AlexandriaX 2026: The First Shared Task on Dialectal Arabic Machine Translation

Abdellah El Mekki[λ] **AbdelRahim A. Elmadany**[λ] **Samar M. Magdy**[λ]
Saad Ezzini[α] **Mo El-Haj**[β,η] **Mustafa Jarrar**[γ]
Zaid Alyafeai[δ] **Bernard Ghanem**[δ] **Muhammad Abdul-Mageed**[λ,κ]

[λ]The University of British Columbia [κ]Canada Research Chair in NLP and ML
[α]King Fahd University of Petroleum and Minerals [β]Lancaster University [η]VinUniversity
[γ]Hamad Bin Khalifa University [δ]King Abdullah University of Science and Technology
{abdellah.elmekki, muhammad.mageed}@ubc.ca

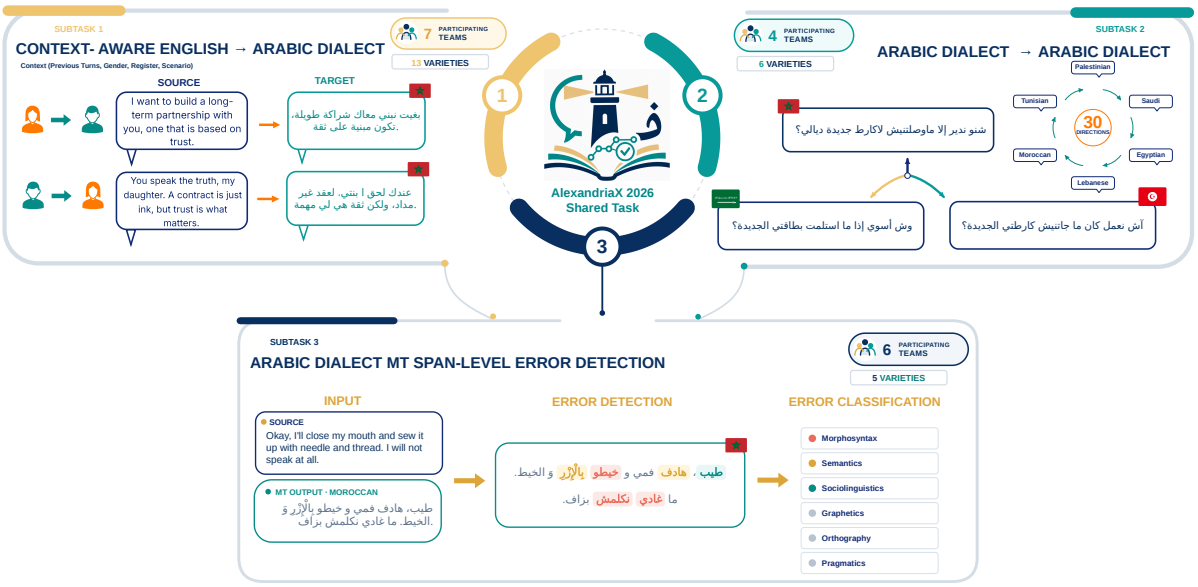

Figure 1: Overview of the three AlexandriaX 2026 subtasks: context-aware English-to-Dialectal Arabic translation, cross-dialect Arabic translation, and span-level detection and classification of translation errors.

Abstract

Dialectal Arabic machine translation (MT) remains challenging despite recent progress in Arabic language technologies, particularly because effective translation requires modeling not only semantic content but also dialectal variation, conversational context, speaker and addressee characteristics, and sociolinguistic appropriateness. Moreover, conventional MT metrics provide limited insight into the linguistic errors produced by dialectal systems. We present the AlexandriaX 2026 Shared Task on Dialectal Arabic MT, which addresses these challenges through three complementary subtasks: (1) context-aware English-to-Dialectal Arabic dialogue translation across 13 Arabic varieties, (2) cross-dialect Arabic translation in the financial domain covering six Arabic dialects, and (3) span-level MT error detection and classification using linguistically motivated error categories across five Arabic varieties. The shared task attracted 38 registrations for Subtask 1, 33 for Subtask 2, and 35 for Subtask 3. Twelve unique teams submitted their system description papers, all of which we accepted for publication. The best system on Subtask 1 achieved 30.42 spBLEU in the constrained track and 33.49 spBLEU in the unconstrained track. On Subtask 2, the top system obtained 28.40 spBLEU. On Subtask 3, the best system achieved an overall score of 49.82, outperforming the 24.29 baseline. Taken together, the results of the leading systems across the three subtasks highlight the benefits of explicit dialect modeling, context-aware generation, retrieval and reranking, and specialized approaches to interpretable MT error analysis. All the resources of AlexandriaX 2026 shared task are publicly available, including data, baselines, and evaluation code on our project page: https://alexandriax.dlnlp.ai.

Proceedings of The Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026) - Volume 2 Shared Tasks, pages 771–785
October 28-29, 2026 ©2026 Association for Computational Linguistics

1 Introduction

Machine translation (MT) for Arabic has improved substantially in recent years, driven by advances in neural models and large language models (LLMs) (Kadaoui et al., 2023). However, this progress remains uneven across Arabic varieties. Most resources and modeling efforts have historically focused on Modern Standard Arabic (MSA) (Bergman and Diab, 2022; Nagoudi et al., 2022; Elmadany et al., 2023), while current systems remain considerably less reliable on Dialectal Arabic (DA) and often struggle to generalize across dialects and sub-dialects (Khalifa et al., 2016; EL Mekki et al., 2026). This gap is especially important because dialects are the primary medium of everyday communication across the Arab world and exhibit substantial variation across regions and communities (Zaidan and Callison-Burch, 2014; Bouamor et al., 2018).

Dialectal Arabic introduces challenges that go beyond standard sentence-level translation. Linguistic choices can depend on the target dialect, speaker and addressee gender, register, domain, persona, and preceding dialogue context. A translation may therefore preserve the source meaning while still sounding unnatural or inappropriate because of incorrect gender agreement, unsuitable formality, dialect mismatches, or the use of MSA constructions where dialectal forms are expected. High-quality dialectal MT consequently requires models that can account for conversational and sociolinguistic context and generalize across dialects with different levels of resource availability (EL Mekki et al., 2026).

Improving translation quality alone is not sufficient; it is also necessary to understand how translation systems fail. Common automatic MT metrics provide useful aggregate scores but offer limited insight into the linguistic nature of errors (Popović and Ney, 2011). This limitation is particularly important for dialectal Arabic, where translation errors may involve semantics, morphology, orthography, pragmatics, or sociolinguistic appropriateness (Magdy et al., 2026). Manual error analysis can capture these distinctions, but it is costly and difficult to scale.

