summary of cross-task behavior but may mask trade-offs between generation quality and translation performance.

We observe substantial performance variation across dialects. Moroccan and Egyptian outperform Syrian, Saudi, and Palestinian in MonoGen and XGen, while MT performance remains consistent. The highest MT score is achieved for Saudi using the Aya IT model selected at the 30% checkpoint. Notably, despite having more IT data, Syrian underperforms Egyptian, suggesting that continual pretraining data volume plays a more critical role than IT data in dialectal differentiation.

Dialect	Sys.	Gen	DA→EN	EN→DA	DA→MSA	MSA→DA
EGY	C1	0.39	53.4	**34.3**	**50.6**	**43.7**
	C2	**0.45**	43.3	31.2	39.1	36.4
	P	**0.45**	**47.7**	33.6	43.9	39.5
MAR	C1	0.54	51.0	**26.8**	44.1	**35.3**
	C2	**0.58**	44.6	25.8	37.9	33.1
	P	0.57	**44.9**	22.7	**39.3**	34.9
PSE	C1	0.09	**58.0**	34.0	**42.9**	40.1
	C2	**0.10**	50.0	30.5	41.8	**42.4**
	P	**0.10**	50.4	**36.9**	41.6	40.5
SAU	C1	0.09	**57.9**	**36.2**	66.3	**56.9**
	C2	**0.14**	52.2	32.6	53.2	42.8
	P	0.10	**58.4**	34.1	**65.3**	55.3
SYR	C1	0.17	52.7	**31.0**	**44.4**	36.8
	C2	**0.21**	48.2	25.7	34.6	37.4
	P	0.18	**54.1**	**31.0**	40.3	**37.6**

Table 6: Official shared task results of our Contrastive-1 (C1), Contrastive-2 (C2), and Primary (P) systems across dialects. Gen denotes overall generation performance (monolingual and cross-lingual) evaluated using ADI2. MT results are reported using chrF++. Best results per dialect and task are in **bold**.

Dialect	Adequacy	Fluency
EGY	2.1	3.6
MAR	2.0	3.2
PSE	1.6	3.5
SAU	2.1	3.4
SYR	1.6	3.4

Table 7: Human evaluation in terms of adequacy and fluency of our Primary system across dialects.

6.2 Official Results

Table 6 presents the official shared task results using automatic metrics. No single system dominates all tasks across dialects: our Contrastive-1 (C1) and Primary (P) systems are the best on several MT directions, whereas Contrastive-2 (2) consistently yields the highest generation scores. Compared to other participants, our systems perform particularly strongly on MT across dialects, whereas generation remains more competitive across teams.

Table 7 presents human evaluation results in terms of adequacy and fluency of our primary systems across dialects. Compared to other participants, our system achieves the best scores across all dialects on both dimensions, with the exception of fluency on Moroccan.

7 Error Analysis

Figure 1 in Appendix C shows t-SNE projections of average sentence-level hidden representations for the five target dialects and MSA. Each plot is constructed from 600 parallel samples per dialect from the MADAR Dev set, using representations produced by the corresponding model. The visualizations align with the findings in Section 6.1: Egyptian and Moroccan dialects form more compact and well-separated clusters, consistent with their stronger MonoGen and XGen performance, while Syrian, Saudi, and Palestinian dialects exhibit more diffuse and overlapping representations, indicating greater representational ambiguity.

8 Conclusion

In this paper, we evaluated multiple strategies for improving DA performance in LLMs as part of our participation in the AMIYA Shared Task (closed data track). Across training setups, instruction-tuned models consistently outperformed base models on the AL-QASIDA benchmark, often achieving peak performance using only 30% of the task data. These results suggest that the volume of data during continual pretraining may have a stronger effect on the performance than IT data on DA tasks.

Limitations

In this work, we explored several model training approaches on a large list of models, but we acknowledge the limitations that should be considered when interpreting the results. Although we rely exclusively on the datasets provided by the organizers, the amount of data varies across dialects, with some dialects (e.g., Moroccan) dominating the corpus, while others are underrepresented. We try to mitigate this imbalance through down-sampling and data synthesis, but these methods do not fully resolve the issue, and models continue to have uneven performance across dialects.

