Current state of LLMs for Arabic dialectal machine translation

Josef Jon[1] Rawan Bondok[2] Ondřej Bojar[1]
[1]Charles University, Faculty of Mathematics and Physics, Prague, Czechia
jon@ufal.mff.cuni.cz
[2]TU Darmstadt, Germany
rawanessam34@gmail.com

Abstract

This work presents an evaluation of large language models (LLMs) for English to dialectal Arabic machine translation on the MADAR dataset. We evaluate both translation directions (English to Arabic and vice-versa) on 16 Arabic dialects. Our experiments cover a diverse set of models, including specialized Arabic models (Jais, Nile), multilingual models (Gemma, Command-R, Mistral, Aya), and commercial APIs (GPT-4.1). We employ multiple evaluation metrics: BLEU, CHRF, COMET (both reference-based and reference-less variants) and GEMBA (LLM-as-a-judge), as well as a small-scale manual evaluation, to assess translation quality. We discuss the challenges of automatic MT evaluation, especially in the context of Arabic dialects. We also evaluate the ability of LLMs to classify the dialect used in a text. The study offers insights into the capabilities and limitations of current LLMs for dialectal Arabic machine translation, particularly highlighting the difficulty of handling dialectal diversity, although the results may be influenced by possible training data contamination, which is always a concern with LLMs.

1 Introduction

Machine translation (MT) for dialects of Arabic remains a challenging problem despite significant progress in neural and large language model (LLM)–based approaches. Arabic is characterized by a high degree of linguistic complexity, including rich morphology, flexible word order, and diglossia between Modern Standard Arabic (MSA) and numerous regional dialects. While MSA is widely used in formal writing and media, everyday communication across the Arab world primarily relies on dialectal varieties, which differ substantially both from MSA and from each other. Moreover, even in a dialect spoken in a single country, there can be large regional variations.

In the last few years, we have witnessed rapid advances in multilingual LLMs, which have demonstrated strong performance on a wide range of natural language processing tasks, including MT (Kocmi et al., 2025b). However, their capabilities for dialectal Arabic are still insufficiently understood. Most existing evaluations (Saeed, 2025) focus on MSA, leaving open questions about how well current LLMs handle the full spectrum of Arabic dialectal variation.

Dialectal Arabic MT poses unique challenges beyond data scarcity. Dialects show substantial lexical, phonological, and syntactic variation and lack of standardized orthography. Even within a single country, there can be considerable variations depending on region, speaker, and domain. These factors complicate both model training and evaluation. In particular, automatic evaluation metrics that rely on a single reference translation may fail to capture valid alternative renderings, leading to unreliable quality estimates. While this is a general problem of reference-based metrics, it is especially pronounced for dialectal Arabic, where multiple translations may be equally acceptable but differ significantly at the surface form.

We present a comprehensive evaluation of 16 large language models for Arabic↔English machine translation across 16 Arabic dialects using the MADAR test set (Bouamor et al., 2018). We evaluate both translation directions and cover a diverse set of models, including Arabic-specialized LLMs, strong multilingual models, and commercial API-based systems. We employ BLEU, ChrF, COMET and LLM-as-a-judge metrics, as well as a small-scale manual evaluation. Through quantitative and qualitative analysis, we investigate model performance patterns across dialects, translation directions, and evaluation metrics, shedding light on the current capabilities and limitations of LLMs for Arabic MT. Our findings show both the progress made by recent LLMs and the persistent challenges

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 329–363

posed by dialectal diversity. By providing a systematic comparison, this study aims to serve as a basis for selecting LLMs for MT in various scenarios, as a reference point for future research on Arabic machine translation and as an encouragement for development of more robust models and evaluation methodologies tailored to dialectal Arabic. We note that due to the unavailability of the training data for many of the models, we do not know if the MADAR test set was used during the training of the LLMs. This is a major shortcoming of this work, which must be taken into account when basing any decisions on our results.