There is therefore a need for automatic and interpretable translation-quality analysis that can identify erroneous spans and characterize the types of errors being made. The Linguistically Motivated Multidimensional Quality Metrics (LQM) framework provides a basis for this goal by organizing translation errors into sociolinguistic, pragmatic, semantic, morphosyntactic, orthographic, and graphetic categories (Magdy et al., 2026).

Motivated by these challenges, the AlexandriaX 2026 Shared Task on Context-Aware Dialectal Arabic MT and MT Evaluation focuses on context-aware English-to-Dialectal Arabic translation, Arabic cross-dialect translation, and automatic span-level translation error detection and classification. Together, these tasks aim to advance MT systems that are dialect-aware, context-sensitive, robust across Arabic varieties, and accompanied by interpretable mechanisms for automatically characterizing their errors.

Across the three subtasks, results highlight the value of dialect-specific adaptation, context-aware generation, retrieval and candidate selection, and hybrid approaches to error analysis.

The remainder of this paper is organized as follows: Section 2 describes the AlexandriaX 2026 Shared Task, while Section 3 outlines the participation rules and evaluation framework. Section 4 presents the participating teams alongside their results, and Section 5 discusses the findings. Finally, Section 6 concludes the paper. In addition, Appendix A reviews related work in the literature.

2 Task Description: AlexandriaX 2026

The objective of the AlexandriaX 2026 Shared Task is to advance the development and evaluation of Machine Translation (MT) systems for Dialectal Arabic through three complementary subtasks: context-aware English-to-Dialectal Arabic dialogue translation, cross-dialect Arabic MT and dialectal Arabic MT error detection and classification. Figure 1 provides an overall description of the three subtasks along with a data point example from each subtask. The first subtask requires systems to translate English dialogue turns into one of 13 Arabic varieties while leveraging conversational context and metadata such as domain, speaker and addressee gender, and persona information. The second subtask focuses on translating short financial-domain queries across six Arabic varieties, including both seen and unseen dialect pairs, to evaluate cross-dialect generalization. The third subtask focuses on automatic MT error analysis for five Arabic dialects, where systems identify erroneous spans in translated text and classify them into six LQM-inspired linguistic categories.

2.1 Subtask 1: Context-Aware Dialectal Arabic Dialogue Machine Translation

The goal of this subtask is to encourage the development of MT systems capable of generating natural and contextually appropriate Dialectal Arabic from English dialogue. Unlike conventional sentence-level MT, systems are expected to make use of the broader conversational context and available metadata, including the target dialect, domain, speaker and addressee gender, and persona information (role, profession). The subtask covers 13 Arabic varieties representing Egypt (EG), Jordan (JO), Lebanon (LB), Libya (LY), Mauritania (MR), Morocco (MA), Oman (OM), Palestine (PS), Saudi Arabia (SA), Sudan (SD), Syria (SY), Tunisia (TN), and Yemen (YE). It further spans 11 domains, including agriculture, commerce, education, healthcare, legal and financial services, logistics, workplace communication, and tourism.

Produced translations should preserve the meaning of the English source while sounding natural in the target dialect and appropriate with respect to lexical, morphological, pragmatic, and sociolinguistic choices. The subtask includes two tracks: a *constrained track*, in which participants may only use the Alexandria data (EL Mekki et al., 2026) and systems with at most 5B parameters, and an *unconstrained track*, which allows the use of external data and models without a parameter limit.

Data Collection and Annotation The data for this subtask are derived from the **Alexandria** dataset (EL Mekki et al., 2026), a large-scale English-Dialectal Arabic parallel corpus consisting of multi-turn conversational scenarios. The dataset was constructed through a combination of controlled LLM-based source generation and community-driven human translation and revision.

More specifically, the Alexandria data was created through three stages: English conversation generation, human dialectal Arabic translation, and revision and quality control. English dialogues were generated with Gemini-2.5 Pro (Comanici et al., 2025) using country-, domain-, persona-, and gender-specific prompts. Native or primary dialect speakers then translated the dialogues into natural local Arabic, preserving meaning, register, gender, and appropriate code-switching. Finally, a second contributor from the same country reviewed each translation for semantic accuracy, dialectal authenticity, consistency, and linguistic quality.

The final Alexandria resource comprises approximately 107K parallel turns across 13 Arab countries and 11 domains. For the shared task, training and development data are provided to participants for system development and local evaluation. The final evaluation was performed on a private test set covering approximately 1,000 turns per dialect, and the participants had access only to its English source. Importantly, Libyan and Sudanese Arabic are absent from the training and development data but included in the test set, allowing the subtask to evaluate generalization to dialectal varieties unseen during development.

2.2 Subtask 2: Cross-Dialect Arabic Machine Translation

The primary objective of Subtask 2 is to advance MT across Arabic dialects. While dialect-English and dialect-MSA MT have received growing attention, direct cross-dialect translation remains an under-resourced and challenging problem. Arabic dialects diverge substantially across regions in phonology, morphology, lexica, and syntax. At the same time, speakers of different dialects frequently interact in business, commerce, and digital services, creating a practical need for direct cross-dialect translation systems that do not rely on English or MSA as an intermediate pivot.

In this subtask, systems translate short queries from a source Arabic variety into a designated target variety. Translations must preserve semantic intent, domain terminology, and numerical entities while using natural target-dialect morphosyntax and vocabulary, without reverting to MSA or simply copying the source.

Data Collection and Setup Building upon previous efforts in Arabic financial NLP benchmarks, such as AraFinNLP 2024 (Malaysha et al., 2024), the benchmark for Subtask 2 is constructed from an expanded multi-dialect version of the **ArBanking77** dataset (Jarrar et al., 2023), a localized Arabic extension of the Banking77 intent-classification benchmark. The dataset comprises short customer-service queries in the financial and banking domain across MSA and several major regional dialects, including Egyptian, Lebanese, Moroccan, Palestinian, Saudi, and Tunisian Arabic. All queries were localized and translated by native speakers of the corresponding varieties.

The official evaluation was conducted on a blind test set comprising 3,574 instances across 30 directed language-pair combinations (derived from

3,518 distinct source dialectal Arabic queries). To test systems' intrinsic cross-dialect generalization and adaptation abilities, no dedicated parallel training or development splits covering all 30 directed pairs were officially provided. Participants were tasked with developing strategies that generalize across dialects using zero-shot prompting, cross-dialect mining, data augmentation, or multilingual transfer learning.

2.3 Subtask 3: Machine Translation Error Detection and Classification

Subtask 3 focuses on fine-grained error analysis of dialectal Arabic MT. Given an English source sentence and its machine-generated translation into a target Arabic dialect, participating systems are required to identify erroneous span(s) in the translation and assign an error category to each detected span. The task follows the Linguistically Motivated Multidimensional Quality Metrics (LQM) framework of Magdy et al. (2026), covering six error categories: *sociolinguistics*, *pragmatics*, *semantics*, *morphosyntax*, *orthography/writing conventions*, and *graphetics*.