Additionally, the use of synthetically generated data for instruction tuning introduces the risk of noise and hallucinations, and may not accurately reflect real-world usage of these dialects. We also note that our findings may not generalize to dialects not represented in the training data. Moreover, in some analyses, we rely on average performance across tasks to summarize cross-task behavior; while this provides a simple aggregate view, it may obscure trade-offs between generation quality and translation performance.

Finally, we limit our experiments to mid-sized models (7–9B parameters) and, due to time constraints, do not conduct a more in-depth analysis of the differences between the curriculum-based and mixed pretraining strategies explored in this work.

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 and Information Technology*, 11(3):319–335.

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of Arabic dialects. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

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, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. NADI 2022: The third nuanced Arabic dialect identification shared task. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 85–97, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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).

Mohamed Al-Badrashiny and Mona Diab. 2016. LILI: A simple language independent approach for language identification. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Osaka, Japan.

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.

Naif Alanazi, Mohammed Al-Batineh, and Hussein Abu-Rayyash. 2025. Saudial: The saudi arabic dialects game localization dataset. *Data in Brief*, 62:111906.

Emad A. Alghamdi, Reem Masoud, Deema Alnuhait, Afnan Y. Alomairi, Ahmed Ashraf, and Mohamed Zaytoon. 2025. AraTrust: An evaluation of trustworthiness for LLMs in Arabic. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8664–8679, Abu Dhabi, UAE. Association for Computational Linguistics.

Bashar Alhafni, Sarah Al-Towaity, Ziyad Fawzy, Fatema Nassar, Fadhl Eryani, Houda Bouamor, and Nizar Habash. 2024. Exploiting dialect identification in automatic dialectal text normalization. In

Proceedings of the Second Arabic Natural Language Processing Conference, pages 42–54, Bangkok, Thailand. Association for Computational Linguistics.

Rawan Nasser Almatham, Kareem Mohamed Darwish, Raghad Al-Rasheed, Waad Thuwaini Alshammari, Muneera Alhoshan, Amal Almazrua, Asma Al Wazrah, Mais Alheraki, Firoj Alam, Preslav Nakov, Norah A. Alzahrani, Eman Albilali, Nizar Habash, Abdelrahman Mustafa El-Sheikh, Muhammad Elmallah, Hamdy Mubarak, Zaid Alyafeai, Mohamed Anwar, Haonan Li, and 24 others. 2025. BALSAM: A platform for benchmarking Arabic large language models. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 258–277, Suzhou, China. Association for Computational Linguistics.

Yazeed Alnumay, Alexandre Barbet, Anna Bialas, William Darling, Shaan Desai, Joan Devassy, Kyle Duffy, Stephanie Howe, Olivia Lasche, Justin Lee, Anirudh Shrinivason, and Jennifer Tracey. 2025. Command r7b arabic: A small, enterprise focused, multilingual, and culturally aware arabic llm. *Preprint*, arXiv:2503.14603.

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 Sibaee, 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.

Mohamed Anwar, Abdelhakim Freihat, George Ibrahim, Mostafa Awad, Abdelrahman Atef Mohamed Ali Sadallah, Gurpreet Gosal, Gokul Ramakrishnan, Sarath Chandran, Biswajit Mishra, Rituraj Joshi, Ahmed Frikha, Etienne Goffinet, Abhishek Maiti, Ali El Filali, Sarah Al Barri, Samujjwal Ghosh, Rahul Pal, Parvez Mullah, Awantika Shukla, and 41 others. 2025. Jais 2: A family of Arabic-centric open large language models. Technical report, IFM.

Yasser Ashraf, Yuxia Wang, Bin Gu, Preslav Nakov, and Timothy Baldwin. 2025. Arabic dataset for LLM safeguard evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5529–5546, Albuquerque, New Mexico. Association for Computational Linguistics.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, and 6 others. 2024. ALLaM: Large language models for Arabic and English. *Preprint*, arXiv:2407.15390.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaian, Ali Alammari, Zaki Alawami, and 7 others. 2025. ALLam: Large language models for arabic and english. In *The Thirteenth International Conference on Learning Representations*.

