Alexandria: A Multi-Domain Dialectal Arabic Machine Translation Dataset for Culturally Inclusive and Linguistically Diverse LLMs

Abdellah El Mekki[1], Samar M. Magdy[1], Houdaifa Atou[3], Ruwa AbuHweidi[4], Baraah Qawasmeh[5], Omer Nacar[6], Thikra Al-hibiri[7], Razan Saadie[8], Hamzah Alsayadi[9], Nadia Ghezaiel Hammouda[10], Alshima Alkhazimi[11], Aya Hamod[12], Al-Yas Al-Ghafri[11], Wesam El-Sayed[13], Asila Al sharji[11], Mohamad Ballout[8], Anas Belfathi[14], Karim Ghaddar[8], Serry Sibaee[15], Alaa Aoun[8], Areej Asiri[7], Lina Abureesh[4], Ahlam Bashiti[4], Majdal Yousef[4], Abdulaziz Hafiz[16], Yehdih Mohamed[17], Emira Hamedtou[17], Brakehe Brahim[17], Rahaf Alhamouri[18], Youssef Nafea[19], Aya El Aatar[3], Walid Al-Dhabyani[20, 21], Emhemed Hamed[22], Sara Shatnawi[23], Fakhraddin Alwajih[1], Khalid Elkhidir[24], Ashwag Alasmari[7], Abdurrahman Gerrio[22], Omar Alshahri[25], AbdelRahim A. Elmadany[1], Ismail Berrada[3], Amir Azad Adli Alkathiri[11], Fadi A Zaraket[8, 26], Mustafa Jarrar[27, 4], Yahya Mohamed El Hadj[26, 28], Hassan Alhuzali[16], Muhammad Abdul-Mageed[1,2]

[1]The University of British Columbia, [2]Canada Research Chair in NLP and ML, [3]Mohammed VI Polytechnic University, [4]Birzeit University, [5]Western Michigan University, [6]Tuwaiq Academy, [7]King Khalid University, [8]American University of Beirut, [9]Ibb University, [10]University of Hail, [11]University of Technology and Applied Sciences, [12]Arab Open University, [13]Minia University, [14]Nantes University, [15]Prince Sultan University, [16]Umm Al-Qura University, [17]University of Nouakchott, [18]Fatabyyano, [19]Independent Researcher, [20]Hadhramout University, [21]Cairo University, [22]Misurata University, [23]Al-Balqa Applied University, [24]University of Khartoum, [25]Sultan Qaboos Higher Centre for Culture and Science, [26]Arab Center for Research and Policy Studies, [27]Hamad Bin Khalifa University, [28]Institut Supérieur du Numérique
{abdellah.elmekki, muhammad.mageed}@ubc.ca

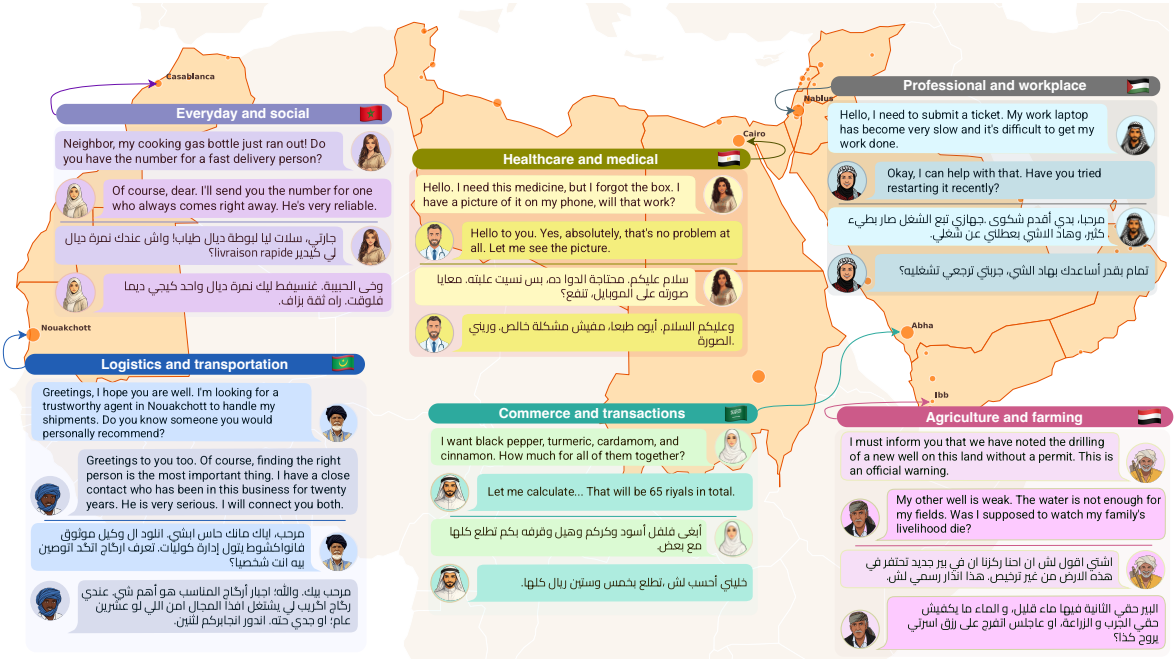

Figure 1: Geographic distribution of Alexandria project participants by city across the Arab world. Point diameter is proportional to participant volume. Representative examples (abbreviated to two-turn interactions) are provided to demonstrate the dataset's coverage across diverse Arabic dialects, domains, and genders.

Abstract

Arabic is a highly diglossic language where most daily communication occurs in regional dialects rather than Modern Standard Arabic (MSA). Despite this, machine translation (MT) systems often generalize poorly to dialectal input, limiting their utility for millions of speakers. We introduce **Alexandria**, a large-scale, community-driven, human-translated dataset designed to bridge this

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32567–32592

gap. Alexandria covers 13 Arab countries and 11 high-impact domains, including *health*, *education*, and *agriculture*. Unlike previous resources, Alexandria provides unprecedented granularity by associating contributions with city-of-origin metadata, capturing authentic local varieties beyond coarse regional labels. The dataset consists of parallel English-Dialectal Arabic multi-turn conversational scenarios annotated with speaker-addressee gender configurations, enabling the study of gender-conditioned variation in dialectal use. Comprising 107K total turns, Alexandria serves as both a training resource and as a rigorous benchmark for evaluating MT and Large Language Models (LLMs). Our automatic and human evaluation benchmarks the current capabilities of Arabic-aware LLMs in translating across diverse Arabic dialects and sub-dialects while exposing significant persistent challenges. The Alexandria dataset, the creation prompts, the translation and revision guidelines, and the evaluation code are publicly available in the following repository: https://github.com/UBC-NLP/Alexandria

1 Introduction

Machine translation (MT) has evolved from a computational convenience into a critical infrastructure for *digital inclusion*, granting diverse populations access to information, technology, and services. Driven by neural sequence-to-sequence models and large-scale training data, recent advances have substantially improved MT quality for many high-resource language pairs (Tiedemann and Thottingal, 2020; Vaswani et al., 2017; Kocmi et al., 2025). Within Arabic, research has also made steady progress on MT involving *Modern Standard Arabic (MSA)*, the lingua franca used in formal writing and broadcast media in the Arab world (Alqudsi et al., 2014; Nagoudi et al., 2022). Yet, a persistent sociolinguistic gap remains: Arabic is *diglossic* and most everyday communication occurs in regional spoken dialects (Ferguson, 1959; Bassiouney, 2020). These dialects can vary widely across countries and even across cities within the same country (Abdul-Mageed et al., 2020), with systematic lexical, morphological, and syntactic differences (Behnstedt and Woidich, 2013). As a result, MT systems trained predominantly on MSA or English-centric resources often generalize poorly to dialectal input, missing vernacular forms and meanings, and thereby limiting the practical utility of MT systems for millions of Arabic speakers (Kadaoui et al., 2023; Harrat et al., 2019).

Recent resources aiming to narrow this gap, most notably PADIC (Meftouh et al., 2015) and MADAR (Bouamor et al., 2018), remain constrained by their design choices and resulting coverage. MADAR is largely organized around travel- and tourism-oriented expressions, while PADIC emphasizes controlled collection and standardized writing practices, improving consistency but limiting naturalistic variation. These choices reduce the extent to which the datasets reflect locally situated usage and fine-grained (e.g., address forms conditioned on gender and social distance, or shifts in register and code choice).

Similar concerns arise even in widely used multilingual evaluation suites such as FLORES+ (Team et al., 2022), which has recently served as a standard benchmark for low-resource MT. Prior work reports issues affecting annotation and translation quality in FLORES+ (Taguchi et al., 2025). For Arabic in particular, analyses suggest that some "dialect" portions may be overly MSA-leaning (e.g., Moroccan Arabic entries reported as essentially MSA), attenuating the very dialect-specific cues the benchmark aims to test (Abdulmumin et al., 2024).

To address this, we introduce **Alexandria**, a large-scale, human-translated, community-driven dataset designed to capture the richness of *dialectal Arabic* across 11 domains with *high social impact*, including health, education, agriculture, and finance. Alexandria includes data from *13 Arab countries*, and associates contributions with city-of-origin metadata, moving beyond coarse regional labels (e.g., "Levantine", "North African") to support analyses at finer geographic granularity. Alexandria is the outcome of a **community project** involving **55 participants** from the **13 Arab countries**. By involving participants tied to specific cities and local varieties, our collection protocol prioritizes authentic, localized realizations of dialectal forms rather than region-level abstractions. The dataset consists of parallel English-Dialectal Arabic **multi-turn conversational scenarios** translated to reflect locally relevant contexts. Crucially, each turn is additionally annotated with speaker-addressee gender

Dataset	# Sentence Pairs / Turns	# Dialects (countries)	Granularity	Src Type	Direction	# Domains	Avg. words	Distinct-2	LC	CS	GD	PR
PADIC (Meftouh et al., 2015)	38K	6	Country	Sentence	Eng ↔ Dialect	1	6.77	0.782	✗	✗	✗	✗
MADAR (Bouamor et al., 2018)	100K	13	City	Sentence	Eng ↔ Dialect	1	5.73	0.768	✗	✗	✗	✗
FLORES+ (Team et al., 2022)	16K	9	Country	Sentence	Eng ↔ Dialect	3	18.39	0.898	✗	✗	✗	✗
Alexandria (ours)	**107K**	**13**	**City**	**Multi-turn**	**Eng ↔ Dialect**	**11**	**13.23**	**0.826**	✓	✓	✓	✓

Table 1: Comparison of Alexandria against existing parallel datasets for Arabic dialects. (LC = Local Context; CS = Code-Switching; GD = gender-direction annotations; PR = persona roles).

configuration (e.g., female-to-male), enabling the study of gender-conditioned variation in dialectal language use.

Alexandria, comprising 107K total turns (34,488 conversations), serves two complementary uses for the NLP community: **(i) Training:** It provides human-translated, domain-diverse conversational data that can be used to train and adapt MT and dialogue-oriented models toward dialectal Arabic in realistic settings. **(ii) Evaluation:** It serves as a benchmark for assessing MT systems and LLM translators under domain, register variation, and speaker-addressee gender configuration conditions, enabling fine-grained analyses of how current models handle dialectal forms and culturally grounded references. To the best of our knowledge, Alexandria is the *first and largest project of its kind*, offering unprecedented granularity in terms of domains, gender annotation, register levels, and city-specific dialectal diversity. By grounding machine translation in the realistic scenarios of the Arab world, we aim to make language technology more accessible, accurate, and culturally inclusive.