Data Split The data is derived from the LQM resource (Magdy et al., 2026) and covers five English-to-dialectal-Arabic translation directions: Egyptian (`ENG_EGY`), Emirati (`ENG_UAE`), Mauritanian (`ENG_MAU`), Moroccan (`ENG_MOR`), and Palestinian (`ENG_PAL`). The dataset is split into 1,125 training, 138 development, and 145 test instances. Each training and development instance contains an English source sentence, its Arabic MT output, and one or more error annotations represented by character-level `start` and `end` offsets together with the corresponding LQM category. For the blind test set, the gold error spans and categories were withheld for official evaluation.

3 Rules and Evaluation

3.1 Participation and Submission Guidelines

To ensure fairness during the final submission phase, all registered participants were given access to the training and development data to build and evaluate their models across all subtasks. We refer to this stage as the *development phase*. Subsequently, we released the unlabeled blind test data and gave participants approximately ten days to submit their system predictions. All test-set evaluations were conducted on CodaBench. Except for the constrained track in Subtask 1, participants were permitted to use any models (open- or closed-source) and external resources to build their final systems. Once the test phase concluded, submissions were finalized before the final leaderboard was released. For Subtask 1, participants were asked to specify the track to which they were submitting.

3.2 Evaluation Method

For evaluating submissions, we report the following metrics: For Subtasks 1 and 2, we follow EL Mekki et al. (2026) and use reference-based surface form SacreBLEU metrics: spBLEU (using the FLORES-200 SentencePiece tokenizer) (Costa-jussà et al., 2024) and chrF++ ($\beta = 2$, word order=2) (Popović, 2015). Participants had access to per-dialect breakdowns of these metrics on CodaBench when submitting their systems.

For Subtask 3, we follow the evaluation framework of Magdy et al. (2026), reporting *Exact Match F_1*, *Overlap F_1*, and *Error Class Macro-F_1*. The first two metrics assess span localization under exact and partial overlap criteria, respectively, while *Error Class Macro-F_1* evaluates error-category prediction. We additionally introduce an *Overall Score* as the official ranking metric, providing a single measure that jointly captures span localization and error classification. It is computed as:

$$\textit{Overall Score} = \frac{1}{2}\left(F_{\text{overlap}} + F_{\text{class}}\right),$$

where F_{overlap} denotes *Overlap F_1* and F_{class} denotes *Error Class Macro-F_1*. The two components are weighted equally, favoring systems that perform well on both span detection and error classification.

4 Shared Task Teams and Results

4.1 Participating Teams

The AlexandriaX 2026 shared task attracted significant interest, with 38 teams registering for Subtask 1, 33 teams registering for Subtask 2, and 35 teams registering for Subtask 3. Actual participation rates varied across the subtasks. Seven teams submitted their final predictions for the test phase of Subtask 1, four teams for Subtask 2, and six teams for Subtask 3. Two teams participated in all three subtasks. Table 1 presents the participating teams with references to their submitted system description papers, their affiliations, and the subtasks

Team Name	Affiliation	Subtask 1		Subtask 2	Subtask 3
		C	U		
ALEXIS (Ibrahim et al., 2026)	Alexandria University			✓	
Alkhder (Alkhder et al., 2026)	Sakarya University	✓	✓		
ArabicMTDiagnostics (Yassine, 2026)	Lebanese University				✓
Axiom (Hossain, 2026)	CUET				✓
CUNI (Jon and Bojar, 2026)	CUNI		✓		
FCDS (Aly et al., 2026)	Alexandria University	✓			
L3IA	USMBA			✓	
LahjaMT (Abdallah and El-Beltagy, 2026)	Newgiza University, Nawy AI lab	✓	✓		
NAMAA Community (Eldin et al., 2026)	NAMAA	✓	✓	✓	✓
REGLAT (Labib et al., 2026)	Elsewedy University of Technology				✓
Rosetta (Esmaeil et al., 2026)	Tanta University	✓	✓		
Thakaa (Alamr et al., 2026)	Thakaa	✓	✓	✓	✓
TTLab (Abusaleh et al., 2026)	Goethe University Frankfurt				✓

Table 1: AlexandriaX 2026 test phase participants by subtask and track. ✓ denotes submitted predictions; uncited teams did not submit a system paper. For Subtask 1, **C** and **U** indicate constrained and unconstrained tracks.

to which they submitted their systems. In total, 13 unique teams participated and submitted their final systems, of which 12 also submitted system description papers.

4.2 Baselines

For each subtask, we provided participants with baseline code and reference benchmark scores. For Subtask 1, we used NileChat-3B (El Mekki et al., 2025), fine-tuned on the Alexandria training dataset using QLoRA (Dettmers et al., 2023).

For Subtask 2, we established two complementary baselines: *Leave-as-is* (Identity Copy) and a zero-shot translation using Gemma-4-31B-IT. More details presented in Appendix B.2.

Finally, for Subtask 3, we used NileChat-3B fine-tuned using QLoRA to produce structured outputs corresponding to the LQM categories.

4.3 Submitted Systems and Results

Subtask 1: Context-Aware Dialectal Arabic Dialogue Machine Translation. Most systems adapted an Arabic-aware or multilingual pretrained model to the Alexandria data and conditioned translation on the target dialect, dialogue history, and speaker metadata. Parameter-efficient fine-tuning was the dominant approach: Thakaa, LahjaMT, Rosetta, and Alkhder adapted NileChat-3B-Base using LoRA (Hu et al., 2022) or QLoRA (Dettmers et al., 2023); FCDS fine-tuned Gemma-3-4B-IT (Team et al., 2025); CUNI used Gemma-4-31B-IT (Gemma Team et al., 2026); and NAMAA Community evaluated AraT5v2, NLLB-200, NileChat, Qwen, and Gemma (Alamr et al., 2026; Abdallah and El-Beltagy, 2026; Esmaeil et al., 2026; Alkhder et al., 2026; Aly et al., 2026; Jon and Bojar, 2026; Eldin et al., 2026).

Rank	Team	Average spBLEU	Average chrF++
1	Thakaa	33.49	48.13
2	CUNI	32.50	47.05
3	LahjaMT	28.54	44.02
4	NAMAA Community	27.41	42.58
5	Rosetta	25.09	41.02
–	Baseline	21.17	37.45
6	Alkhder	20.36	36.64

Table 2: Official results for the unconstrained track of Subtask 1. Systems are ranked by mean spBLEU across the 13 target Arabic varieties; mean chrF++ is reported as a secondary metric. Alkhder submitted the same constrained system to both tracks, hence its identical scores in Table 3.