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 (WANLP19)*, Florence, Italy.

Abdelaziz Bounhar and Abdeljalil El Majjodi. 2025. Atlaset dataset for moroccan darija: From data collection, analysis, to model trainings. *Hugging Face Blog*.

Team Cohere, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphael Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, and 207 others. 2025. Command a: An enterprise-ready large language model. *ArXiv*, abs/2504.00698.

Ryan Cotterell and Chris Callison-Burch. 2014. A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 241–245, Reykjavik, Iceland.

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 Expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Shereen ElSayed and Mona Farouk. 2020. Gender identification for egyptian arabic dialect in twitter using deep learning models. *Egyptian Informatics Journal*, 21.

Fadhl Eryani, Nizar Habash, Houda Bouamor, and Salam Khalifa. 2020. A spelling correction corpus for multiple Arabic dialects. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4130–4138, Marseille, France. European Language Resources Association.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. 2025. Fanar: An Arabic-centric multimodal generative AI platform.

Charles F Ferguson. 1959. Diglossia. *Word*, 15(2):325–340.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nizar Habash, Mona Diab, and Owen Rambow. 2012. Conventional Orthography for Dialectal Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 711–718, Istanbul, Turkey.

Nizar Habash, Salam Khalifa, Fadhl Eryani, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, Sara Hassan, Faisal Al shargi, Sakhar Alkhereyf, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh, and Hind Saddiki. 2018. Unified Guidelines and Resources for Arabic Dialect Orthography. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Faris Hijazi, Somayah AlHarbi, Abdulaziz AlHussein, Harethah Abu Shairah, Reem AlZahrani, Hebah AlShamlan, Omar Knio, and George Turkiyyah. 2024. ArabLegalEval: A multitask benchmark for assessing Arabic legal knowledge in large language models. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 225–249, Bangkok, Thailand. Association for Computational Linguistics.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2016. Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, pages 1–31.

Serena Jeblee, Weston Feely, Houda Bouamor, Alon Lavie, Nizar Habash, and Kemal Oflazer. 2014. Domain and dialect adaptation for machine translation into egyptian Arabic. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, pages 196–206, Doha, Qatar.

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.

Salam Khalifa, Nizar Habash, Fadhl Eryani, Ossama Obeid, Dana Abdulrahim, and Meera Al Kaabi. 2018. A morphologically annotated corpus of emirati Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

Juhao Liang, Zhenyang Cai, Jianqing Zhu, Huang Huang, Kewei Zong, Bang An, Mosen Alharthi, Juncai He, Lian Zhang, Haizhou Li, Benyou Wang, and Jinchao Xu. 2024. Alignment at pre-training! towards native alignment for arabic LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Karen McNeil and Miled Faiza. 2011. Tunisian Arabic Corpus : Creating a Written Corpus of an "Unwritten" Language. In *Proceedings of the Workshop on Arabic Corpus Linguistics (WACL)*.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Aissam Outchakoucht and Hamza Es-Samaali. 2021. Moroccan dialect -darija- open dataset. *Preprint*, arXiv:2103.09687.

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, Shahd Abdelmoneim, Anjali Kantharuban, Otba Alsboul, Salima Lamsiyah, Kelly Marchisio, and Kenton Murray. 2026. AMIYA shared task: Arabic Modeling In Your Accent at VarDial 2026. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. 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.

Abdelrahman Sadallah, Junior Cedric Tonga, Khalid Almubarak, Saeed Almheiri, Farah Atif, Chatrine Qwaider, Karima Kadaoui, Sara Shatnawi, Yaser Alesh, and Fajri Koto. 2025. Commonsense reasoning in Arab culture. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7695–7710, Vienna, Austria. Association for Computational Linguistics.

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.

Ahmed Salama, Houda Bouamor, Behrang Mohit, and Kemal Oflazer. 2014. YouDACC: the Youtube Dialectal Arabic Comment Corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1246–1251, Reykjavik, Iceland.