2 Related Work

Arabic Diglossia and the MT Gap. Despite significant progress in Neural MT, Arabic continues to struggle with the challenges of diglossia and data scarcity (Zbib et al., 2012; Sajjad et al., 2020; Kadaoui et al., 2023). Early parallel corpora remain limited in scale: PADIC (Meftouh et al., 2015) provides approximately 6,400 parallel sentences per dialect across five Maghrebi and Levantine varieties, while MADAR (Bouamor et al., 2018) translates 2,000 sentences into 25 city-specific dialects. Even the FLORES+ benchmark (Team et al., 2022) includes fewer than 1,000 dev/test sentences per 9 covered Arabic dialects, leaving many dialects underrepresented. Recent efforts, such as WMT24++ (Deutsch et al., 2025), have introduced human-written references and post-edits for several languages, including Egyptian and Saudi Arabic. Furthermore, existing benchmarks are often limited by narrow domains, short sentence lengths, and a lack of context-sensitive translations (Malaysha et al., 2024; Abdulmumin et al., 2024; Taguchi et al., 2025). Furthermore, recent work has addressed gender-aware MT for Arabic (Elaraby et al., 2018; Alhafni et al., 2022) to mitigate gender bias; however, these efforts have been limited to MSA. Our Alexandria dataset addresses these gaps by providing a large-scale corpus covering 13 Arab countries and 11 domains, featuring conversation-based contexts and granular city-level metadata. Table 1 contrasts Alexandria with existing datasets across dialect coverage, domain diversity, and annotation strategies.

Arabic-Capable LLMs and Evaluation. The rise of LLMs has shifted research focus toward adapting models to specific communities to better reflect their unique linguistic and cultural nuances. This has prompted the release of several evaluation datasets, such as Palm (Alwajih et al., 2025a), Pearl (Alwajih et al., 2025b), and AraDice (Mousi et al., 2025), which assess diverse cultural dimensions and modalities. In terms of modeling, several methodologies have recently emerged to showcase the adaptation of LLMs to the Arab world, notably NileChat (El Mekki et al., 2025) and Fanar (Team et al., 2025). We position Alexandria as a vital contribution to this landscape; it serves not only as a benchmark but also as a powerful tool for the adaptation of conversational LLMs tailored to the specific needs of the Arab world.

3 Alexandria Dataset Creation

The Alexandria dataset was created through a six-month community-driven effort involving 55 team members (29 women, 26 men)[1] from 13 Arab countries[2]. Participants were involved to represent local, city-anchored dialectal varieties; the full list of covered sub-dialects appears in Table A.2 (Appendix). Each country team was coordinated by a country lead, who supported member on-boarding and localized annotation guideline examples while pre-

[1] Self-reported; binary categories

[2] 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), Yemen (YE).

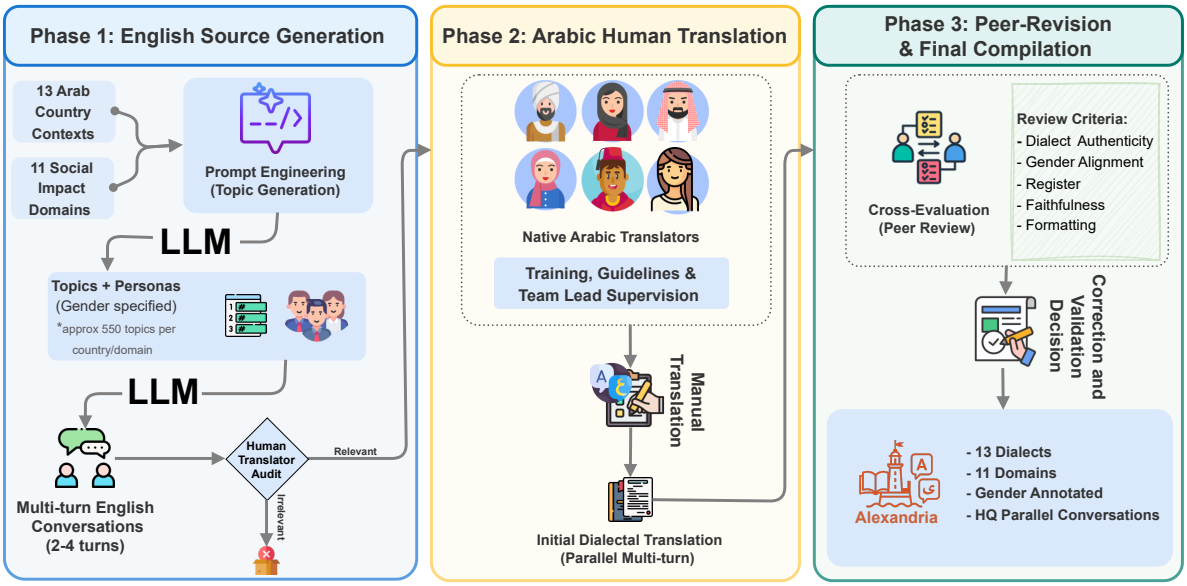

Figure 2: The data creation workflow for the Alexandria dataset. The process illustrates three key phases: *(i)* English source generation, *(ii)* human translation into Dialectal Arabic, and *(iii)* peer-revision and correction.

serving a shared annotation schema across dialects. We employed a structured coordination process with weekly checks to ensure consistent progress. (see Appendix A.2 for project management details).

The dataset consists of turn-aligned parallel multi-turn conversations between English and dialectal Arabic, spanning across 11 domains relevant to public service and everyday life.[3] Additionally, conversations are constructed around persona profiles and include speaker-addressee gender configurations as metadata. Figure 2 illustrates our workflow for creating Alexandria. In the following sections, we describe each phase in detail.

3.1 Alexandria English Sources

3.1.1 English Source Generation

To address common limitations in prior dialectal Arabic MT resources, such as short utterances, narrow topical coverage, and limited conversational context (Bouamor et al., 2018), we construct Alexandria using a controlled source-generation pipeline. Specifically, we use `Gemini-2.5 Pro` (Comanici et al., 2025) to generate multi-turn English conversational scenarios conditioned on the target country and domain. The process is carried out in two phases to promote topical diversity and minimize near-duplicate conversations.

Phase 1: Topic Creation. For each country-domain pair, we use a shared prompt template (with country- and domain-specific variables) to generate 550 topic specifications spanning diverse personas and scenarios. Concretely, we first generate 55 subdomains and then produce 10 topics per subdomain, each paired with a persona specification (e.g., role, speaker/addressee, and gender attributes), yielding $55 \times 10 = 550$ topics per country-domain.

Phase 2: English Conversation Creation. Given these topic specifications, we prompt `Gemini-2.5 Pro` to generate 2-4 turn English dialogues conditioned on the target country local culture and domain. We constrain the model's generations to produce spoken dialogues only that are free of personally identifiable information (PII). To reduce lexical leakage from Arabic into the English sources and to encourage semantic (rather than transliteration-based) transfer, we ask the model to use English paraphrases for culturally specific expressions (e.g., "*God willing*" rather than the transliterated Arabic "*inshallah*").

We iteratively refine our shared prompt template in pilot runs, using feedback from several team members to improve cultural plausibility, domain coverage, and linguistic naturalness. Applied across all 13 countries, our prompt configuration, instantiated with country- and domain-specific variables, produced 6,050 conversations per country under the 2-4 turns constraint ($\sim$ 3 turns/conversation on average). Examples from Phase 1 and Phase 2

[3]We include domains from the set *{Agriculture/Farming, Commerce/Transactions, Construction/Real Estate, Education/Academia, Energy/Resources, Everyday/Social, Healthcare/Medical, Legal/Financial, Logistics/Transportation, Professional/Workplace, Tourism/Hospitality}*.

are shown in Figure A.3 (Appendix).

3.1.2 English Source Quality Assurance

To ensure the high quality, diversity, and cultural validity of the generated English source conversations, we utilized a two-step pipeline consisting of automated screening and human validation. Automated checks filtered the data for format compliance, length constraints, and PII. We also verified lexical and semantic diversity, confirming low redundancy across the generated conversations. Crucially, before translation, human annotators audited the source materials to filter out linguistic artifacts, factual inconsistencies, and cultural mismatches (e.g., dialect-inappropriate idioms or regional inaccuracies). This targeted human review resulted in the exclusion of approximately 2.94% of the generated sentences. A comprehensive breakdown of our diversity metrics, cultural screening procedures, and representations of domain coverage can be found in Appendix A.1.

3.2 Human Dialectal Arabic Translation

Once the English conversations were generated, we distributed them to the corresponding country teams. Each conversation was translated by a primary translator, who produced a single dialectal Arabic translation for each turn. To promote quality and consistency across contributors and dialects, we implemented the following procedures:

Participant Selection and Training. We recruited translators who *(i)* self-identified as primary speakers of the target local dialect and *(ii)* reported advanced proficiency in English. The entire cohort was drawn from academic settings (primarily MA/MSc and PhD students across disciplines). We conducted an initial live training session covering task requirements and guideline conventions; the recording and comprehensive written guidelines were shared for later reference. After one to two weeks, we provided targeted feedback based on sampled translations, focusing on recurrent issues. Throughout the process, country leads conducted ongoing quality control by reviewing submitted translations, providing corrective guidance, and coordinating revisions when needed. Leads also reported periodic progress updates.

Translation Guidelines. We provided detailed translation guidelines instructing participants to render each turn in their local Arabic dialect while preserving the meaning of the English source. Contributors were asked to use Arabic script and to avoid rewriting turns into formal MSA; instead, they were encouraged to use colloquial phrasing typical of their variety, without enforcing a single standardized orthography. Translations were produced at the turn level, with instructions to *(i)* maintain **semantic faithfulness**, *(ii)* adhere to the assigned **persona attributes** (e.g., role/occupation) and **speaker-addressee gender configuration**; and *(iii)* follow an appropriate **social register**. We also provided guidance on *code-switching:* participants could include commonly used borrowed terms (rendered in Latin script) when they are conventional in the target community and lack a natural dialectal alternative, including terms in English or other locally prevalent contact languages such as French or Spanish. While manual translation was prioritized, we permitted the use of AI assistance under narrowly defined conditions, accompanied by rigorous human post-editing to ensure correctness and dialectal authenticity (see Appendix A.5 for analysis of tool usage). We additionally conducted a post-task survey to document translation challenges and contributor feedback (Appendix A.6).

3.3 Translation Revision

Revision Guidelines. To improve data quality, we conducted a peer review and revision phase. Each translated conversation was assigned to a second participant from the same country for cross-evaluation. Reviewers assessed each turn along six dimensions: *dialectal authenticity*, *speaker-addressee gender alignment*, *register appropriateness*, *semantic faithfulness*, *punctuation*, and *code-switching* consistency.[4] Reviewers then issued an overall decision: *Accept*, *Minor edit* (mechanical corrections such as punctuation/typos), or *Major issue* (substantive problems affecting meaning, register, or metadata alignment). When a translation used a regional dialect different from the reviewer's own, the rubric restricted edits to mechanical corrections only; reviewers were instructed not to alter the dialect-specific phrasing or vocabulary. Items flagged with major issues were escalated for follow-up (revision by the original translator and/or adjudication by the country lead). Reviewers assigned a difficulty score used for test-set selection; specifically, this metric was employed to penalize simplistic and trivial turns such as "thank you," thereby

[4]We provided a rubric with examples to guide contributors in this process.