Systems differed in how they represented dialogue context. CUNI translated whole conversations jointly, obtaining an approximately 0.7 chrF++ development gain over turn-level translation. FCDS performed best when trained with the two most recent English turns but decoded with the complete source and model-generated target history. In contrast, Alkhder found in a post-submission analysis that self-generated history performed worse than both gold history and no history, illustrating the risk of error propagation. Rosetta addressed this mismatch by corrupting reference histories during training, while LahjaMT found

Rank	Team	Average spBLEU	Average chrF++
1	Thakaa	30.42	45.42
2	LahjaMT	28.30	43.81
3	FCDS	26.48	42.30
4	Rosetta	26.10	41.79
5	NAMAA Community	23.26	39.03
–	Baseline	21.17	37.45
6	Alkhder	20.36	36.64

Table 3: Official results for the constrained track of Subtask 1. Systems are ranked by mean spBLEU across the 13 target Arabic varieties; mean chrF++ is reported as a secondary metric.

Rank	Team	Global spBLEU	Global chrF++
1	Thakaa	28.40	42.15
2	ALEXIS	25.84	39.02
3	NAMAA Community	24.53	39.04
4	L3IA	14.07	29.75
–	Baseline (Zero-Shot Gemma-4-31B)	24.69	38.44
–	Baseline (Leave-as-is)	12.81	27.31

Table 4: Official results for Subtask 2 (global corpus-level spBLEU and chrF++ evaluated across all 3,574 blind test instances on CodaBench).

only modest gains from metadata and retrieved demonstrations. Overall, context was beneficial when systems accounted for the difference between clean training histories and noisy model-generated histories at inference time.

Inference-time selection was another major source of performance improvement. Thakaa's unconstrained system used candidate-constrained MBR (Kumar and Byrne, 2004): four systems provided selectable translations, while seven additional systems contributed only pseudo-reference evidence. This yielded 33.49 spBLEU, 1.25 points above the best complete candidate. CUNI applied MBR to a 48-candidate pool, improving from 31.42 to 32.47 spBLEU, while its subsequent genetic optimization added only 0.05 points. FCDS similarly improved from 25.52 to 26.48 by selecting among 50 sampled translations. LahjaMT and NAMAA instead routed each dialect to a selected checkpoint-prompt configuration. These results show that selection can substantially improve translation, but it depends strongly on candidate diversity: several correlated systems may reinforce one another without providing independent evidence.

Dialect-matched supervision was particularly important for Libyan and Sudanese Arabic, which were absent from the original training and development splits. Thakaa routed these dialects to a fold-in LoRA adapter trained with their released public-test examples, increasing constrained-track spBLEU from 29.59 to 30.42. LahjaMT's external-data specialists produced a further 0.24-point gain in the unconstrained track. Broader external-data use was less reliable: Rosetta's MADAR (Bouamor et al., 2018) and PADIC (Meftouh et al., 2015) adaptation improved only Libyan and Moroccan while reducing the overall score, and CUNI's synthetic and WMT-derived data did not improve its strongest standalone checkpoint. Across systems, Mauritanian was frequently among the weakest varieties, while code-switching and non-standardized spelling remained important sources of disagreement with the single references. Overall, the results favor task-specific adaptation, carefully designed context, and targeted dialect supervision over model scale or data volume alone.

The official Subtask 1 results, including the baseline results, are presented in Tables 2 and 3, while Tables 7 and 6 present the country-level breakdown scores for the constrained and unconstrained tracks, respectively.

Subtask 2: Cross-Dialect Arabic Machine Translation. A total of four teams submitted predictions for Subtask 2 (Table 4), with three teams submitting accompanying system description papers (Alamr et al., 2026; Ibrahim et al., 2026; Eldin et al., 2026). Participating systems addressed this Subtask through retrieval-guided generation, adaptation on constructed parallel data, and zero-shot prompting.

Thakaa used a frozen Gemma-4-31B-IT model (Gemma Team et al., 2026) and retrieved up to 32 nonparallel examples written in the requested target dialect from the released input pool. The examples supplied target-side vocabulary, morphology, spelling, and style, while a guided prompt asked the model to synthesize a source-faithful translation. Retrieval raised spBLEU from 24.68 to 27.94, and guided synthesis produced the winning score of 28.40. The method was transductive because it used the released test inputs as retrieval evidence. Although reciprocal-neighbor filtering was applied, a retrospective audit found two effective cross-dialect counterparts among the 3,574 prompts.

ALEXIS adapted NLLB-200-distilled-600M (Costa-jussà et al., 2024) on 70,139 examples constructed from ArBanking77. The corpus combined human-written MSA–Palestinian pairs, intent-matched but unverified cross-dialect pairs, and synthetic Egyptian and Lebanese data. Increasing adaptation capacity and training

Rank	Team	Overall Score	Exact Match F_1	Overlap F_1	Error Class Macro-F_1
1	Thakaa	**49.82**	**23.19**	54.26	**45.38**
2	NAMAA Community	46.41	22.39	**54.48**	38.34
3	TTLab	40.91	19.40	46.00	35.82
4	REGLAT	40.21	0.41	42.72	37.70
5	Axiom	39.26	16.64	48.13	30.40
6	ArabicMTDiagnostics	34.84	12.22	38.18	31.51
-	Baseline	24.29	2.94	28.39	20.18

Table 5: Official results for Subtask 3. Systems are ranked by the Overall Score, computed as the mean of Overlap F1 and Error Class Macro-F1. Exact Match F1 is reported as an additional measure of span-localization accuracy.

Team	EG	JO	LB	LY	MA	MR	OM	PS	SA	SD	SY	TN	YE	Avg.
Thakaa	37.54	38.80	**35.75**	29.45	27.57	**20.18**	**38.10**	**36.04**	**36.95**	30.32	**42.92**	**33.16**	**28.57**	**33.49**
CUNI	**38.29**	**40.10**	34.69	**30.85**	**28.11**	18.79	29.31	35.36	36.05	**32.38**	38.68	32.79	27.12	32.50
LahjaMT	32.59	34.93	31.81	24.70	24.29	18.00	28.61	31.69	31.39	25.53	37.02	29.62	20.85	28.54
NAMAA Community	32.52	33.92	29.53	23.75	22.83	15.89	26.18	30.26	30.80	25.66	35.27	27.25	22.51	27.41
Rosetta	30.41	32.26	28.09	20.14	21.84	10.83	24.78	29.56	29.58	20.49	33.64	24.50	20.11	25.09
Baseline	26.22	27.61	24.34	18.18	14.97	9.83	20.33	24.27	25.22	18.47	29.34	19.65	16.73	21.17
Alkhder	25.69	26.57	21.92	17.82	18.10	10.19	19.08	21.57	22.56	17.21	26.45	21.76	15.79	20.36

Table 6: Subtask 1 spBLEU scores by target Arabic variety for the unconstrained track. Avg. is the arithmetic mean across the 13 varieties. Alkhder's scores match Table 7 because the same system was submitted to both tracks.

duration yielded larger gains than further data expansion, and the official system ranked second with 25.84 spBLEU and 39.02 chrF++.