The main contributions of this work are:

- A comprehensive evaluation of 16 LLMs on Arabic-English translation across 16 dialects
- Manual analysis of selected examples
- Analysis of automatic evaluation metrics for dialectal Arabic MT
- Insights into dialect-specific challenges and model performance patterns
- Public release of results and analysis scripts[1]

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 describes the MADAR dataset and the evaluated models. Section 4 presents experimental settings and evaluation metrics. Section 5 presents and analyzes results. Section 6 discusses the implications of the results. Section 7 summarizes our findings.

2 Related Work

The work on Arabic MT copies the general trends in the field of MT. Early work relied on rule-based and phrase-based statistical MT systems, often combined with morphological analysis (Habash, 2010; Habash and Hu, 2009).

In recent years, neural machine translation (NMT), particularly Transformer-based architectures (Vaswani et al., 2017), has become the dominant paradigm for all MT, including Arabic (Almahairi et al., 2016; Durrani et al., 2016). The newest iteration of neural network-based approaches, LLMs, have also been finetuned with success for Arabic translation (Nagoudi et al., 2022a,b). A number of recent LLMs with Arabic capabilities are listed in Table 1.

Other works have addressed the specific issue of dialectal Arabic MT, either through pivoting via MSA (Sajjad et al., 2013; Salloum and Habash, 2013), or by finetuning the models (Zbib et al., 2012; Farhan et al., 2020; Nagoudi et al., 2021; Baniata et al., 2018; Heakl et al., 2024). The translation of Arabic dialects remains a low-resource problem due to limited parallel data and high linguistic variation. To address this, several datasets have been introduced, including MADAR, which covers dialects from multiple countries aligned with MSA and English (Bouamor et al., 2018).

One of the regular evaluation campaigns that covers Arabic dialects (Tunisian and Levantine) is IWSLT (Ahmad et al., 2024; Abdulmumin et al., 2025). LLMs have proven to be effective in this setting as well (Ben Kheder et al., 2024).

Model	Reference
Arabic-specialized models	
Jais-2-8B-Chat	Anwar et al. (2025)
Jais-2-70B-Chat	Anwar et al. (2025)
Nile-Chat-12B	Shang et al. (2025)
c4ai-command-r7b-arabic	Alnumay et al. (2025)
Multilingual models	
Aya Expanse 8B	Dang et al. (2024)
Aya Expanse 32B	Dang et al. (2024)
c4ai-command-r-08-2024	Cohere Labs (2024a)
c4ai-command-r-v01	Cohere Labs (2024b)
Command-A-Translate-08-2025	Kocmi et al. (2025a)
Gemma-3-4B-IT	Team (2025)
Gemma-3-27B-IT	Team (2025)
EuroLLM-9B-Instruct	Martins et al. (2025)
Mistral-Small-3.2-24B-Instruct	Mistral-Team (2025)
Qwen3-4B-Instruct-2507	Yang et al. (2025)
Llama-3.3-70B-Instruct	Grattafiori et al. (2024)
Commercial API models	
GPT-4.1-nano	OpenAI et al. (2024)
GPT-4.1-mini	OpenAI et al. (2024)

Table 1: Overview of evaluated large language models.

3 Data and Models

3.1 Dataset Overview

MADAR is a parallel corpus designed for evaluating machine translation systems across Arabic dialects. The dataset includes 32 test sets covering multiple domains, spanning English and the following dialects: Modern Standard Arabic (MSA), Moroccan (MA), Tunisian (TN), Algerian (DZ), Libyan (LY), Syrian (SY), Jordanian (JO), Palestinian (PS), Lebanese (LB), Qatari (QA), Omani (OM), Saudi (SA), Yemeni (YE), Iraqi (IQ), Egyptian (EG) and Sudanese (SD). We use the version from AraBench repository (Sajjad et al., 2020).

3.2 Evaluated Models

We evaluate 16 LLMs, shown in Table 1.