Domain	Levant				Gulf			Nile		Maghreb			
	JO	LB	PS	SY	SA	OM	YE	EG	SD	LY	MA	MR	TN
Agriculture/Farming	825	1140	1770	931	1162	915	529	583	163	231	570	970	481
Commerce/Transactions	750	1004	1595	749	1020	650	579	506	201	160	445	757	401
Construction/Real Estate	859	995	1761	861	1161	974	696	660	225	271	574	673	485
Education/Academia	816	1191	1513	831	1017	1079	563	549	170	220	601	863	551
Energy/Resources	786	1048	1715	928	1177	937	587	625	189	243	447	719	470
Everyday/Social	967	1215	1697	787	1020	888	642	604	175	210	595	824	550
Healthcare/Medical	727	1240	1728	781	1043	895	548	487	164	253	556	948	522
Legal/Financial	693	1006	1566	757	857	753	496	539	177	174	481	642	412
Logistics/Transport	842	1020	1512	950	1234	842	629	646	189	187	593	877	515
Professional/Workplace	845	1220	1810	959	1112	866	549	645	178	253	480	709	526
Tourism/Hospitality	720	1161	1596	884	1004	815	608	608	190	216	567	878	460
Total	**8,830**	**12,240**	**18,263**	**9,418**	**11,807**	**9,614**	**6,426**	**6,452**	**2,021**	**2,418**	**5,909**	**8,860**	**5,373**

Table 2: Post-revision turn statistics in the Alexandria dataset across domains and dialects.

ensuring a more rigorous evaluation set.

Revision Insights. In the revision phase, which was strictly human-only, 68.4% of turns remained unchanged, 30.6% required minor edits, and 1% were flagged for major issues. For turns that were edited, the mean normalized edit distance (turn-level Levenshtein distance divided by character count) was 16.9%. Beyond structural edits, we assessed the quality of the final output across three dimensions: *dialectal authenticity* (9.03/10), *register appropriateness* (9.40/10), and *semantic faithfulness* (9.36/10). These high scores, averaged across all target countries, demonstrate that the final output met strict standards of native-speaker authenticity. Table A.3 (Appendix) illustrates examples of corrections made.

To assess revision reliability, we measured inter-rater agreement on an overlapping subset in which two reviewers from the same country independently reviewed 50-110 shared turns. We report agreement on the three-way decision label (Accept/Minor/Major): the mean exact match rate is 68.2%, and the mean Gwet's AC1 score is 0.65[5].

Preprocessing and Normalization. We applied a three-step cleaning procedure to the revised turns: *(i)* NFKC Unicode normalization, *(ii)* punctuation normalization, and *(iii)* whitespace cleaning.

Final examples of the Alexandria parallel conversations are presented in Table A.1 (Appendix).

3.4 Alexandria Characteristics

Dataset Statistics. The final dataset comprises 34,488 multi-turn conversations, totaling 107K turns. Table 2 summarizes Alexandria after the revision phase, broken down by country and domain. Dataset size varies by country, largely reflecting differences in contributor availability across country teams. Despite these differences, each country covers 11 domains (i.e., no domain is missing for any country/dialect group). On average, a dialectal turn contains 13.23 words. Details about the code-switching coverage and gender distribution are presented in Appendix A.3. In Table 1, we compare Alexandria to prior resources along size, domain coverage, and available annotations.

Data Splits and Release. The Alexandria dataset is partitioned into four splits: *training*, *public development*, *public test*, and *private test*.[6] To ensure balanced representation, the public development and test sets are stratified equally across dialect groups, genders, and translators. Specifically, each country-domain pair contributes ∼100 turns (∼30 conversations) to the public test and ∼ 50 turns to public development, yielding ∼1,100 test turns and ∼550 dev turns per country. The remaining data are allocated to the training and private test sets.

4 Evaluation

4.1 Evaluation Setup

We use Alexandria public test set to evaluate English↔Arabic translation across a diverse set of API access (closed-weight) and open-weight Arabic-capable LLMs. Exploiting Alexandria's conversational structure and metadata (persona role and speaker→addressee gender configuration), we evaluate three input settings: *(i) Turn-level*, translating a single turn in isolation; *(ii) Context-level*, translating a turn given the preceding source-side

[5]We report Gwet's AC1 (Gwet, 2008) instead of Cohen's Kappa due to the high class imbalance in our data (a high prevalence of "Accept" decisions). In such distributions, Cohen's Kappa penalizes high agreement (the "Kappa Paradox"), whereas Gwet's AC1 provides a more robust estimate of chance agreement.

[6]A private test set is withheld to facilitate future open evaluations (e.g., leaderboards, shared tasks).

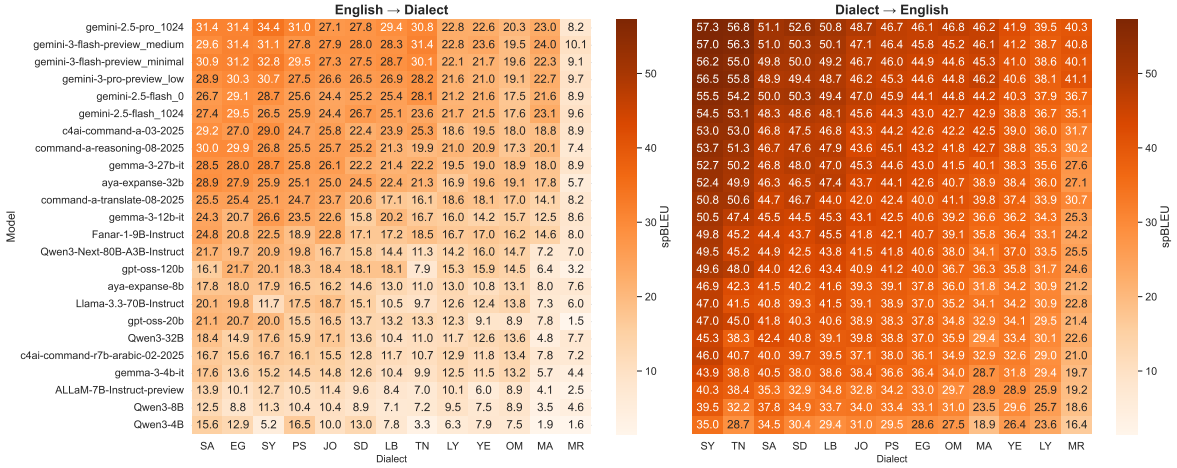

Figure 3: Context-aware MT performance (spBLEU) across 13 dialects. Results reveal a significant directional asymmetry: models perform consistently stronger on Dialect → English (right) than English → Dialect (left). Maghrebi dialects (e.g., MR, MA, TN) remain the most challenging across all models.

dialogue turns; and *(iii) Conversation-level*, translating the entire conversation in a single pass with explicit turn delimiters. In all settings, we prepend the relevant metadata to the input. Figure B.2 (Appendix) provides the generation prompt templates.

We evaluate 24 Arabic-capable LLMs including the `Gemini`, `Qwen 3`, `Gemma`, and `Command A` (Table B.1, Appendix), including both standard and "reasoning" variants. Unless otherwise noted, decoding uses greedy generations (temperature=0).

4.2 Automatic Evaluation

We report reference-based surface metrics using SacreBLEU: spBLEU (Goyal et al., 2022) (SentencePiece with the flores200 tokenizer), and chrF++ (Popović, 2017) (robust for rich morphology). We avoid model-based metrics such as COMET (Rei et al., 2020) due to the limited reliability of dialectal Arabic.

4.3 Human Evaluation

For human evaluation, we focus on English→dialect, which is more sensitive to dialectness and lexical homogenization (i.e., MSA leakage). Native speakers evaluate outputs from six selected LLMs[7] on *(i) semantic adequacy* (5-point Crosslingual Semantic Text Similarity [XSTS] scale (Agirre et al., 2012)), to measure meaning preservation regardless of variety; *(ii) gender accuracy* (Pass/Fail/NA); and *(iii) dialectness & Fluency* (1–5). Further details on the scoring rubrics are provided in Appendix B.1.

[7]These models were selected based on the diversity of their automatic evaluation scores.

We selected 1–3 evaluators per country.[8] Each evaluator rated ∼500 items, stratified across models and domains. For countries with ≥2 evaluators, evaluator pairs additionally rated an overlapping subset of ∼50 items to estimate consistency. We computed inter-rater agreement for three criteria: gender accuracy achieved a mean Gwet's AC1 of 0.970 (averaged across countries), while semantic adequacy and dialectness/fluency yielded Intraclass Correlations (ICC(2,k)) of 0.45 and 0.56, respectively, indicating fair-to-moderate agreement.

5 Results and Discussion

We focus on the context-level conversation setting, which better matches conversational MT: the system translates the current turn given only the preceding dialogue history, without access to future turns (Figure B.1 in Appendix presents comparison against the other settings). While conversation-level translation (full-conversation input) yields higher raw scores, it represents a more permissive, offline setting. Because spBLEU and chrF++ are highly correlated in our experiments (Pearson $r = 0.90$), we report spBLEU in the main text for compactness and include chrF++ in Appendix C.1.

5.1 Automatic Evaluation Results

Per-Dialect Results. Figure 3 reports context-level spBLEU across the 13 country-level dialect groups, with scores averaged over city-level varieties and domains. We observe a clear directional asymmetry: dialect→English achieves consistently

[8]EG, JO, SD, TN, and YE each had one evaluator.

higher scores than English→dialect. Among the evaluated models, the Gemini variants (specifically `Gemini-2.5-pro` and `Gemini-3-flash`) achieve the strongest performance across both directions. Performance also varies substantially by dialect group: models tend to perform best on Egyptian and Levantine varieties (e.g., SY, LB, JO), possibly due to training data availability for these varieties, while Maghrebi varieties pose a greater challenge, with Mauritanian yielding the lowest scores.

Per-Sub-Dialect Results. We further evaluate performance at the sub-dialect (city-level) granularity. We select the three best-performing LLMs from the previous section (based on context-level spBLEU averaged across dialect groups and domains) and evaluate them on each sub-dialect; Figure 4 summarizes the results. Focusing on *within-country* variation, we observe that sub-dialect rankings are broadly consistent across models: while absolute scores differ, the relative ordering of sub-dialects within a country is largely stable, suggesting systematic sub-dialect difficulty that generalizes across model families. Results from other models are provided in Figure C.2 (Appendix).

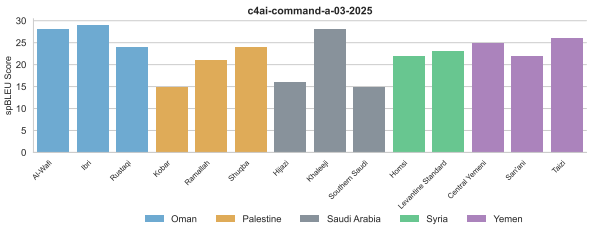

Figure 4: Intra-country performance variance (English → Sub-Dialect). Scores for selected sub-dialects reveal systematic difficulty gaps within countries (e.g., urban vs. rural Palestinian varieties), with consistent model rankings across sub-dialects.