NAMAA Community used Gemini-2.5-Flash zero-shot, with prompts specifying the source and target varieties and instructions to preserve meaning while avoiding MSA drift and unnecessary copying. A validation and repair stage handled empty, malformed, truncated, or inappropriate identity outputs. The system ranked third with 24.53 spBLEU and 39.04 chrF++.

The fourth participating team, **L3IA**, ranked fourth on the official leaderboard with 14.07 spBLEU and 29.75 chrF++, outperforming the Leave-as-is baseline (12.81 spBLEU / 27.31 chrF++).

Together, these systems show that retrieval, mining, synthesis, and prompting can compensate for missing parallel data, but their effectiveness depends on the provenance and reliability of the resulting evidence. Nonparallel text is useful as target-dialect guidance, mined or synthetic pairs provide weak supervision, and only sufficiently reliable data should be treated as reference translations.

Table 4 presents the official results for the systems submitted to Subtask 2, along with the baselines. A breakdown of the results per target dialect variety is presented in Table 8. A further description of the breakdown is provided in Appendix B.1.

Subtask 3: Machine Translation Error Detection and Classification. The six systems submitted to Subtask 3 followed three main approaches: discriminative span tagging, generative structured prediction, and hybrid pipelines combining generative semantic reasoning with encoder-based localization (Alamr et al., 2026; Eldin et al., 2026; Abusaleh et al., 2026; Labib et al., 2026; Hossain, 2026; Yassine, 2026).

TTLab and Axiom formulated the task as MARBERTv2-based (Abdul-Mageed et al., 2021) BIO sequence labeling. This anchored predictions to the input and avoided generating offsets explicitly. TTLab combined focal loss (Lin et al., 2020), direction-specific thresholds, and light boundary post-processing, ranking third with 40.91 Overall. Axiom used inverse-frequency class weighting and confidence-thresholded decoding, ranking fifth with 39.26. REGLAT separated binary span detection from category classification and ensembled MARBERTv2 and CAMeLBERT-DA through a shared character grid. It ranked fourth with 40.21 Overall, but aggressive gap bridging and dilation produced only 0.41 exact-match F1 despite 42.72 overlap F1, which the authors characterized as a "dilation trap."

ArabicMTDiagnostics retained a generative formulation, fine-tuning NileChat-3B-Base with QLoRA to emit JSON spans and categories. Input-

Team	EG	JO	LB	LY	MA	MR	OM	PS	SA	SD	SY	TN	YE	Avg.
Thakaa	**33.07**	**36.17**	**32.44**	**26.72**	24.14	17.67	**35.51**	**33.22**	**32.86**	**28.17**	**39.75**	29.41	**26.37**	**30.42**
LahjaMT	32.59	34.93	31.81	23.32	**24.29**	**18.00**	28.61	31.69	31.39	23.81	37.02	**29.62**	20.85	28.30
FCDS	29.89	33.30	30.22	21.31	20.64	17.27	26.95	30.44	29.97	19.19	35.34	27.63	22.12	26.48
Rosetta	30.85	33.39	29.70	19.82	21.42	12.13	26.07	30.55	30.56	21.59	34.73	26.12	22.36	26.10
NAMAA Community	26.82	29.93	26.28	15.62	18.46	15.46	23.12	26.27	26.14	15.66	33.97	24.71	19.90	23.26
Baseline	26.22	27.61	24.34	18.18	14.97	9.83	20.33	24.27	25.22	18.47	29.34	19.65	16.73	21.17
Alkhder	25.69	26.57	21.92	17.82	18.10	10.19	19.08	21.57	22.56	17.21	26.45	21.76	15.79	20.36

Table 7: Subtask 1 spBLEU scores by target Arabic variety for the constrained track. Avg. is the arithmetic mean across the 13 varieties.

Team	Egyptian	Lebanese	Moroccan	Palestinian	Saudi	Tunisian	Avg.
Thakaa	**36.47**	**34.98**	24.32	20.48	26.55	16.44	**26.54**
ALEXIS	13.53	14.68	**33.59**	**29.35**	**37.55**	**27.23**	25.99
NAMAA Community	33.18	29.05	18.57	17.76	25.67	14.63	23.14
L3IA	15.10	14.58	16.28	11.51	13.13	7.19	12.97
Baseline (Leave-as-is)	13.96	17.32	8.90	10.94	12.34	5.11	11.43

Table 8: Subtask 2 spBLEU scores grouped by target dialect. Each score aggregates results from the five corresponding source dialects, and Avg. is the unweighted macro-average (arithmetic mean) of the six displayed target-dialect scores, which differs from the global corpus-level score in Table 4 due to dialect sample size disparity.

side character annotations helped with offset generation, while category definitions and examples addressed minority classes. However, union voting over five sampled predictions over-generated spans, resulting in 34.84 Overall and the lowest overlap F1. These experiments showed that structured generation can support joint reasoning, but remains vulnerable to malformed output, inaccurate offsets, and excessive recall.

The two leading systems combined heterogeneous predictors. Thakaa used GPT-5.5 (OpenAI, 2026) to propose semantically motivated error regions and MARBERTv2 to refine their boundaries and corroborate additional spans. Agreement-gated additions, fallback handling, and fragment repair raised Overall from 46.91 for the raw GPT-5.5 output to the winning 49.82, with 54.26 overlap F1 and 45.38 class macro-F1. NAMAA combined MARBERTv2, GPT-5.6, and Gemini-2.5-Flash through 2-of-3 character-level agreement. It obtained the highest overlap F1, 54.48, but ranked second with 46.41 Overall because its class macro-F1 was lower at 38.34.

Two challenges affected all submitted systems. First, the categories were highly imbalanced: sociolinguistics represented approximately 58% of the training spans, whereas graphetics appeared only twice and was absent from development. Systems therefore used focal loss, class weighting, label smoothing, category guidance, and calibration, but rare classes remained difficult. Second, precise localization was a major bottleneck. Generative systems struggled with offsets, while token classifiers were constrained by subword boundaries; overly broad post-processing improved overlap but harmed exact match. Thakaa found that category accuracy reached 75.35% once a span was matched, indicating that discovering and localizing errors was often harder than classifying them. Overall, the strongest systems assigned semantic proposal to large generative models and retained span verification and boundary control with dialect-aware encoders.

Table 5 presents the official results for Subtask 3 submissions alongside the baseline results, while Tables 9 and 10 show the dialect breakdown for span localization and error classification, respectively.