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, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, and 13 others. 2023. Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Guokan Shang, Hadi Abdine, Ahmad Chamma, Amr Mohamed, Mohamed Anwar, Abdelaziz Bounhar, Omar El Herraoui, Preslav Nakov, Michalis Vazirgiannis, and Eric P. Xing. 2025a. Nile-chat: Egyptian language models for Arabic and Latin scripts. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 306–322, Suzhou, China. Association for Computational Linguistics.

Guokan Shang, Hadi Abdine, Yousef Khoubrane, Amr Mohamed, Yassine Abbahaddou, Sofiane Ennadir, Imane Momayiz, Xuguang Ren, Eric Moulines, Preslav Nakov, Michalis Vazirgiannis, and Eric Xing. 2025b. Atlas-chat: Adapting large language models for low-resource Moroccan Arabic dialect. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 9–30, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Mohamedou Cheikh Tourad, Rahaf Alhamouri, Rwaa 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 Alsalka. 2019. Compression vs traditional machine learning classifiers to detect code-switching in varieties and dialects: Arabic as a case study. *Natural Language Engineering*, 26.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

BigScience Workshop. 2024. bloom-7b1 (revision 6232703).

Wajdi Zaghouani and Anis Charfi. 2018. ArapTweet: A Large Multi-Dialect Twitter Corpus for Gender, Age and Language Variety Identification. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Omar F Zaidan and Chris Callison-Burch. 2011. The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic With High Dialectal Content. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 37–41.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine Translation of Arabic Dialects. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 49–59, Montréal, Canada.

A Details Experimental Details

A.1 Continual Pretraining Setup

Across both CPT strategies, we use a batch size of 8 with gradient accumulation over 8 steps. Models are optimized using AdamW with a cosine learning rate schedule and a warm-up ratio of 0.03. Training is conducted in bfloat16 precision with gradient checkpointing enabled. The maximum sequence length is set to 256 tokens. We adopt a higher peak learning rate of 3×10^{-5} for Mixed CPT and for the initial curriculum CPT stage (training on EN and MSA), followed by a reduced learning rate of 1×10^{-5} during the dialect-only curriculum CPT stage to mitigate catastrophic forgetting.

To analyze intermediate adaptation behavior, additional checkpoints are saved at approximately 30% of training progress for both the Mixed and the dialect-only curriculum CPT strategies. Final checkpoints are saved at the end of each training run and used for subsequent instruction tuning and evaluation.

A.2 Instruction Tuning Setup

The instruction tuning setup was standardized across all experiments, using a maximum sequence length of 256, a learning rate of 3×10^{-5}, a batch size of 8 with gradient accumulation over 8 steps, a weight decay of 0.01, and training for a single epoch to avoid the risk of overfitting. Models were optimized using the AdamW optimizer. All instruction tuning experiments were implemented using Hugging Face's Transformers.

B IT Synthetic Data Generation Prompts

MonoGen prompt Given the following question: **text**, generate a natural and contextually appropriate response written in **DA**.

- Only generate the response; do not generate anything else.
- The response must be a reasonable reply to the given prompt.
- Do NOT reply in MSA no matter what DA and only DA.

Where **DA** represents the dialect of the monolingual sample, and **text** refers to the monolingual sample embedded in the DA prompt from the Qasida repository.

XGen prompt Generate an English question where text is the direct answer.

Requirements:

- The question must be specific enough that text is the complete, natural response
- You MUST tell the user to answer in DA
- The question itself MUST be in English.
- Output the English question only while naturally requesting to respond in DA.
- Only say “Respond in DA” without saying please.

Where **DA** is the dialect of the sample, and **text** is the monolingual sample.