Per-Domain Results. Figure 5 shows English→dialect spBLEU across 11 domains, averaged over countries. Model rankings are highly consistent across domains: top-tier models such as `Gemini-3-flash` and `Command A` perform best throughout, while smaller open-weight models such as `ALLaM-7B` and `Fanar-1-9B` remain in the lower tier. The limited crossing of performance curves suggests little domain-specific specialization, with overall model strength being the main driver of performance. Dialect→English results are shown in Figure C.4 in the Appendix.

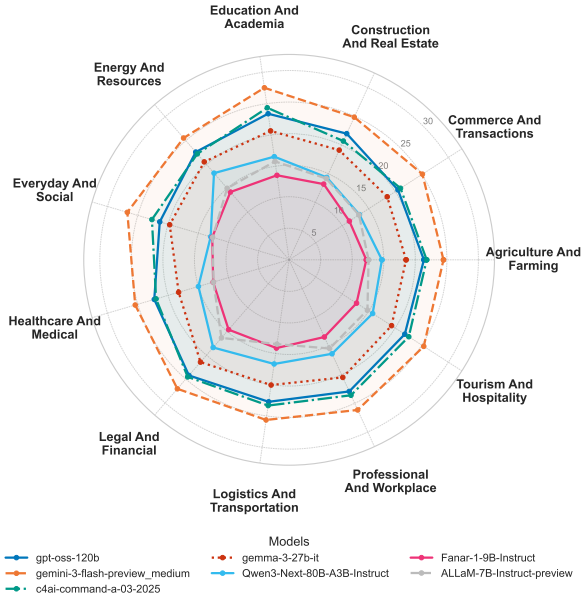

Figure 5: Domain robustness analysis (English → Dialect). The plot illustrates average per-domain spBLEU scores for a subset of models across all 11 domains.

Impact of prompt metadata and comparison against NLLB. Because most of our experiments use conditioned prompts with metadata, we ran a single-turn ablation with three settings: *None*, *Partial* (gender only), and *Full*. We also compared these results with NLLB-200-3.3B (Team et al., 2022), which supports 9 of the 13 Alexandria dialects. As shown in Table 3, the evaluated LLMs outperform NLLB across all supported dialects, even without metadata. The effect of metadata, however, is mixed: it helps in some cases (e.g., c4ai-command-a-03-2025) but is inconsistent across models and dialects. For gemma-3-12b-it, differences across settings are small and sometimes negative, while for aya-expanse-32b the *Full* setting is often worse than *None* or *Partial*. Overall, metadata can help, but its benefits are model- and dialect-dependent rather than uniformly additive.

Additional Factors Influencing Translation Performance. We investigated the impact of specific linguistic characteristics and model configurations on translation quality. First, we found a positive correlation between translation performance and a dialect's lexical overlap with MSA; models perform better on dialects closer to MSA, while structurally distant varieties (e.g., Maghrebi dialects) pose significant challenges. Second, the presence of intra-sentential code-switching (e.g., using Latin script loanwords) substantially degrades translation performance across most evaluated dialects. Detailed correlation analyses and dialect-specific

Model	Metadata	EG	JO	LB	MA	PS	SA	SY	TN	YE
NLLB-200-3.3B	**N/A**	17.16	18.44	17.06	9.82	18.57	17.96	22.24	10.64	11.17
gemma-3-12b-it	**None**	25.54	23.54	20.82	11.33	24.12	25.65	25.79	16.84	17.39
	Partial	25.22	23.77	**21.30**	11.47	23.19	25.97	25.79	16.30	16.91
	Full	25.11	23.14	20.96	11.34	23.39	24.39	24.90	12.00	17.50
aya-expanse-32b	**None**	25.06	22.26	19.77	11.52	21.31	23.54	26.31	15.97	14.01
	Partial	25.32	22.66	20.49	10.96	21.84	23.82	26.25	15.91	14.09
	Full	24.35	21.52	14.51	7.54	21.90	23.27	24.68	8.34	12.99
c4ai-command-a-03-2025	**None**	28.78	24.27	20.20	18.60	**25.34**	28.88	**27.74**	**23.91**	18.58
	Partial	29.07	24.31	20.88	19.19	25.06	28.98	27.70	20.36	19.34
	Full	**29.45**	**25.41**	21.13	**20.01**	25.29	**29.40**	26.96	20.79	**20.51**

Table 3: Single-turn English-to-Dialect spBLEU scores across nine Arabic dialects. LLMs are evaluated under metadata ablations: *None*, *Partial* (gender only), and *Full* (complete metadata). **Bold** denotes the overall column maximum; underline denotes the maximum per model.

breakdowns regarding these linguistic factors are provided in Appendix C.1.1. Finally, we evaluated the impact of explicit LLM reasoning (the "thinking process"). While reasoning generally hurts translation performance for most models, it notably yields improvements for *gemini-3-flash* (e.g., +2.0 spBLEU for English-to-Dialect). Full comparative results for the reasoning experiments are detailed in Appendix C.1.2.

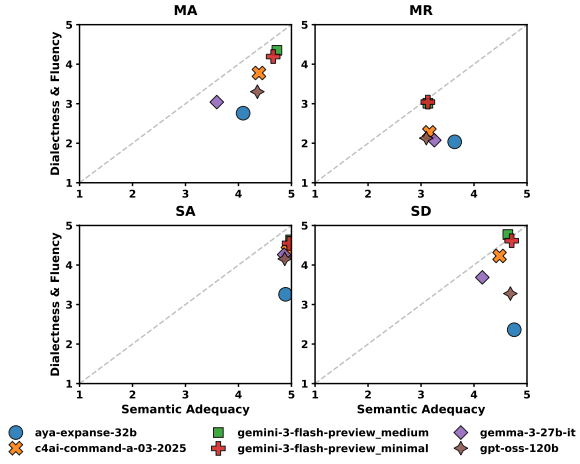

Figure 6: Human evaluation results of Semantic Adequacy vs. Dialectness.

5.2 Human Evaluation Results

Detailed per-model and per-country human evaluation results are provided in Tables C.3, C.4, and C.5 (Appendix). Overall, gender accuracy is high (>90%), suggesting that models effectively adhere to explicit gender constraints when provided in the prompt. In contrast, models exhibit significant performance drops in semantic adequacy and, most notably, dialectness/fluency. Across dialects, average semantic adequacy remains above 3/5, whereas dialectness/fluency is substantially lower, dropping to ∼2/5 for some model-country pairs. Figure 6 plots semantic adequacy against dialectness/fluency for four representative dialect groups: Moroccan, Mauritanian, Saudi, and Sudanese (Figure C.5 in Appendix presents results for the remaining dialects). Most data points lie below the diagonal identity line, indicating that models preserve meaning more reliably than they produce dialect-authentic output. We also observe systematic differences across dialects: Saudi and Sudanese varieties tend to achieve higher scores on both axes, while Mauritanian remains the most challenging, with dialectness often near 2.0 even when semantic adequacy exceeds 3.0. Among the evaluated models, `Gemini-3-flash` and `Command-A` consistently define the Pareto frontier (offering the strongest adequacy-dialectness trade-off), whereas `gpt-oss-120b` demonstrates comparatively lower dialectness/fluency across all regions.

6 Conclusion

In this work, we introduce Alexandria, a culturally inclusive benchmark covering 13 Arab countries, designed to evaluate the dialectal capabilities of Arabic-aware LLMs. Curated by a community of 55 researchers, the dataset comprises 107K turns (34,488 conversations) across 11 domains. Our evaluations reveal critical gaps in existing models regarding regional dialects, technical terminology, and gender alignment. By releasing Alexandria, we provide a robust framework to address these limitations, fostering the development of more accurate and culturally sensitive language technologies.

Limitations

- **Gender Imbalance in Scenarios:** Our dataset exhibits an imbalance in gender transfer directions, specifically a lower frequency of female-to-female interactions (12.60%) compared to other categories. This disparity is an artifact of the Phase 1 generation process; while our prompts explicitly requested diverse personas, the source LLM exhibited a latent bias toward generating mixed-gender or male-dominant scenarios. However, to ensure fair evaluation, the test and development sets were explicitly curated in a balanced manner across gender directions.
- **Technical Lexical Gaps and MSA Leakage:** Annotators reported significant hurdles when translating technical terminology (e.g., in mining, geology, or corporate logistics) that lacks direct dialectal equivalents. In these instances, translators frequently resorted to MSA or code-switching to convey scientific concepts, which may introduce a formal register bias in technical domains.
- **Restricted Closed-Models Evaluation:** We integrated LLMs into our framework following their demonstrated effectiveness in translation tasks, particularly among proprietary architectures. Due to budget constraints, we could not assess the full range of closed-source models and instead focused our evaluation exclusively on the Gemini suite.

Ethics Statement

All Alexandria translations were produced by community participants under a pre-established authorship agreement. Specifically, contributors who translated and revised a minimum of $3,000$ sentences each are included as co-authors to ensure full credit for their substantial labor. Participants who contributed to the project but did not meet this threshold are recognized in the Acknowledgments. To maintain ethical standards, we used English source sentences free of personally identifiable information (PII) and provided participants with rigorous guidelines regarding local norms, data privacy, and informed consent.

Acknowledgments

Muhammad Abdul-Mageed acknowledges support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[9] and UBC Advanced Research Computing-Sockeye.[10]

In addition to the authors who provided translations for this project, we gratefully acknowledge the help and contributions of the following individuals: Mazen Al-Asali, Doaa Qawasmeh, Vatimetou Mohamed Lemin, Abdallah Al-Ameen, Muhammed Saeed, Hawraa Ramadhan, Ro'a Nafi, Ghadeer Shalash, Saja Ayyad, Lina Hamad, Asia Albarghouthi, Manar Shawahnii, Mohammed Anwar Al-Ghrawi, Aminetou Yacoub, Sumayah Alsakiti, Raheeq Mousa, Sondos Khieriah, and Itidal Fares. Also, we thank Mohammed Akallouch for polishing and enhancing the visual presentation of the paper's main figure.

We also acknowledge support from the Google Cloud Research Credits program (Award GCP19980904), which was utilized during data generation and evaluation API calls.

References

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

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo Putini, Miehleketo Mathebula, Matimba Shingange, Tajuddeen Gwadabe, and Vukosi Marivate. 2024. Correcting FLORES evaluation dataset for four African languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 570–578, Miami, Florida, USA. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

[9] https://alliancecan.ca

[10] https://arc.ubc.ca/ubc-arc-sockeye

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022. The Arabic parallel gender corpus 2.0: Extensions and analyses. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1870–1884, Marseille, France. European Language Resources Association.

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.

Arwa Alqudsi, Nazlia Omar, and Khalid Shaker. 2014. Arabic machine translation: a survey. *Artificial Intelligence Review*, 42(4):549–572.

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. 2025a. 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.

Fakhraddin Alwajih, Samar M. Magdy, Abdellah El Mekki, Omer Nacar, Youssef Nafea, Safaa Taher Abdelfadil, Abdulfattah Mohammed Yahya, Hamzah Luqman, Nada Almarwani, Samah Aloufi, Baraah Qawasmeh, Houdaifa Atou, Serry Sibaee, Hamzah A. Alsayadi, Walid Al-Dhabyani, Maged S. Al-shaibani, Aya El aatar, Nour Qandos, Rahaf Alhamouri, and 18 others. 2025b. Pearl: A multimodal culturally-aware Arabic instruction dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23048–23079, Suzhou, China. Association for Computational Linguistics.