5 Discussion

Across the three subtasks, several common lessons emerged. First, the strongest systems generally benefited more from targeted adaptation and careful inference strategies than from simply increasing model size or adding more data. In Subtask 1, dialect-specific adaptation, effective use of dialogue context, and candidate selection were consistently useful, while broad external-data augmentation produced mixed results. Context was most helpful when systems accounted for the mis-

Team	EGY	MAU	MOR	PAL	UAE	Avg.
Thakaa	**35.19** / 62.79	17.48 / 37.78	**33.63 / 57.79**	12.90 / 53.21	19.05 / **59.64**	**23.65 / 54.24**
NAMAA Community	29.82 / **64.86**	**17.89** / 35.91	24.24 / 56.23	**19.51 / 55.45**	**21.25** / 58.46	22.54 / 54.18
TTLab	23.93 / 58.63	17.78 / 30.32	16.16 / 46.48	18.00 / 38.46	20.00 / 51.64	19.17 / 45.11
REGLAT	2.02 / 43.03	0.00 / 36.45	0.00 / 44.02	0.00 / 44.69	0.00 / 45.27	0.40 / 42.69
Axiom	24.19 / 45.50	12.50 / **38.88**	21.88 / 50.90	13.91 / 49.80	10.96 / 52.05	16.69 / 47.43
ArabicMTDiagnostics	10.69 / 32.98	16.16 / 33.75	6.61 / 34.87	8.93 / 42.69	17.61 / 44.09	12.00 / 37.68
Baseline	3.25 / 24.56	0.00 / 29.88	2.20 / 37.21	2.02 / 16.63	6.06 / 33.88	2.71 / 28.43

Table 9: Subtask 3 span-localization performance by target dialect, reported as Exact Match F_1 / Overlap F_1. Avg. is the arithmetic mean of the five dialect-specific scores. The best result for each metric and dialect is shown in bold.

Team	EGY	MAU	MOR	PAL	UAE	Avg.
Thakaa	**50.00**	40.78	**42.48**	**43.55**	48.98	**45.16**
NAMAA Community	40.35	29.27	36.36	29.27	**52.50**	37.55
TTLab	37.61	26.67	36.36	30.00	44.62	35.05
REGLAT	36.36	**59.52**	16.28	24.74	49.18	37.22
Axiom	32.26	23.21	31.25	19.13	42.47	29.66
ArabicMTDiagnostics	30.53	34.34	21.49	30.36	38.99	31.14
Baseline	16.26	22.00	17.58	16.16	27.27	19.85

Table 10: Subtask 3 error-classification performance by target dialect, measured using Error Class Macro-F_1. Avg. is the unweighted mean of the five dialect-specific scores. The best result for each dialect is shown in bold.

match between clean training histories and model-generated histories at inference time.

Subtask 2 similarly showed that retrieval, synthetic or mined supervision, and prompting can compensate for limited parallel data, but the quality and provenance of the evidence matter. Retrieved target-dialect examples were useful for guiding vocabulary, morphology, and style, whereas noisy or synthetic pairs were less reliable when treated as equivalent to human references.

Subtask 3 highlighted error localization as a major remaining challenge. Generative systems could reason well about error types but struggled with precise offsets, while token-based systems were better anchored to the input but remained sensitive to span boundaries. The strongest approaches therefore combined generative semantic reasoning with encoder-based localization. More broadly, performance varied substantially across dialects, reinforcing the need for dialect-specific modeling rather than treating Dialectal Arabic as a uniform setting.

For future editions, we see several priorities. Evaluation should move beyond single-reference lexical metrics toward multiple references or human/dialect-sensitive evaluation, since valid lexical, orthographic, and code-switching variation can be penalized by current metrics. Future tasks should also expand low-resource and subdialect coverage, and enlarge and balance the Subtask 3 annotations.

6 Conclusion

The AlexandriaX 2026 Shared Task provides a comprehensive benchmark for dialectal Arabic machine translation through three complementary subtasks. Subtask 1 evaluates context-aware translation from English into 13 Arabic varieties, using dialogue history and metadata such as domain, speaker and addressee gender, and persona information. Subtask 2 evaluates direct translation among six Arabic dialects, while Subtask 3 assesses the span-level detection and linguistic classification of machine translation errors across five English-to-dialect directions. Together, these subtasks address translation quality, cross-dialect generalization, contextual appropriateness, and interpretable error analysis.

The shared task received strong community participation from diverse international teams across the different subtasks, with competitive systems that largely outperformed the provided baselines using innovative methods including parameter-efficient and full fine-tuning, LLM prompting, ensemble and routing strategies, data augmentation, and context modeling. Nonetheless, substantial headroom remains and results across the various subtasks underscore that dialectal Arabic MT and fine-grained error localization remain open challenges. The AlexandriaX 2026 Shared Task provides a foundation for systematic evaluation and continued progress in dialectal Arabic machine translation, supporting the development of more robust and inclusive language technologies for Arabic-speaking communities across diverse dialects and contexts.

Limitations

We acknowledge that the AlexandriaX 2026 Shared Task has several limitations:

- **Coverage of Arabic dialects and subdialects.** While the task covers a broad set of Arabic varieties, it primarily models them at the country or regional level, omitting intra-dialectal and subdialectal variation across cities, age groups, and speech communities. As a result, strong country-level performance may not generalize across all speakers of a given variety.

- **Uneven dialect and domain coverage across subtasks.** Dialect and domain distributions vary significantly across subtasks: Subtask 1 covers 13 varieties across multiple domains, Subtask 2 targets fewer varieties in the banking and finance domain, and Subtask 3 evaluates five English-to-dialect directions. Performance here may not reflect broader dialectal MT capabilities across other genres or settings.

- **Limitations of reference-based evaluation.** Subtasks 1 and 2 rely on automatic reference-based metrics (spBLEU and chrF++), which struggle to capture semantic adequacy, pragmatic nuance, and valid dialectal variation. This is especially problematic given the high lexical and orthographic variance inherent to Dialectal Arabic. Furthermore, these metrics exclude key natural phenomena such as code-switching and dialectal language mixing.

- **Limited coverage of fine-grained linguistic phenomena.** Although Subtask 3 offers span-level error detection and LQM-based classification, it is constrained by a small test set of only five varieties and six linguistic categories.

Acknowledgments

Muhammad Abdul-Mageed acknowledges support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2026-07098), the Social Sciences and Humanities Research Council of Canada (SSHRC; 895-2020-1004), Canada Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[1] and UBC ARC-Sockeye. Mustafa Jarrar acknowledges that this work was partially supported by internal research grants from the College of Humanities and Social Sciences at Hamad Bin Khalifa University.

References

Mohamed A. Abdallah and Samhaa R. El-Beltagy. 2026. LahjaMT at AlexandriaX-2026: A Context-Aware English-to-Dialectal Arabic Machine Translation System with Lightweight Routing of LoRA Experts. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

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.

Ali Abusaleh, Bhuvanesh Verma, and Alexander Mehler. 2026. TTLab at AlexandriaX-2026: A Fine-Tuned Surface Tagger for Arabic Machine-Translation Error-Span Detection and Classification. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

Abdullah Alabdullah, Lifeng Han, and Chenghua Lin. 2026. Ara-HOPE: Human-centric post-editing evaluation for dialectal Arabic to Modern Standard Arabic translation. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 157–171, Rabat, Morocco. Association for Computational Linguistics.