	EGY				MAR				PSE				SAU				SYR			
Model	**MonoGen**	**XGen**	**MT**	**Avg**	**MonoGen**	**XGen**	**MT**	**Avg**	**MonoGen**	**XGen**	**MT**	**Avg**	**MonoGen**	**XGen**	**MT**	**Avg**	**MonoGen**	**XGen**	**MT**	**Avg**
ALLaM	0.33	0.03	44	0.27	0.16	0.00	44	0.20	0.06	0.00	43	0.16	0.07	0.00	60	0.22	0.07	0.00	45	0.17
Aya	0.43	0.03	31	0.26	0.39	0.04	27	0.23	0.07	0.02	28	0.12	0.12	0.02	41	0.18	0.24	0.04	29	0.19
Bloom	0.34	0.00	11	0.15	0.22	0.00	11	0.11	0.04	0.00	11	0.05	0.04	0.00	12	0.05	0.08	0.00	11	0.06
CommandR	0.48	**0.05**	30	0.28	0.42	0.05	29	0.25	0.08	0.02	31	0.14	0.10	0.02	45	0.19	0.28	0.03	31	**0.21**
Fanar-B	**0.72**	0.03	29	**0.35**	**0.48**	0.02	27	**0.26**	**0.15**	0.01	27	0.14	**0.20**	0.01	37	0.19	**0.32**	0.02	28	**0.21**
Fanar-I	0.16	**0.05**	41	0.21	0.21	**0.07**	36	0.21	0.03	0.05	39	0.16	0.04	0.04	46	0.18	0.09	**0.05**	39	0.18
Gemma	0.50	0.00	19	0.23	0.37	0.00	20	0.19	0.08	0.00	20	0.09	0.13	0.00	22	0.12	0.18	0.00	20	0.13
Jais-2	0.19	0.05	**51**	0.25	0.21	0.06	**46**	0.24	0.02	0.05	**49**	**0.19**	0.04	0.03	**61**	**0.23**	0.02	0.04	**50**	0.19
LLaMa-B	0.27	0.00	11	0.13	0.28	0.00	12	0.13	0.07	0.00	11	0.06	0.08	0.00	15	0.08	0.13	0.00	12	0.08
LLaMa-I	0.41	**0.05**	28	0.25	0.39	0.04	25	0.23	0.08	**0.06**	27	0.14	0.05	**0.05**	38	0.16	0.13	**0.05**	27	0.15
Qwen3-B	0.51	0.01	27	0.26	0.46	0.01	24	0.24	0.10	0.01	25	0.12	0.09	0.01	36	0.15	0.27	0.01	25	0.18
Qwen3-I	0.11	0.00	17	0.09	0.08	0.00	14	0.07	0.02	0.00	15	0.06	0.03	0.00	22	0.08	0.04	0.00	15	0.06

Table 8: Zero-shot baseline results across for all models across dialects in terms of ADI2 for MonoGen and XGen, chrF++ for MT. Avg denotes the macro-average performance across the three tasks after normalizing chrF++ to $[0, 1]$. B and I denote base and instruct models, respectively. Best scores per dialect are in **bold**.