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*.

Reem Bassiouney. 2020. *Arabic Sociolinguistics: Topics in Diglossia, Gender, Identity, and Politics, Second Edition*, 2 edition. Georgetown University Press.

Peter Behnstedt and Manfred Woidich. 2013. Arabic dialectology. In *The Oxford Handbook of Arabic Linguistics*. Oxford University Press.

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).

Team Cohere, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, and 26 others. 2025. Command a: An enterprise-ready large language model. *Preprint*, arXiv:2504.00698.

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 3 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

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.

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed Indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387, Goa, India. NLP Association of India.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.

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 10967–10991, Suzhou, China. Association for Computational Linguistics.

Mostafa Elaraby, Ahmed Y. Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. 2018. Gender aware spoken language translation applied to english-arabic. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6.

Charles A. Ferguson. 1959. Diglossia. *WORD*, 15(2):325–340.

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

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 8 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Salima Harrat, Karima Meftouh, and Kamel Smaili. 2019. Machine translation for arabic dialects (survey). *Information Processing & Management*, 56(2):262–273. Advance Arabic Natural Language Processing (ANLP) and its Applications.

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.

Tom Kocmi, Arkady Arkhangorodsky, Alexandre Berard, Phil Blunsom, Samuel Cahyawijaya, Théo Dehaze, Marzieh Fadaee, Nicholas Frosst, Matthias Galle, Aidan Gomez, Nithya Govindarajan, Wei-Yin Ko, Julia Kreutzer, Kelly Marchisio, Ahmet Üstün, Sebastian Vincent, and Ivan Zhang. 2025. Command-a-translate: Raising the bar of machine translation with difficulty filtering. In *Proceedings of the Tenth Conference on Machine Translation*, pages 789–799, Suzhou, China. 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.

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. 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, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 9 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. 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.

Chihiro Taguchi, Seng Mai, Keita Kurabe, Yusuke Sakai, Georgina Agyei, Soudabeh Eslami, and David Chiang. 2025. Languages still left behind: Toward a better multilingual machine translation benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages

20142–20154, Suzhou, China. 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. *Preprint*, arXiv:2501.13944.

Gemma Team. 2025. Gemma 3.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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 Alexandria MT Dataset Creation

A.1 English Source Quality Assurance

We screen the generated English conversations using automated checks (format compliance, length bounds, and heuristic PII detection), followed by targeted human review to remove outputs that violate privacy, realism, or guideline constraints.

Diversity and Redundancy. We assess lexical variety using the proportion of unique bigrams (unique/total) computed per domain, identifying the ratio to range from 0.47 to 0.62. To estimate semantic redundancy, we embed each conversation (mean pooled sentence embeddings)[11], and compute cosine similarity over all English conversation pairs; the mean similarity is 0.20, suggesting limited near-duplication/semantic and topical diversity. Figure A.1 visualizes the diversity of the generated English sources using t-SNE.

Linguistic and Cultural Screening. Because LLM-generated sources can contain artifacts (e.g., unnatural phrasing, implausible cultural details, or mismatches with persona/gender specifications), participants were instructed to audit each source conversation before translation and to flag and discard items that violated the guidelines. On average, 2.94% of the sentences were skipped by the participants and marked as irrelevant.

English Sources Quality Check To verify the quality of the English source conversations prior to translation, we incorporated a human validation step into the translators' workflow. Specifically, we added a dedicated "Comments" column in each participant's assigned spreadsheet so that translators could flag problematic turns and provide concrete feedback. We instructed participants to skip any conversation that *(i)* did not align with the target community's cultural context or *(ii)* included culture-specific references that were incorrect for that country/dialect. This process surfaced several culture-mismatch cases. For instance, in the Jordanian track, an annotator flagged an expression that is common in Syrian usage but not in Jordanian Arabic (e.g., "green light, dead") and noted that the surrounding scenario relied on assumptions that do not reflect the local context. In the Saudi track, a translator identified a factual inconsistency where the conversation described performing Umrah in Medina rather than in Mecca. Such cases were excluded from the translation set to prevent downstream evaluation from being confounded by culturally inaccurate or factually incorrect source content.

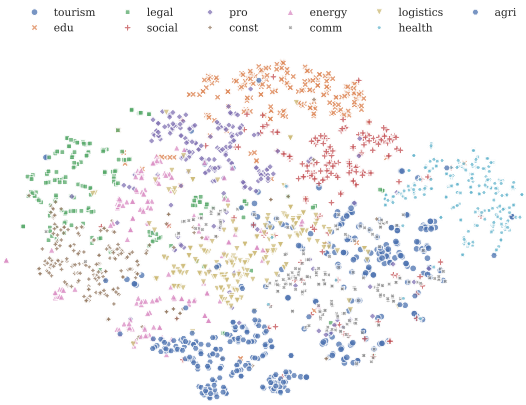

Figure A.1: t-SNE projection of generated English conversations, shown for the Moroccan context as a representative example. Conversations are grouped by domain using source embeddings from our two-phase pipeline; these trends are consistent across the other countries.

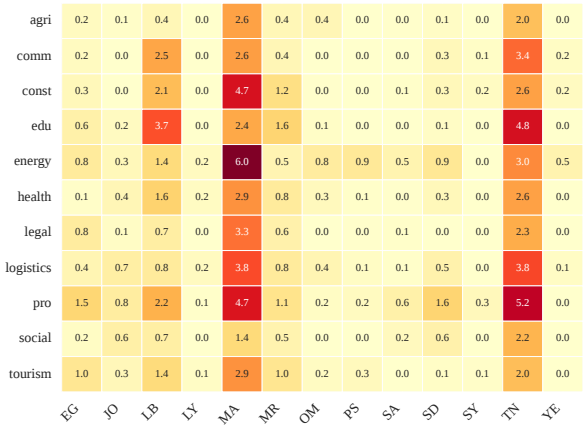

Figure A.2: Average Code-Mixing Index across dialects and domains in Alexandria dataset.

A.2 Project Management and Feedback Loops

To support data quality and maintain steady progress over the project lifecycle, we used a structured coordination process. We held weekly meetings with all participants to review progress and surface recurring issues; feedback from these meetings was used to iteratively refine both the transla-

[11]The model used is `all-MiniLM-L6-v2`, Available at https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

tion guidelines and the annotation platform. Day-to-day coordination occurred through a dedicated Slack workspace, complemented by bi-weekly reminders to keep the workflow on track. Additionally, country leads met every 3–4 weeks to review team-specific progress, address bottlenecks, and consolidate high-level observations and recommendations for subsequent iterations.

A.3 Alexandria Characteristics

Code-Switching Rates. Analysis of code-switching rates reveals that Moroccan and Tunisian varieties consistently exhibit higher code-mixing, while Lebanon shows moderate levels. Most other dialect groups remain low. For a detailed breakdown of the Code-Mixing Index (Das and Gambäck, 2014) across dialects and domains, see Figure A.2.

Gender Direction Distribution. The final version of Alexandria includes four speaker-addressee gender configurations: F→M (33.19%), M→F (32.78%), M→M (21.43%), and F→F (12.60%).

A.4 Annotation Platform

The entire data collection and revision workflow was executed using a spreadsheet-based infrastructure (Google Sheets). The generated English conversations were shuffled and partitioned into batches of 300 conversations (approximately 1,000 turns), with each batch exported to a dedicated sheet. Once a participant finishes a sheet, we can assign them another sheet.

Translation Interface. Each conversation was annotated with metadata indicating the participating personas and the gender direction for each turn. The interface provided translators with a checkbox to discard an entire conversation if any constituent sentence was deemed irrelevant or problematic. Translators entered their translations in a designated column, strictly adhering to the specified gender and social register. An additional field was provided for translators to log specific notes or linguistic observations for each turn. Figure A.4 shows a screenshot from one of the translation sheets.

Revision Interface. For the peer-revision phase, reviewers received separate sheets containing the source English text and the anonymized dialectal translations collected during the previous phase. Reviewers were tasked with populating specific columns for quality scores and, where necessary, providing corrected translations. A notes column was also available for qualitative feedback. Figure A.5 shows a screenshot from one of the revision sheets.

Access Control and Monitoring. To ensure data integrity and privacy, access to each sheet was strictly limited to the assigned participant and their team leader. We implemented a centralized progress tracking system (on Google Sheets) that aggregated statistics from all individual sheets. This system logged daily metrics to monitor throughput at both the country and participant levels. Additionally, team leaders were provided with a customized dashboard view of this global report, enabling real-time oversight of their respective teams' progress.

A.5 AI Usage in Translation Phase

AI as an Auxiliary Assistance Tool To preserve the authenticity of the dialectal data, our protocol strictly defined generative AI as an *auxiliary assistance tool* rather than a primary translation source. The initial protocol prioritized fully manual translation; however, we refined this policy to allow a machine-translation-assisted workflow specifically for technical domains such as **Energy, Mining, and Logistics**. In these cases, AI served as a comprehension and lexical aid to help translators "bootstrap" initial MSA drafts. This workflow was permitted only under the condition that the final output was a product of rigorous manual post-editing and adaptation to ensure dialectal integrity.

Adoption and Frequency of Assistance Based on a survey conducted with **28 participants**, the adoption of these auxiliary tools was widespread; self-reported data indicates that 85.7% of participants utilized an AI system or translation utility as part of their workflow, while only 14.3% relied solely on fully manual translation. Among those who utilized AI for assistance, the degree of reliance varied. While 29.2% of users reported low reliance (1–100 sentences) and an equal percentage reported moderate reliance (100–500 sentences), 37.5% indicated high reliance exceeding 500 sentences for initial drafting.

Tool Selection for Assistance Participants often employed multiple systems simultaneously to verify assistance outputs. Google Translate was the

Figure A.3: Examples of the topic and conversation generation process across six countries. The process defines 55 high-level subdomains, each expanding into 10 specific topics along with their personas (gender & roles); a conversation is then generated for each topic-persona pair.