[1] https://github.com/cepin19/arabic_llms

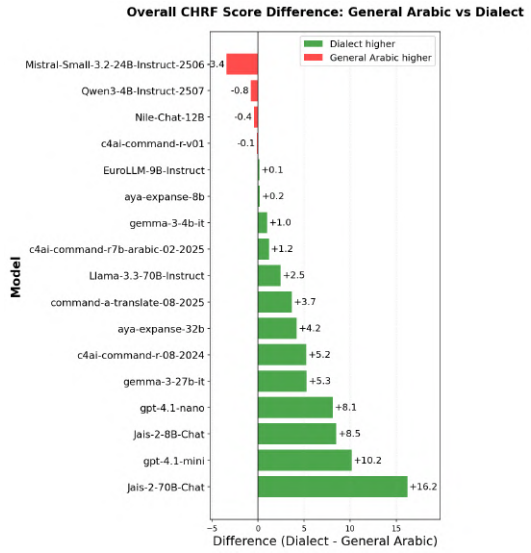

Figure 1: ChrF difference between using dialect-specific vs. general prompt on complete concatenated test set. Green signifies the dialect-specific prompt scored better.

4 Experimental Settings

4.1 Metrics

We compute **BLEU** (Papineni et al., 2002) and **ChrF** (Popović, 2015) scores using SacreBLEU (Post, 2018). In order to compute **COMET** (Rei et al., 2020) scores we use the original implementation[2] and the `XCOMET-XL` model. **GEMBA** (Kocmi and Federmann, 2023b,a) is an LLM-as-a-judge approach, where an LLM is prompted to assess translation quality. We use direct assessment without reference, with a score scale from 0 to 100. We instruct the evaluation model to subtract up to 50 points for the wrong dialect used in the translation (see Appendix B for the exact prompt). For **human evaluation**, we use direct assessment as well, with two separate 0-100 scores for accuracy and dialectalness.

4.2 Translation Directions

We translate the test sets in both directions (English to/from Arabic). For the translation into Arabic, we use **two different prompts: general Arabic** and **dialect-specific**. In the dialect-specific, we instruct the model to translate into the dialect specified by the country of the origin of the translation, e.g. *Translate into Egyptian Arabic: text*. In the general Arabic prompt, we only instruct the model to translate into Arabic. The exact prompt is shown in Appendix B. For the Arabic to English direction, we only use one prompt, as we do not specify the source language or dialect. We compute the scores both on particular files as well as merged datasets, where all the test sets for the given dialect are concatenated.

4.3 Dialect classification

We use some of the models (*Jais-2-70B-Chat*, *gemma-3-27b-it*, *aya-expanse-32b* and *Llama-3.3-70B-Instruct*) to classify the dialect of both the produced machine translations and the reference human translations. While GPT models showed a promising performance across tasks, we omit them from these experiments due to budget constraints. We aim to assess if these LLMs can identify the dialects (by classifying the reference translations), and, provided they can, we use them to evaluate dialectalness of the produced machine translations. The classification prompt is shown in Appendix B.

4.4 Inference Setup

The OpenAI models are accessed through OpenAI API with default parameters. We deploy other models using VLLM with default decoding parameters.

4.5 Postprocessing

We have noticed that some LLMs produce repetitions of a single token for some of the inputs. In case the produced output is more than 5 times longer than the source, either in tokens or in characters, we crop the output to the input token length.

We use CAMeL Tools (Obeid et al., 2020) to normalize both the references and the translations. We use Alef Maksura, Alef, Teh Marbuta and Hamza normalizations and we also convert Arabic numerals to Latin ones.

5 Results and Analysis

In Figure 1, we compare system-level (i.e., on concatenation of all test sets) ChrF scores for general vs. dialect-specific prompts across models. The figure demonstrates which models react to dialect-specific prompts and to what extent. Many of the models are able to distinguish between dialects and produce more appropriate translations based on the dialect specification, resulting in higher overall ChrF scores when the target dialect is specified.