	EGY				MAR				PSE				SAU				SYR			
Model	**MonoGen**	**XGen**	**MT**	**Avg**	**MonoGen**	**XGen**	**MT**	**Avg**	**MonoGen**	**XGen**	**MT**	**Avg**	**MonoGen**	**XGen**	**MT**	**Avg**	**MonoGen**	**XGen**	**MT**	**Avg**
ALLaM-IT	**0.67**	0.27	46	**0.47**	**0.73**	0.36	41	0.50	**0.21**	0.05	44	**0.23**	**0.24**	0.06	51	0.27	0.15	0.13	41	0.23
Aya-IT	0.62	0.16	**51**	0.43	0.64	0.39	**46**	0.50	0.15	0.01	46	0.21	0.19	0.03	**61**	**0.28**	**0.25**	0.10	**46**	**0.27**
Bloom-Curr-IT	0.64	0.14	41	0.40	0.65	0.20	34	0.40	0.17	0.02	38	0.19	0.08	0.01	51	0.20	0.20	0.09	33	0.21
CommandR-IT	0.53	0.3	16	0.33	0.69	**0.52**	13	0.45	0.14	0.05	22	0.14	0.14	**0.07**	31	0.17	0.22	0.15	11	0.16
Fanar-Curr-IT	0.58	0.15	37	0.37	0.66	0.27	33	0.42	0.15	0.02	39	0.19	0.06	0.02	45	0.18	0.22	0.07	35	0.21
Fanar-Mix-IT	**0.67**	0.13	32	0.37	0.71	0.34	31	0.45	0.17	0.02	32	0.17	0.17	0.03	42	0.21	0.21	0.09	30	0.20
Fanar-IT	0.57	0.12	18	0.29	0.62	0.32	15	0.36	0.18	0.05	19	0.14	0.05	0.03	24	0.11	0.17	0.08	16	0.14
Gemma-Curr-IT	0.51	0.17	22	0.30	0.68	0.28	18	0.38	0.17	0.04	27	0.16	0.06	0.05	27	0.13	0.18	0.10	28	0.19
Gemma-Mix-IT	0.64	**0.41**	33	0.46	0.67	0.30	29	0.42	0.15	0.04	35	0.18	0.22	0.05	42	0.23	0.21	0.05	29	0.18
Jais-2-IT	0.44	0.27	44	0.38	0.44	0.36	39	0.40	0.16	0.04	42	0.21	0.13	**0.07**	50	0.23	0.14	0.11	39	0.21
LLaMa-Curr-IT	0.27	0.12	39	0.26	0.41	0.29	35	0.35	0.10	**0.07**	40	0.19	0.04	0.05	48	0.19	0.13	0.07	38	0.19
LLaMa-Mix-IT	0.29	0.12	38	0.26	0.41	0.27	34	0.34	0.10	0.02	40	0.17	0.04	0.04	47	0.18	0.12	0.09	35	0.19
LLaMa-IT	0.49	0.25	43	0.39	0.72	0.51	38	**0.54**	0.14	**0.07**	40	0.20	0.04	0.03	51	0.19	0.23	**0.20**	37	**0.27**
Qwen3-Curr-IT	0.25	0.07	39	0.24	0.68	0.27	33	0.43	0.16	0.02	38	0.19	0.04	0.01	46	0.17	0.23	0.10	35	0.23
Qwen3-IT	0.25	0.05	39	0.23	0.67	0.27	33	0.42	0.13	0.02	38	0.18	0.06	0.01	46	0.18	0.23	0.09	36	0.23

Table 9: Evaluation results of partially trained models using 30% of the data across dialects in terms of ADI2 for MonoGen and XGen and chrF++ for MT. Avg denotes the macro-average performance across the three tasks after normalizing chrF++ to $[0, 1]$. Curr and Mix denote Curriculum and Mixed CPT, respectively, and IT denotes instruction tuning. Best scores per dialect are in **bold**.