Dialect / Domain	English Context (Source)	Dialectal Arabic (Target)	Speaker
Sudan Everyday & social life	I've started making a big jug of cold hibiscus tea every morning. It's the only way to get through this heat.	انا بديت أعمل جك عصير كركدي كبير كل صباح، دي الطريقة الوحيدة مع السخانة دي	♀→♂
	That's a very good idea. I should do that for the children. They get so thirsty.	دي فكرة سمحة شديد، لازم عشان كده للأولاد، بيعطشو شديد	♂→♀
Egypt Construction & real estate	Ahmed. The container just cleared customs in Alexandria. Can you have a truck ready to load it tomorrow morning?	يا احمد، الصندوق عدى حالا من الجمارك في اسكندرية. تقدر تجهز مقطورة بكرة الصبح عشان نحمله؟	♂→♂
	Yes, of course. I have a truck available. Send me the release order, and we'll be there first thing, God willing.	ايوة طبعا. انا عندي مقطورة. ابعتلي رقم امر الافراج وهنبقى هناك من النجمة بأمر الله.	♂→♂
Morocco Professional & workplace	Hello, IT support, how may I help you?	الو، Assistance informatique ، كيفاش نقدر نعاونك؟	♀→♂
	Hello, I can't log into my account. I think I've forgotten my password.	الو، منقدرش ندخل لل compte ديالي. واقيلا نسيت ل mot de passe ديالي.	♂→♀
	No problem at all, we can sort this out. Are you in front of your computer right now?	ماشي مشكل غنحاولو نلقاو الحل. واش انتي حدا ل pc دابا؟	♀→♂
Mauritania Commerce & transactions	Excuse me, is the hibiscus juice made fresh here?	عفوا، يعدل هون عصير امبصام اجديد؟	♂→♀
	Yes, madam. We prepare it fresh every morning.	أهيه، madam . نحن انعدلوه كل صبحاية اجديد.	♀→♂
Lebanon Agriculture & farming	Here, I saved some of my local zucchini seeds from last year's harvest. They have the best flavor.	هون، انا احتفظت بشوي بزر كوسى البلدي تعولي من موسم السنة الماضية. طعمتا أطيب شي.	♂→♀
	Oh, thank you so much! You're a lifesaver. These are much better than the ones they sell at the store.	يا سلام، شكرا كتير! أنقذتني. هودي أحسن بكتير من يلي عم يبيعوهن بالسوق.	♀→♂

Table A.1: Cross-dialectal dialogue examples from Alexandria dataset.

conv_id	sentence_id	sentence	dialectal_translation	genders	conversation_direction	participants	whole_conversation	skip_conversation	notes
0-1-95	1	What day works best for you? We have availability tomorrow afternoon or Thursday morning.		Male, Female	Male -> Female	Shopper - Customer Service Agent	CustomerServiceAgent: What day works best for you? We have availability tomorrow afternoon or Thursday morning. Shopper: Tomorrow afternoon is perfect, God willing. Around what time? CustomerServiceAgent: We can set the appointment between 2 PM and 4 PM. Is that suitable? Shopper: Yes, that works for me.	☐	
0-1-95	2	Tomorrow afternoon is perfect, God willing. Around what time?			Female -> Male				
0-1-95	3	We can set the appointment between 2 PM and 4 PM. Is that suitable?			Male -> Female				
0-1-95	4	Yes, that works for me.			Female -> Male				
0-0-208	1	Good morning! Would you like to try some fresh goat cheese? I made it this morning.		female, female	female -> female	Seller, Customer	Seller: Good morning! Would you like to try some fresh goat cheese? I made it this morning. Customer: Oh, it looks delicious. Is that the soft fresh cheese next to it? Seller: Yes, that's our popular fresh cheese. It's very creamy. Which one would you like? I'll give you a good price. Customer: I'll take a small block of the goat cheese and half a kilo of the fresh cheese, please.	☐	
0-0-208	2	Oh, it looks delicious. Is that the soft fresh cheese next to it?			female -> female				
0-0-208	3	Yes, that's our popular fresh cheese. It's very creamy. Which one would you like? I'll give you a good price.			female -> female				
0-0-208	4	I'll take a small block of the goat cheese and half a kilo of the fresh cheese, please.			female -> female				

Figure A.4: Screenshot of the Translation Interface (Google Sheets). Translators are provided with the English source, gender direction, and persona metadata, and enter the dialectal translation in the designated column.

conv_id	sentence_id	sentence	dialectal_translation	conversation_direction	dialectness	gender_accuracy	register_accuracy	faithfulness	correct_translation	overallDecision	testSet	notes
B0-0-3-547	1	For the mid-term exams, I suggest we schedule the science and math exams on separate days. Having them back-to-back is too much pressure on the students.	بالنسبة للامتحانات ديال نص الدورة، كنقترح باش نديرو لامتحان ديال العلوم والعلوم كل فنهار. الا دارو واحد من ورا واحد غادي نضغطو التلاميذ.	male -> female		☐						
B0-0-3-547	2	That's a valid point. Let's look at the calendar. We could move the science exam to Thursday, that would give them a day in between. I agree with that.	عندك الصح. بلاتي نشوفو لcalanderier. نقدرو نحولو الامتحان ديال العلوم نهار الخميس، هكا غايولي عندهم نهار بيناتهم. متافق معاك.	female -> male		☐						
B0-0-2-530	1	Good morning, Architect. I've prepared this Zellige sample for you. Please take a look at the color and the quality of the cut.	صباح الخير، المهندسة. راه وجدت ليك واحد الشانتيو ديال الزليج. عافاك شوفي اللون والقطاع واش مزيانين؟	male -> female		☐						
B0-0-2-530	2	This is excellent. The craftsmanship is perfect and the glaze is consistent. This is exactly what we need. Please proceed with this standard for the fountain area.	هادشي مزيان. الصنعة مخبرة والدهان مقاد كيف كيف. هادشي بالضبط لي محتاجين ليه. عافاك كمل هكا فجيبة النافورة.	female -> male		☐						
B0-0-0-476	1	The proposed minimum price for olives is too low. The cost of fertilizer has gone up, and our work must be valued.	أقل ثمن لي عطونا فالزيتون قليل بزاف. ثمن لانكري طلع، والخدمة ديالنا خاصها تقدر.	male -> male		☐						
B0-0-0-476	2	I understand, but if we set the price too high, the big distributors will just buy from the Spanish cooperatives instead.	انا فاهم هادشي، ولكن الا زدنا فالثمن بزاف، الشاري الكبير غادي يمشي يشري من عند الجمعية الصينليونية بلاصتنا.	male -> male		☐						
B0-0-0-476	3	We need to find a balance. Our quality is better. We should not sell ourselves short.	خاصنا نلقاو توازن. الجودة عندنا حسن. ماخصناش نبيعو راسنا رخاص.	male -> male		☐						

Figure A.5: Screenshot of the Peer-Revision Interface. Reviewers assess translations based on dialectness, gender accuracy, register, and faithfulness, and provide corrections where necessary.

dominant tool, utilized by 87.5% of AI users (21 participants), primarily for its speed in retrieving technical terminology. LLMs served as frequent supplementary aids, with various versions of ChatGPT used by 41.7% of the cohort. Other tools used strictly for lexical retrieval included online dictionaries like Cambridge or Linguee, and alternative LLMs such as Gemini or Qwen.

Primary Modes of Assistance The specific modes of assistance identified by participants reinforce that AI served as a lexical and comprehension bridge rather than a replacement for human translation. The most frequent applications included

Country	Covered Subdialects
Egypt	Egyptian Arabic (Cairene)
Jordan	Jordanian Irbidi
Lebanon	Lebanese Standard
Libya	Libyan Arabic (Misrati/Central)
Mauritania	Mauritanian Hassaniya
Morocco	Moroccan Standard Darija Dialect
Oman	Omani Al-Wafi
	Omani Ibri (Al Nahda)
	Omani Rustaqi
	Omani Seebi (Al Mawaleh)
	Omani Suri (Bani Khuzaymah)
Palestine	Palestinian Albira (Urban)
	Palestinian Arabic (Aboud Falahi)
	Palestinian Arabic (Kobar Falahi)
	Palestinian Arabic (Ni'lin Falahi)
	Palestinian Arabic (Noba Falahi)
	Palestinian Arabic (Ramallah Falahi)
	Palestinian Arabic (Shuqba Falahi)
	Palestinian Arabic (Silwad Falahi)
	Palestinian Arabic (Surif Falahi)
	Palestinian Nabulsi (Urban)
Saudi Arabia	Saudi Arabic (Southern)
	Saudi Arabic Hijazi
	Saudi Arabic Khaleeji
Sudan	Sudanese Standard
Syria	Syrian Arabic (Homsi)
	Syrian Arabic (Levantine Standard)
Tunisia	Tunisian
Yemen	Yemeni Arabic (Central)
	Yemeni San'ani
	Yemeni Taiz

Table A.2: Arabic Subdialects by Country covered in the Alexandria project.

lexical support for specific technical terms (21 mentions) and improving the comprehension of complex source English text (15 mentions). AI was used significantly less for drafting entire sentences (7 mentions), and in all such cases, these drafts served as "base" versions for human-led dialectal adaptation.

The MSA Bridge Strategy A critical finding from translator feedback was the emergence of the *MSA Bridge Strategy*, where AI functioned as an intermediate step. Because commercial AI models frequently struggle to produce authentic local dialects, translators used AI to generate an intermediate MSA version of the text to ensure technical accuracy. This strategy ensured that while AI provided the technical "bridge," the linguistic integrity and final dialectal variety remained entirely human-validated.

A.6 Qualitative Analysis of Translation Challenges

We conducted a qualitative survey of the translators to identify linguistic and non-linguistic friction points in the English-to-Dialectal Arabic translation pipeline. The feedback highlights three primary categories of challenges:

Lexical Gaps and Domain Specificity A significant hurdle reported by annotators was the lack of direct dialectal equivalents for technical and specialized terminology (e.g., in mining, geology, or corporate logistics). Dialectal Arabic is predominantly a spoken register used for daily communication; consequently, translators frequently resorted to MSA or code-switching to convey scientific concepts. Where direct equivalents were absent, translators utilized periphrasis, replacing single English words with descriptive phrases, which introduced structural divergence between the source and target.

Fidelity vs. Fluency Trade-offs The annotation guidelines' requirement for strict semantic faithfulness often conflicted with the goal of producing natural, conversational dialect. Annotators noted that preserving the syntactic structure of English resulted in "translationese"—phrasing that is grammatically correct but pragmatically unnatural in a dialectal context. Idioms and fixed expressions proved particularly difficult to map, requiring significant rewording to maintain the original intent without sacrificing the colloquial tone.

Source Ambiguity and Sociocultural Mismatch Issues inherent to the source text further complicated the process. Annotators cited ambiguity and vague references in the English source as a cause for interpretation delays. Furthermore, cultural disparities posed a distinct challenge; scenarios depicting gender roles uncommon in the target culture (e.g., female electricians or delivery personnel) were perceived as unnatural to the Arab world. This highlights a critical sociolinguistic challenge: accurately translating the *meaning* of a sentence while navigating the *cultural expectations* embedded in the target dialect.

B Evaluation

B.1 Human Evaluation

Our human evaluation assessed English-to-Dialect translations across three decoupled dimensions: *Semantic Adequacy*, *Gender Accuracy*, and *Dialectness & Fluency*. Native speakers of the target dialects followed the specific scoring protocols detailed below.

Country	English Source Sentence	Pre-revised Translation	Revised Translation	Speaker
Yemen	Please look up at the half-moon windows above the doorway.	لو سمحتم شوفوا للطاقات اللي على شكل هلال فوق المدخل.	لو سمحتي شوفي للطاقات اللي على شكل هلال فوق البوابة.	👩→👩
Saudi Arabia	Here is the problem. You can see the water dripping from this pipe behind the washing machine.	هنا المشكلة. تشوف الماي ينقط من هالهوز ورا الي الغسالة .	هنا المشكلة، شايف الموية تنقط من هذي الماصورة ورى المغسلة .	👨→👨
Palestine	Thank you very much. I'm glad I could contribute, especially with coordinating the user feedback sessions with the development team.	شكراً كثير الك. مبسوطة إني قدرت أساهم، وخصوصاً في تنسيق جلسات user feedback مع فريق development .	شكراً كثير الك. مبسوطة إني قدرت أساهم، وخصوصاً في تنسيق جلسات التغذية الراجعة للمستخدمين مع فريق التطوير .	👩→👨
Jordan	Welcome aboard! Let's go over the payments. The company's commission is 25% on every trip.	أهلاً فيك معنا! خلينا نحكي عن الدفعات. عمولة الشركة ٢٥ % على كل رحلة.	أهلاً فيك معنا! خلينا نحكي عن الدفعات. عمولة الشركة خمسة وعشرين بالمية على كل رحلة.	👩→👨
Oman	Good morning. That's a wise precaution. I can come next Tuesday. We will inject each tree directly to protect it.	صباح الخير، زين سويتي كذا، اقدر اجيك يوم الثلاثاء الجاي، انضرب كل نخلة بمرة.	صباح الخير، زين سويتي كذا، اقدر اجيش يوم الثلاثاء الجاي، عشان نحميها انضرب كل نخلة بمرة .	👨→👩

Table A.3: Illustrative examples of prerevised and revised Arabic translations from the human-only revision phase across different countries.