In Figure 2 we present the English to Arabic dialect-level (i.e. on concatenation of test sets pertaining to one dialect) BLEU scores, for the dialect-specific prompt for all models. ChrF scores for the

[2] https://github.com/Unbabel/COMET

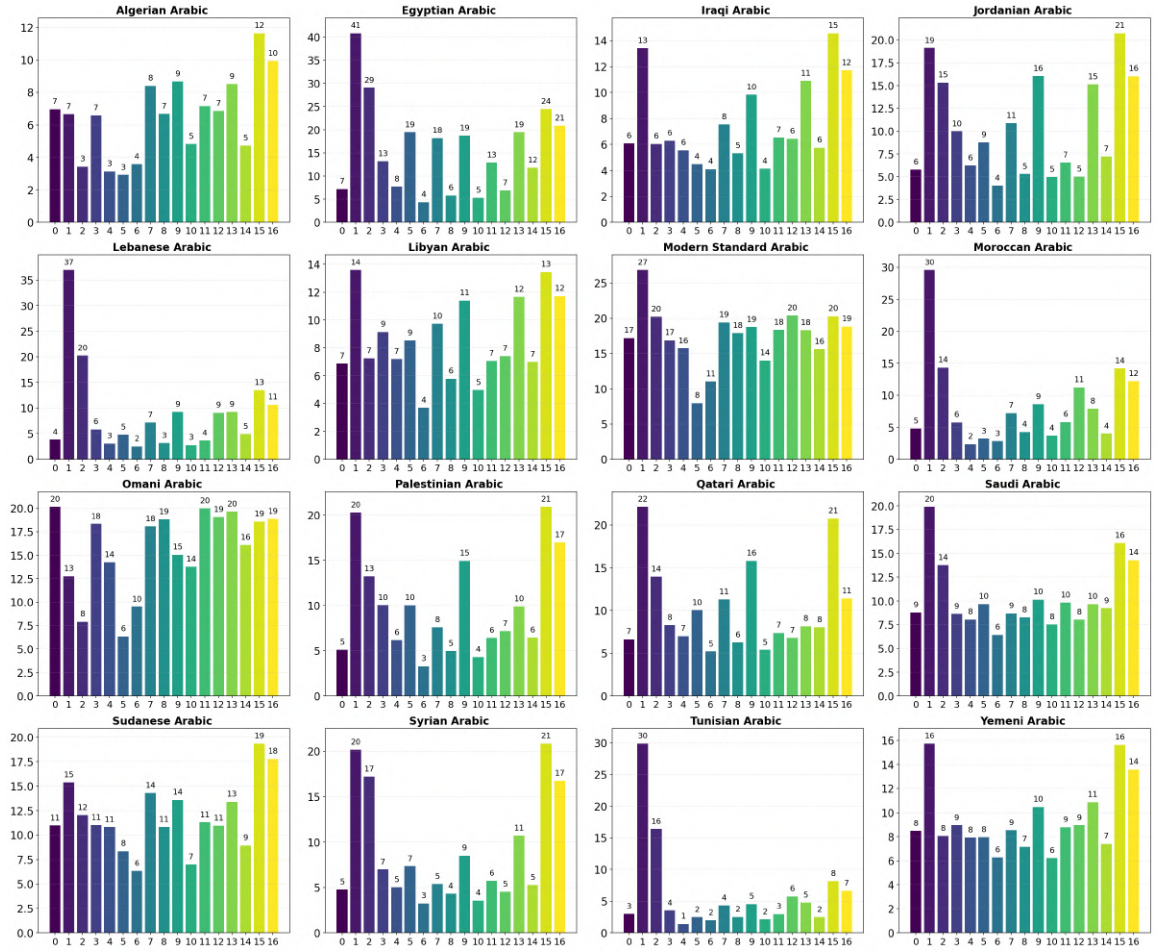

Figure 2: BLEU scores for merged dialect test sets for all models in English to Arabic, using the dialect-specific prompt. The indices on the *x* axis map to the models as follows: 0: EuroLLM-9B-Instruct, 1: Jais-2-70B-Chat, 2: Jais-2-8B-Chat, 3: Llama-3.3-70B-Instruct, 4: Mistral-Small-3.2-24B-Instruct-2506, 5: Nile-Chat-12B, 6: Qwen3-4B-Instruct-2507, 7: aya-expanse-32b, 8: aya-expanse-8b, 9: c4ai-command-r-08-2024, 10: c4ai-command-r-v01, 11: c4ai-command-r7b-arabic-02-2025, 12: command-a-translate-08-2025, 13: gemma-3-27b-it, 14: gemma-3-4b-it, 15: gpt-4.1-mini, 16: gpt-4.1-nano