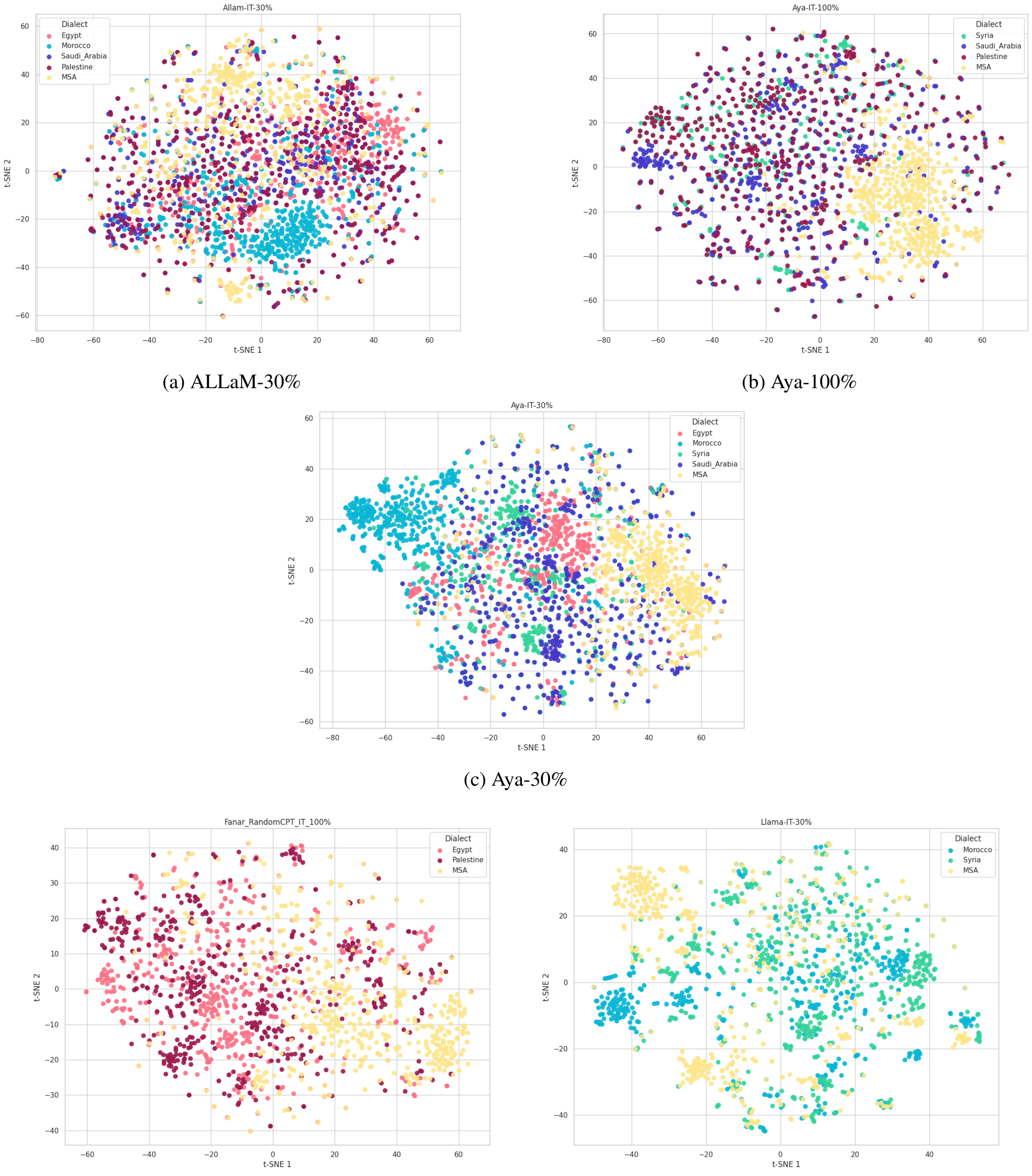
	EGY				MAR				PSE				SAU				SYR			
Model	**MonoGen**	**XGen**	**MT**	**Avg**	**MonoGen**	**XGen**	**MT**	**Avg**	**MonoGen**	**XGen**	**MT**	**Avg**	**MonoGen**	**XGen**	**MT**	**Avg**	**MonoGen**	**XGen**	**MT**	**Avg**
ALLaM-30%	0.67	0.27	46	**0.47**	0.73	0.36	41	**0.50**	**0.21**	0.05	44	**0.23**	**0.24**	0.06	51	0.27	0.15	0.13	41	0.23
ALLaM-50%	0.26	0.17	27	0.23	0.34	0.23	23	0.27	0.11	0.02	27	0.13	0.10	0.07	31	0.16	0.11	0.09	23	0.14
ALLaM-70%	0.24	0.16	23	0.21	0.33	0.19	21	0.24	0.11	0.03	25	0.13	0.08	0.05	29	0.14	0.09	0.08	22	0.13
ALLaM-100%	0.27	0.18	24	0.23	0.35	0.23	23	0.27	0.12	0.05	26	0.14	0.09	0.06	31	0.15	0.10	0.09	22	0.14
Aya-30%	0.62	0.16	51	0.43	0.64	0.39	46	**0.50**	0.15	0.01	**46**	0.21	0.19	0.03	61	**0.28**	0.25	0.10	46	**0.27**
Aya-50%	0.31	0.17	48	0.32	0.43	0.28	43	0.38	0.11	0.02	43	0.19	0.11	0.06	57	0.25	0.16	0.09	43	0.23
Aya-70%	0.31	0.17	47	0.32	0.41	0.27	43	0.37	0.10	0.03	43	0.19	0.11	0.06	56	0.24	0.14	0.11	42	0.22
Aya-100%	0.54	0.18	**53**	0.42	0.65	0.32	**47**	0.48	0.17	0.01	**46**	0.21	0.18	0.02	**63**	**0.28**	0.20	0.09	**48**	0.26
CommandR-30%	0.53	**0.30**	16	0.33	0.69	**0.52**	13	0.45	0.14	0.05	22	0.14	0.14	0.07	31	0.17	0.22	**0.15**	11	0.16
CommandR-50%	0.36	0.17	45	0.33	0.42	0.26	40	0.36	0.13	0.02	43	0.19	0.10	0.05	51	0.22	0.17	0.09	39	0.22
CommandR-70%	0.36	0.19	45	0.33	0.42	0.26	40	0.36	0.12	0.01	44	0.19	0.09	0.05	51	0.22	0.16	0.08	39	0.21
CommandR-100%	0.36	0.21	45	0.34	0.43	0.30	40	0.38	0.13	0.02	43	0.19	0.10	0.06	51	0.22	0.17	0.08	40	0.22
Fanar-Mix-30%	0.68	0.18	44	0.43	0.74	0.26	40	0.47	0.16	0.03	42	0.20	0.21	0.05	50	0.25	0.21	0.10	39	0.23