B.1.1 Semantic Adequacy (XSTS)

Annotators evaluated meaning preservation using a 5-point Crosslingual Semantic Textual Similarity (XSTS) scale (Agirre et al., 2012). They were instructed to ignore grammar, style, or dialect errors for this metric.

5 (Perfect) Meaning is identical; all nuances and tone are preserved.

4 (Good) Core meaning is correct; minor nuances (e.g., *huge* vs. *big*) are lost.

3 (Acceptable) Main message conveyed; non-critical details missing or slightly inaccurate.

2 (Poor) Critical information is missing or wrong; meaning is significantly altered.

1 (Wrong) Unrelated to source, contradictory, or gibberish.

B.1.2 Gender Accuracy

Annotators verified adherence to the specified grammatical gender direction (e.g., Male speaker → Female listener).

Pass (1) Correct use of gendered forms (pronouns, verbs, adjectives).

Fail (0) Incorrect gender marking (e.g., masculine *anta* instead of feminine *anti*).

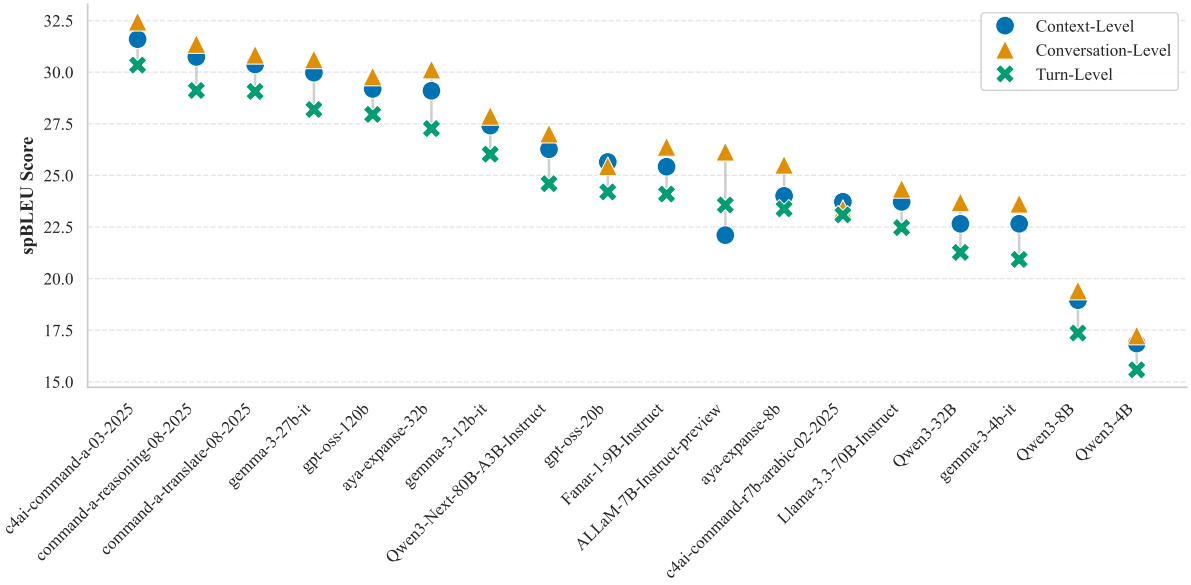

Figure B.1: Comparison of evaluation scenarios across various models. The reported spBLEU scores represent an average across all dialects for both translation directions.

Category	Model	Size
Closed-source LLMs		
Google	Gemini-3-Pro	N/A
	Gemini-3-Flash	N/A
	Gemini-2.5-Pro	N/A
	Gemini-2.5-Flash	N/A
Open-source LLMs		
Google	Gemma-3-27B-IT (Team, 2025)	27B
	Gemma-3-12B-IT (Team, 2025)	12B
	Gemma-3-4B-IT (Team, 2025)	4B
Cohere	Command-A (Cohere et al., 2025)	111B
	Command-A-Translate (Kocmi et al., 2025)	111B
	Command-A-Reasoning (Cohere et al., 2025)	111B
	Command-R7B-Arabic (Alnumay et al., 2025)	7B
	Aya-Expanse-32B (Dang et al., 2024)	32B
	Aya-Expanse-8B (Dang et al., 2024)	8B
Humain	ALLAM-7B-Instruct (Bari et al., 2025)	7B
QCRI	Fanar-1-9B-Instruct (Team et al., 2025)	9B
Qwen	Qwen3-Next-80B-A3B (Yang et al., 2025)	80B
	Qwen3-32B (Yang et al., 2025)	32B
	Qwen3-8B (Yang et al., 2025)	8B
	Qwen3-4B (Yang et al., 2025)	4B
OpenAI	GPT-OSS-120B (OpenAI et al., 2025)	120B
	GPT-OSS-20B (OpenAI et al., 2025)	20B
Meta	Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	70B

Table B.1: List of Arabic-aware open-source and closed-source LLMs evaluated with Alexandria test set.

N/A Sentence is gender-neutral; no specific markers required.

B.1.3 Dialectness & Fluency

Annotators assessed the output by answering the specific question: *"Does this sound like a native speaker of the target dialect (e.g., Moroccan, Levantine)?"*

5 (Native) 100% Authentic. Uses slang/idioms correctly; contains no MSA.

4 (Good) Correct dialect grammar. Phrasing is a bit stiff, but clearly local.

3 (Hybrid) Mixes Dialect and MSA. Phrasing feels awkward or "translated."

2 (MSA) Correct Arabic, but it is Formal (MSA), not Dialect.

1 (Fail) Gibberish, wrong dialect entirely, or not Arabic.

B.1.4 Protocol for MSA Leakage

To isolate meaning from register control, annotators were instructed to score semantic adequacy and dialectness independently.

> *Example:* A correct MSA translation for a request in Moroccan Arabic receives a **Semantic Score of 5** (perfect meaning) but a **Dialect Score of 1–2** (wrong register). XSTS scores are not penalized for dialect errors.

C Results

C.1 Automatic Evaluation Results

C.1.1 Impact of Linguistic Distance and Code-Switching

To better understand the variance in translation quality across different regions, we investigated

Turn-Level Prompt Configuration

```
Translate the English text contained in the JSON input into <DIALECT>.
Input: { "country": "<Country>", "domain": "<Domain>", "participants": ["<Speaker1>",
"<Speaker2>"], "gender_direction": "<Gender>", "speaker": "<Current_Speaker>", "text":
"<Source_Text>" }
Guidelines:
- Return the result strictly in valid JSON.
- Translate to <DIALECT> using Arabic script.
- Do not add any code, explanations, comments, or any other extra text.
- Keep the meaning and tone and respect the gender direction.
- Consider the country, the domain, the participants, and the speaker in your translation.
Output scheme: { "translation": "translated text here" }
```

Context-Level Prompt Configuration

```
Translate the given turn of a conversation from English to <DIALECT>, considering the previous
context if provided.
Input: { "country": "<Country>", "domain": "<Domain>", "participants": [...], "context":
["<Previous_Turn_1>", "<Previous_Turn_2>"], "current_turn": { "speaker": "...", "text": "..."
} }
Guidelines:
... [Same as Turn-Level] ...
- Only translate the "text" field of the "current_turn".
- If a context is provided, do not translate it, and use it to inform your translation.
Output scheme: { "translation": "translation of the text from the current turn" }
```

Conversation-Level Prompt Configuration

```
Translate all turns in the following conversation from English to <DIALECT>.
Input: { "country": "<Country>", "domain": "<Domain>", "participants": [...], "turns": [
{"speaker": "A", "text": "..."}, {"speaker": "B", "text": "..."} ] }
Guidelines:
... [Same as Turn-Level] ...
Output scheme: { "turn_1": "translation of the text from turn_1", "turn_2": "translation of the
text from turn_2", ... }
```

Figure B.2: The three prompt configurations used for the English → Arabic Dialect evaluation. Note that for the reverse direction (Dialect → English), the source/target languages are swapped, and the guideline regarding **Arabic script** is removed.

Country	Lexical Overlap	Correlation with spBLEU
EG	0.17	0.3970
JO	0.21	0.4335
LB	0.15	0.3967
MA	0.10	0.2731
PS	0.19	0.3943
SA	0.23	0.4824
SY	0.22	0.1911
TN	0.14	0.3606
YE	0.19	0.4414

Table C.1: Lexical overlap between dialectal references and NLLB-generated MSA translations, alongside their correlation with spBLEU scores.

Country	Avg spBLEU (No Latin)	Avg spBLEU (Has Latin)	CMI Correlation
EG	27.86	21.86	-0.13
JO	24.17	18.85	-0.06
LB	22.30	17.13	-0.14
MA	20.81	16.33	-0.14
PS	23.75	17.56	-0.05
SA	27.27	28.28	0.02
SY	26.51	29.89	0.02
TN	24.41	19.39	-0.21
YE	19.78	21.78	0.01

Table C.2: Impact of code-switching on translation quality, comparing average spBLEU scores for sentences with and without Latin characters, along with the Code-Mixing Index (CMI) correlation.

two primary linguistic factors: a dialect's distance from MSA and the prevalence of intra-sentential code-switching. First, we measured the lexical overlap between our dialectal reference texts and MSA translations generated by NLLB. As detailed in Table C.1, we observed a moderate positive correlation between MSA lexical overlap and spBLEU scores (from English-to-Dialect using c4ai-command-a-03-2025 experiments) across the dataset (e.g., reaching 0.48 for Saudi and 0.44 for Yemeni). This indicates that models consistently yield higher quality translations for sentences—and by extension, dialects like Levantine and Gulf—that share lexical and structural similarities with MSA. Conversely, structurally distant varieties, such as Maghrebi dialects, pose a significantly greater challenge.

Second, we analyzed the impact of code-switching, which speakers frequently employ using Latin script (e.g., English or French loanwords) to bridge lexical gaps in technical domains. By comparing sentences with and without Latin characters (Table C.2), we found that code-mixing degrades translation performance. For the majority of the evaluated dialects (including Egyptian, Jordanian, Lebanese, Moroccan, Palestinian, and Tunisian), average spBLEU scores dropped substantially in the presence of code-switching, a finding further reinforced by a negative correlation with the Code-Mixing Index.

C.1.2 Impact of LLM Reasoning on Translation Performance

Figure C.1 presents a comparison between three models using two configurations: one with the thinking process and one without. The results show that the thinking process generally does not help and often hurts translation performance, except for *gemini-3-flash*. In this case, reasoning boosts average performance by 2.0 spBLEU points for English-to-Dialect and approximately 0.4 points for Dialect-to-English.