same configuration are located in Appendix A, Figure 8. We see that for many models, BLEU scores under 10, which usually signal unusable translation, are common. Notable exceptions are the Jais-2 and gpt-4.1 models, which result in at least moderate scores for most of the dialects, except the Algerian, Iraqi, Yemeni and Libyan dialects. Even for these models however, many of the dialect BLEU scores are not higher than 20, indicating that the translations are not very similar to the reference. To gauge whether the low scores are indicative of poor translation quality, or are more caused by the orthographic and morphological richness and diversity of the dialects, we carry out a small manual evaluation for the Egyptian dialect, described later.

We show the ChrF scores on individual test sets in Figure 4. We see that `gpt-4.1-mini` and `Jais-2-70B-Chat` models have the highest scores across all test sets, with a very few exceptions. This is also demonstrated by the ChrF-based ranking of

Model	Size	Wins	Avg. Rank
gpt-4.1-mini	N/A	16	2.1
Jais-2-70B-Chat	70B	14	2.9
gpt-4.1-nano	N/A	0	3.5
gemma-3-27b-it	27B	0	5.3
c4ai-command-r-08-2024	32B	0	6.0
aya-expanse-32b	32B	0	6.7
command-a-translate-08-2025	111B	2	7.1
Jais-2-8B-Chat	8B	0	7.8
Llama-3.3-70B-Instruct	70B	0	9.6
c4ai-command-r7b-arabic-02-2025	7B	0	9.6
Nile-Chat-12B	12B	0	10.8
EuroLLM-9B-Instruct	9B	0	11.1
aya-expanse-8b	8B	0	11.7
gemma-3-4b-it	4B	0	12.7
Mistral-Small-3.2-24B-Instruct-2506	24B	0	14.8
c4ai-command-r-v01	35B	0	15.0
Qwen3-4B-Instruct-2507	4B	0	16.4

Table 2: Average rankings of all models based on ChrF scores using dialect-specific prompts.