Fanar-Mix-50%	0.68	0.18	41	0.42	**0.76**	0.23	37	0.45	0.17	0.04	42	0.21	0.18	0.05	43	0.22	**0.26**	0.12	37	0.25
Fanar-Mix-70%	**0.69**	0.19	45	0.44	0.74	0.23	39	0.45	0.19	0.03	44	0.22	0.22	0.04	45	0.24	0.20	0.10	40	0.23
Fanar-Mix-100%	0.68	0.19	46	0.44	0.74	0.24	40	0.46	0.20	0.04	44	0.23	0.22	0.04	44	0.23	0.19	0.11	40	0.23
Gemma-Mix-30%	0.33	0.21	34	0.29	0.40	0.28	31	0.33	0.11	0.05	36	0.17	0.12	**0.08**	41	0.20	0.14	0.10	33	0.19
Gemma-Mix-50%	0.35	0.21	38	0.31	0.44	0.30	34	0.36	0.11	0.05	39	0.18	0.13	**0.08**	43	0.21	0.16	0.1	33	0.20
Gemma-Mix-70%	0.35	0.21	37	0.31	0.42	0.27	34	0.34	0.12	0.05	39	0.19	0.11	**0.08**	41	0.20	0.14	0.11	34	0.20
Gemma-Mix-100%	0.36	0.2	38	0.31	0.43	0.30	34	0.36	0.13	**0.06**	40	0.20	0.11	0.07	43	0.20	0.14	0.10	34	0.19

Table 10: Evaluation results of shortlisted models at instruction-tuning checkpoints (30%, 50%, 70%, and 100%) across dialects in terms of ADI2 for MonoGen and XGen and chrF++ for MT. Avg denotes the macro-average performance across the three tasks after normalizing chrF++ to $[0, 1]$. Best scores per dialect are in **bold**.

Submission	**EGY**	**MAR**	**PSE**	**SAU**	**SYR**
Primary (P)	ALLaM-30%	LLaMa-30%	ALLaM-30%	Aya-30%	Aya-30%
Contrastive-1 (C1)	Aya-30%	Aya-30%	Aya-100%	Aya-100%	Aya-100%
Contrastive-2 (C2)	Fanar-Mix-100%	ALLaM-30%	Fanar-Mix-100%	ALLaM-30%	LLaMa-30%

Table 11: Submitted models per dialect. Percentages denote the training step at which each checkpoint was selected.

C Error Analysis



(a) ALLaM-30%

(b) Aya-100%

(c) Aya-30%

(d) Fanar-Mix-100%

(e) LLaMa-30%

Figure 1: t-SNE projections of the average hidden representations for sentences across the five dialects, categorized by model.