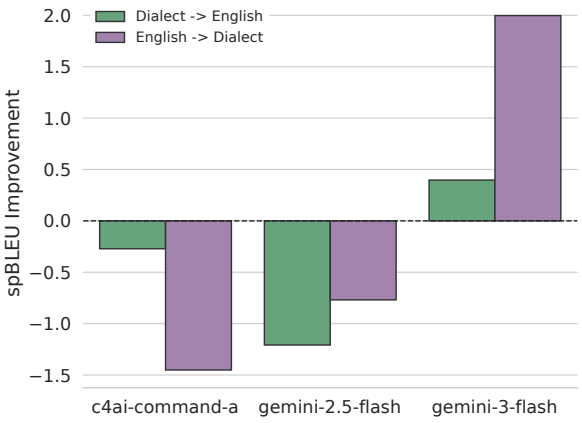

Figure C.1: Impact of reasoning on translation performance. The bar chart shows the spBLEU improvement (or degradation) when reasoning is enabled for the evaluated models compared to their non-reasoning baselines.

C.2 Human Evaluation Results

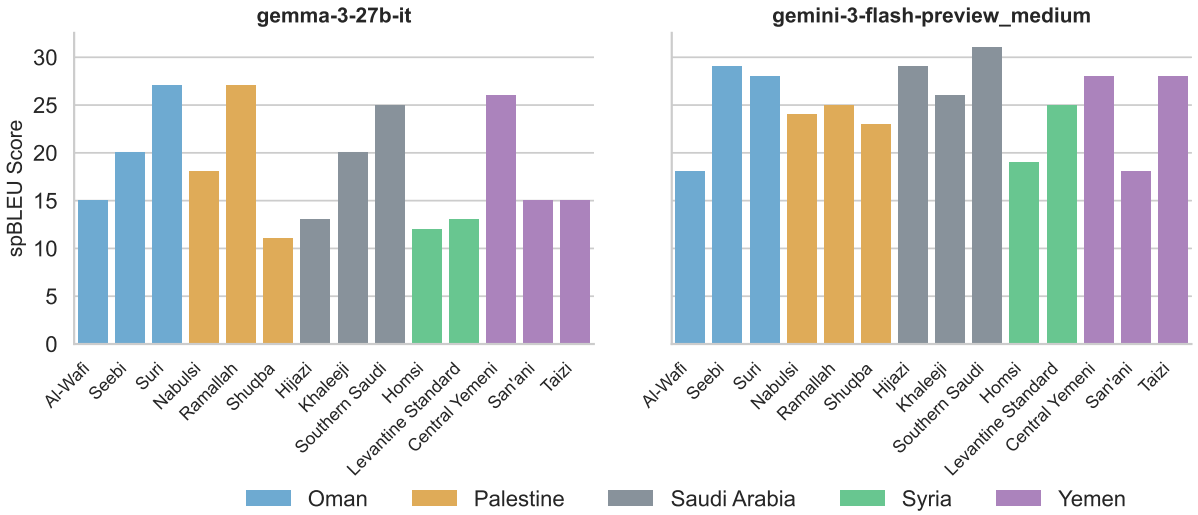

Figure C.2: spBLEU scores for selected LLMs on the Alexandria test set (English → Sub-Dialect). We report results for specific sub-dialects across five countries to highlight intra-country performance discrepancies.

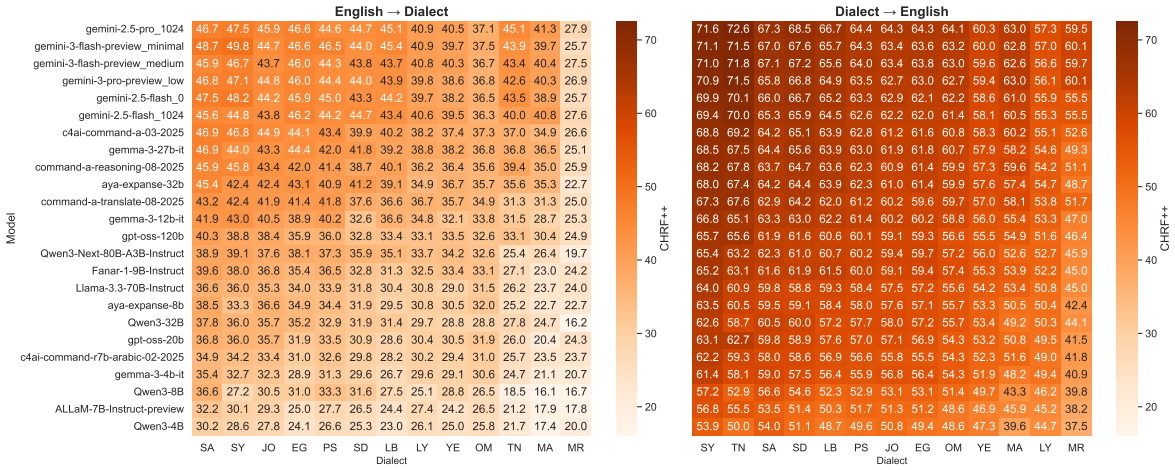

Figure C.3: chrF++ scores for LLM-based machine translation on the Alexandria test set. Results cover 13 dialects in both directions (English → Dialect and Dialect → English).

Model	EG	JO	LB	LY	MA	MR	OM	PS	SA	SD	SY	TN	YE
gemini-3-flash-preview_medium	97.5	100.0	100.0	99.4	100.0	99.4	98.9	100.0	99.6	100.0	95.2	98.8	100.0
gemini-3-flash-preview_minimal	95.1	100.0	99.6	99.4	100.0	99.4	96.1	99.6	99.6	100.0	97.6	95.1	100.0
c4ai-command-a-03-2025	96.3	97.7	98.4	100.0	100.0	99.4	98.9	100.0	99.6	98.8	95.8	95.1	96.2
gpt-oss-120b	93.8	100.0	98.8	96.4	98.2	99.4	98.4	99.6	99.2	98.8	91.7	92.6	100.0
gemma-3-27b-it	91.4	97.7	96.4	97.6	100.0	97.1	98.9	99.2	98.4	95.2	91.1	96.3	100.0
aya-expanse-32b	92.6	100.0	96.0	96.4	98.8	97.1	97.7	98.8	98.4	98.8	86.3	96.3	98.8

Table C.3: Human Evaluation Gender Accuracy (Pass %) across different countries

Model	EG	JO	LB	LY	MA	MR	OM	PS	SA	SD	SY	TN	YE
gemini-3-flash-preview_medium	3.94	5.00	4.78	4.03	4.73	3.13	4.85	4.87	4.98	4.64	4.07	3.21	4.74
gemini-3-flash-preview_minimal	3.86	5.00	4.78	4.01	4.65	3.13	4.90	4.89	4.96	4.71	4.06	3.16	4.74
c4ai-command-a-03-2025	3.72	4.98	4.44	3.73	4.39	3.16	4.88	4.62	4.93	4.48	4.08	2.94	4.47
gpt-oss-120b	3.58	5.00	4.46	3.62	4.36	3.10	4.89	4.62	4.87	4.69	4.00	2.78	4.35
gemma-3-27b-it	3.62	5.00	4.35	3.54	3.59	3.25	4.79	4.50	4.86	4.16	4.03	2.60	4.72
aya-expanse-32b	3.51	5.00	4.50	3.50	4.09	3.64	4.75	4.38	4.89	4.76	3.91	2.72	4.60

Table C.4: Human Evaluation Semantic Adequacy (1-5) across different countries

Model	EG	JO	LB	LY	MA	MR	OM	PS	SA	SD	SY	TN	YE
gemini-3-flash-preview_medium	3.56	4.87	4.57	4.05	4.35	3.01	4.55	4.56	4.63	4.77	3.92	3.24	4.49
gemini-3-flash-preview_minimal	3.62	4.83	4.54	3.95	4.19	3.04	4.41	4.56	4.55	4.61	3.82	3.20	4.28
c4ai-command-a-03-2025	3.31	3.95	4.14	3.30	3.78	2.27	4.22	4.12	4.36	4.23	3.80	2.85	3.66
gpt-oss-120b	3.22	3.41	3.85	2.83	3.30	2.13	4.36	3.73	4.15	3.28	3.61	2.58	3.67
gemma-3-27b-it	3.35	3.90	3.85	3.07	3.04	2.08	4.17	3.86	4.26	3.69	3.66	2.30	3.86
aya-expanse-32b	3.02	3.15	3.37	2.78	2.76	2.04	3.71	3.48	3.26	2.36	3.51	2.22	2.25

Table C.5: Human Evaluation Dialectness & Fluency (1-5) across different countries

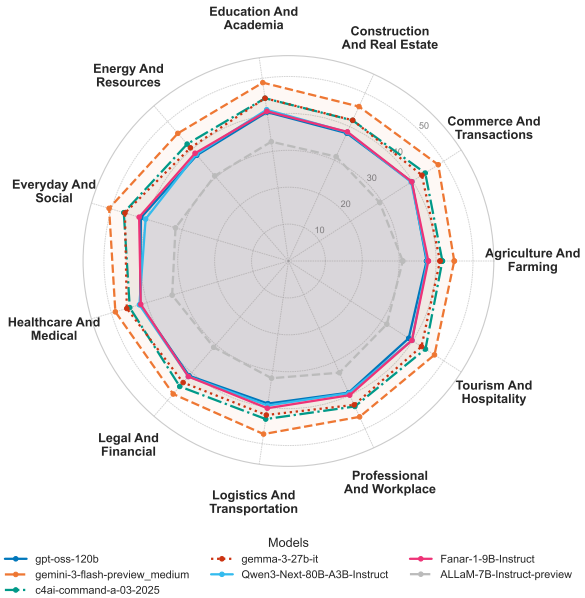

Figure C.4: Domain robustness analysis (Dialect → English). The radar chart illustrates spBLEU scores for a subset of models across all 11 domains, demonstrating consistent performance stratification regardless of the topic.

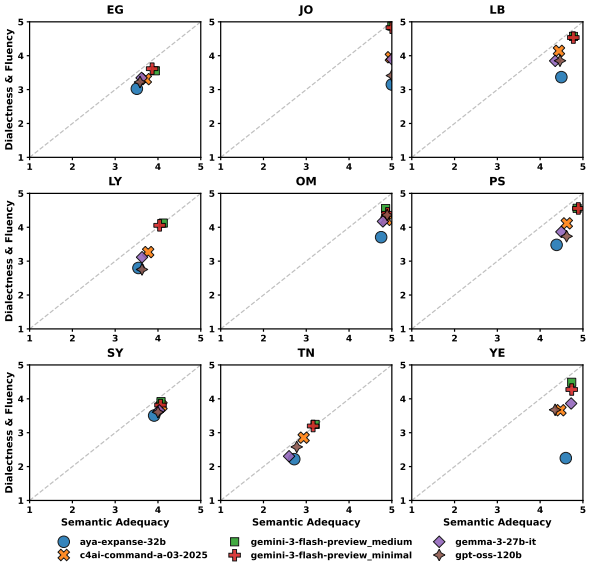

Figure C.5: Human evaluation of Semantic Adequacy vs. Dialectness across the remaining dialects other than the ones presented in the main text. Points below the diagonal ($y = x$) indicate that models consistently achieve higher semantic fidelity than dialectal authenticity.