Meshal Alamr, Maryam Altuwaijri, Hassan Alqaeri, and Abdullah Aldahlawi. 2026. Thakaa at AlexandriaX-2026: Who Gets the Final Say? Assigning Output Authority in Dialectal Arabic Machine Translation and Error Detection. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

Hasan Alkhder, Maher Pasa, Maria Hanini, Amro Najjar, and Ahmet Zengin. 2026. alkhder at AlexandriaX-

[1] https://alliancecan.ca

2026: Context-Aware English-to-Dialectal Arabic Dialogue Translation with QLoRA-Fine-Tuned NileChat-3B. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

Ranwa Aly, Seif Eldin Maan, Farida Aly, and Isis Ashraf. 2026. FCDS at AlexandriaX-2026: Improving Context Aware Dialogue English To Dialectal Arabic Machine Translation Via Context Strategies And Quality Aware Decoding. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

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).

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Abdellah El Mekki, Houdaifa Atou, Omer Nacar, Shady Shehata, and Muhammad Abdul-Mageed. 2025. NileChat: Towards linguistically diverse and culturally aware LLMs for local communities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10978–11002, Suzhou, China. Association for Computational Linguistics.

Abdellah EL Mekki, Samar M. Magdy, Houdaifa Atou, Ruwa AbuHweidi, Baraah Qawasmeh, Omer Nacar, Thikra Al-hibiri, Razan Saadie, Hamzah A. Alsayadi, Nadia Ghezaiel Hammouda, Alshima Mohammed Alkhazimi, Aya Hamod, Al-Yas Yaqoob Al-Ghafri, Wesam El-Sayed, Asila Ismail al Sharji, Mohamad Ballout, Anas Belfathi, Karim Ghaddar, Serry Sibaee, and 28 others. 2026. Alexandria: A multi-domain dialectal Arabic machine translation dataset for culturally inclusive and linguistically diverse LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32567–32592, San Diego, California, United States. Association for Computational Linguistics.

Fatimah Emad Eldin, Khloud Al Jallad, Mona A. Azim, and Omer Nacar. 2026. NAMAA Community at AlexandriaX-2026: Prompting, Fine-Tuning and Agreement Voting for Dialectal Arabic Translation and Evaluation. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

AbdelRahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Octopus: A multitask model and toolkit for Arabic natural language generation. In *Proceedings of ArabicNLP 2023*, pages 232–243, Singapore (Hybrid). Association for Computational Linguistics.

Ashraf Hatim Elneima, AhmedElmogtaba Abdelmoniem Ali Abdelaziz, and Kareem Darwish. 2024. OSACT6 dialect to MSA translation shared task overview. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 93–97, Torino, Italia. ELRA and ICCL.

Nada Esmaeil, Fathima Rena, Sibi Subhash, Osama Elgendy, Mina Naguib, Salma Omar, and Muhammad Arif. 2026. Rosetta at AlexandriaX-2026: LoRA-Adapted NileChat for Context-Aware Dialectal Arabic Dialogue Translation. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 304 others. 2026. Gemma 4 technical report. *Preprint*, arXiv:2607.02770.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Mar-

tins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Mohamed Motasim Hamed, Muhammad Hreden, Khalil Hennara, Zeina Aldallal, Sara Chrouf, and Safwan AlModhayan. 2025. Lahjawi: Arabic cross-dialect translator. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 12–24, Abu Dhabi, UAE. Association for Computational Linguistics.

Md. Ajwad Hossain. 2026. Axiom at AlexandriaX-2026: Confidence-Thresholded Dialectal Encoder Fine-Tuning for MT Error Span Detection and Classification. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Momen Ibrahim, Mariam E. Hassib, Ahmed Salem, and Marwan Torki. 2026. ALEXIS at AlexandriaX-2026 Subtask 2: Capacity Discovery and Calibrated Preference Optimization for Cross-Dialect Arabic MT. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

Mustafa Jarrar, Ahmet Birim, Mohammed Khalilia, Mustafa Erden, and Sana Ghanem. 2023. ArBanking77: Intent detection neural model and a new dataset in modern and dialectical Arabic. In *Proceedings of the ArabicNLP 2023*, pages 276–287, Singapore (Hybrid). Association for Computational Linguistics.

Josef Jon and Ondřej Bojar. 2026. CUNI at AlexandriaX-2026: Finetuning LLMs for dialectal Arabic Translation. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

Josef Jon, Rawan Bondok, and Ondřej Bojar. 2026. Current state of LLMs for Arabic dialectal machine translation. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script*, pages 329–363, Rabat, Morocco. Association for Computational Linguistics.

Karima Kadaoui, Samar M. Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed Oumar El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. TARJAMAT: Evaluation of bard and ChatGPT on machine translation of ten Arabic varieties. In *Proceedings of ArabicNLP 2023*, pages 52–75, Singapore (Hybrid). Association for Computational Linguistics.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A large scale corpus of Gulf Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4282–4289, Portorož, Slovenia. European Language Resources Association (ELRA).

Shankar Kumar and William 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, Boston, Massachusetts, USA. Association for Computational Linguistics.

Mariam Labib, Bakhom Ramzy, Nsrin Ashraf, Ahmed M. Fetouh, and Hamada Nayel. 2026. REGLAT at AlexandriaX-2026: A Two-Stage Multi-Encoder Framework for Character-Level Error Detection and LQM Classification in Dialectal Arabic Machine Translation. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.

Samar M. Magdy, Fakhraddin Alwajih, Abdellah El Mekki, Wesam El Sayed, and Muhammad Abdul-Mageed. 2026. LQM: Linguistically motivated multidimensional quality metrics for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 40470–40493, San Diego, California, United States. Association for Computational Linguistics.

Sanad Malaysha, Mo El-Haj, Saad Ezzini, Mohammed Khalilia, Mustafa Jarrar, Sultan Almujaiwel, Ismail Berrada, and Houda Bouamor. 2024. AraFinNLP 2024: The first Arabic financial NLP shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 393–402, Bangkok, Thailand. Association for Computational Linguistics.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on PADIC: A parallel Arabic DIalect corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 26–34, Shanghai, China.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. TURJUMAN: A public toolkit for neural Arabic machine translation. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 1–11, Marseille, France. European Language Resources Association.

OpenAI. 2026. GPT-5.5 System Card. https://openai.com/index/gpt-5-5-system-card/. Accessed: 2026-08-31.

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 and Hermann Ney. 2011. Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4):657–688.

Nathaniel R. 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*, pages 1–17, Rabat, Morocco. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Sarah Yassine. 2026. ArabicMTDiagnostics at AlexandriaX-2026: Detecting and Classifying Errors in Dialectal Arabic Machine Translation Outputs. In *Proceedings of The Fourth Arabic Natural Language Processing Conference: Shared Tasks*, Budapest, Hungary. Association for Computational Linguistics.

Omar F. Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

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 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada. Association for Computational Linguistics.

Appendices

A Related Work

Dialectal Arabic translation resources and benchmarks. Early work established the value of explicitly modeling Arabic dialects rather than relying solely on MSA resources. Zbib et al. (2012) constructed crowdsourced Egyptian–English and Levantine–English parallel corpora and demonstrated the benefits of dialectal training data. PADIC subsequently supported translation experiments across Maghrebi and Middle Eastern dialects aligned with MSA (Meftouh et al., 2015), while MADAR provided parallel coverage of 25 city-level dialects and an accompanying lexicon (Bouamor et al., 2018). More recently, TARJAMAT evaluated instruction-tuned LLMs across ten Arabic varieties, finding persistent limitations for varieties with scarce public resources (Kadaoui et al., 2023). Jon et al. (2026) extended the empirical picture through bidirectional English–Arabic evaluation on 16 MADAR dialects using automatic metrics and a small-scale human evaluation. Together, these studies motivate evaluating dialectal competence explicitly rather than inferring it from performance on standard Arabic.

Shared evaluations have also addressed dialectal translation. NADI 2023 introduced dialect-to-MSA translation subtasks (Abdul-Mageed et al., 2023), and OSACT6 organized a dedicated dialect-to-MSA translation shared task (Elneima et al., 2024). AMIYA 2026 evaluated dialectal language modeling across five country varieties, including translation between dialectal Arabic and MSA or English alongside instruction following and generation (Robinson et al., 2026). AlexandriaX builds on these efforts through a coordinated evaluation of context-conditioned English-to-dialect translation, direct cross-dialect translation, and automatic linguistic error analysis.

Conversational context and dialect-conditioned generation. The importance of context extends beyond Arabic. In English–Russian subtitle translation, Voita et al. (2019) identified deixis, ellipsis, and lexical cohesion as sources of inconsistency and showed that context-aware models improve performance on targeted evaluations. This work motivates examining translation decisions that cannot be assessed adequately from isolated sentences. For dialectal Arabic, the Alexandria dataset provides human-translated, multi-turn English–Arabic

conversations spanning 13 countries and 11 domains, with speaker–addressee gender annotations and contributor city metadata (EL Mekki et al., 2026). Subtask 1 uses this resource to evaluate translation conditioned on dialogue history and metadata, including gender and persona, under constrained and unconstrained resource settings. The contribution is the shared evaluation protocol and comparison of participating systems built on Alexandria, rather than the introduction of the underlying corpus.

Direct cross-dialect translation. Direct translation among Arabic dialects has precedents in PADIC's cross-dialect experiments (Meftouh et al., 2015) and, more recently, Lahjawi, which introduced a model for translation among 15 dialects (Hamed et al., 2025). In the financial domain, ArBanking77 localized banking queries into MSA and Palestinian Arabic for intent detection (Jarrar et al., 2023). AraFinNLP 2024 expanded this line of work with multi-dialect intent detection and a cross-dialect translation and intent-preservation subtask (Malaysha et al., 2024). Subtask 2 builds on this resource lineage through an expanded ArBanking77-based benchmark covering six dialects and 30 directed translation pairs. Its blind evaluation, without dedicated parallel training or development splits covering all directions, examines generalization across dialect pairs while requiring preservation of banking intent, terminology, and numerical information.

Linguistically informed translation error analysis. Fine-grained evaluation complements aggregate translation scores by identifying where systems fail. xCOMET combines sentence-level quality estimation with error-span detection, providing localized diagnostics alongside overall scores (Guerreiro et al., 2024). For Arabic, Ara-HOPE introduces a human post-editing evaluation framework with five error categories and a decision-tree annotation protocol for dialect-to-MSA translation (Alabdullah et al., 2026). LQM provides a complementary taxonomy organized around sociolinguistics, pragmatics, semantics, morphosyntax, orthography, and graphetics, supported by expert span annotations of translations involving Arabic dialects (Magdy et al., 2026). Subtask 3 adopts these six linguistic categories for automatic detection and classification of erroneous spans in English-to-dialect translations across five varieties. This formulation makes both localization and linguistic classification explicit evaluation targets. Taken together, the three AlexandriaX subtasks connect context-sensitive generation and cross-dialect transfer with interpretable analysis of the resulting translation errors.

B Evaluation and Results

B.1 Evaluation

Subtask 2 Dialectal Breakdown and Score Aggregation Analysis Table 8 presents the translation quality broken down by target dialect variety. We note that the official primary leaderboard score in Table 4 (e.g., 28.40 spBLEU for Thakaa) is computed as the global corpus-level (micro-averaged) metric over all 3,574 test instances pooled together on CodaBench. In contrast, the summary column in Table 8 reports the unweighted macro-average across the six target dialect categories (e.g., 26.54 spBLEU for Thakaa). This difference stems from two factors: (1) sample size imbalance across dialect categories (e.g., Egyptian contains $N = 705$ instances, whereas Tunisian contains only $N = 158$), and (2) dialect-specific generation difficulty. Tunisian proved the most challenging target dialect for every system except ALEXIS (16.44 spBLEU for Thakaa). In the unweighted macro-average, Tunisian receives an equal one-sixth weight (16.7%), which pulls down the macro-average to 26.54. Thakaa's two highest-scoring target dialects (Egyptian and Lebanese, 36.47 and 34.98) together constitute nearly 40% of the test corpus.

B.2 Baselines

For Subtask 2, we established two complementary baselines:

- **Leave-as-is (Identity Copy) Baseline:** This baseline outputs the source dialect sentence unmodified without any rewriting. In cross-dialect Arabic NLP, regional dialects share substantial vocabulary, morphological structures, and MSA cognates. Consequently, an unmodified input can achieve deceptively high surface overlap (e.g., exceeding 21 spBLEU between closely related varieties like Egyptian and Lebanese), creating an "illusion of translation." Because Subtask 2 provided no parallel training data across all 30 evaluated dialect pairs, the *Leave-as-is* baseline is essential to establish the empirical floor of passive lexical similarity (achieving a global corpus

average of 12.81 spBLEU and 27.31 chrF++) and to diagnose whether submitted models execute genuine dialectal transfer or rely on copy-through shortcuts.

- **Zero-Shot Model Baseline:** To provide a standard model-based reference representing an off-the-shelf translation approach, zero-shot prompting with Gemma-4-31B-IT achieves 24.69 spBLEU and 38.44 chrF++ (and similarly Gemini-2.5-Flash at 24.53 spBLEU / 39.04 chrF++ as evaluated by NAMAA Community), contextualizing the explicit performance gains delivered by participants' dedicated retrieval, adaptation, and fine-tuning pipelines.