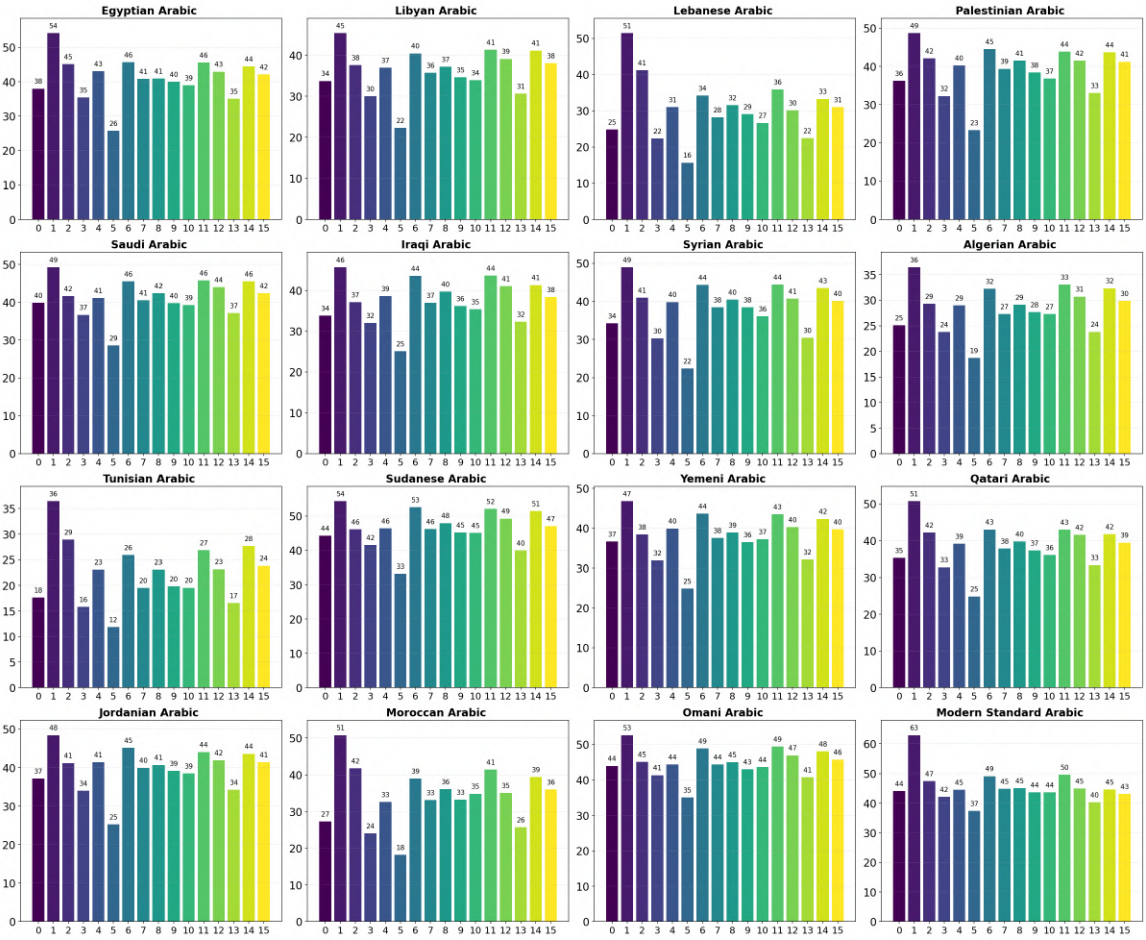

Figure 3: BLEU scores for merged dialect test sets for all models in Arabic to English direction. See Figure 8 for the mapping of the indices on the *x* axis to model names.

each model, computed across all test sets, which is shown in Table 2. Across all models, we see high scores on the MSA test set, showing that most of the models possess good capabilities for translation into MSA, but only some of them are capable of producing dialectal Arabic translations.

The BLEU scores for the reverse direction, Arabic-to-English, are presented in Figure 3. The differences between models are less pronounced here and the overall scores are higher as most models can produce high-quality English translations for a dialectal Arabic source text. `Jais-2-70B-Chat` scores the best on all test sets.

Similar plots for the COMET-XL scores are shown in Appendix A, in Figures 13 and 14 for the reference-based and reference-less versions. These results demonstrate that COMET is not suitable for our use-case, which is caused by the fact that these models do not allow the user to specify source and target languages. This could be partially mitigated by using the reference-based version, but even then, the model largely prefers the MSA translations over the dialectal ones. We make this observation as an argument for including language pair specification into future automatic MT metrics.

Dialect	BLEU		CHRF	
	General	Dialect	General	Dialect
Algerian	7.0	11.6	32.8	40.1
Egyptian	6.0	23.6	32.5	52.2
Iraqi	5.9	14.6	35.8	44.8
Jordanian	5.7	20.7	34.8	50.0
Lebanese	3.3	14.0	30.5	44.4
Libyan	6.4	13.4	33.3	43.1
MSA	16.6	16.4	48.5	48.2
Moroccan	3.2	10.6	27.7	41.2
Omani	21.0	18.6	50.7	46.9
Palestinian	5.4	20.9	33.1	50.5
Qatari	6.6	14.0	35.8	43.9
Saudi	9.5	16.1	40.3	47.3
Sudanese	11.1	19.3	41.0	48.9
Syrian	4.8	20.8	34.3	51.1
Tunisian	2.6	7.6	25.3	36.1
Yemeni	8.5	15.6	37.4	46.3

Table 3: Average merged dialect scores for GPT-4.1-mini: General vs. dialect-specific prompt

Other results are located in Appendix A.

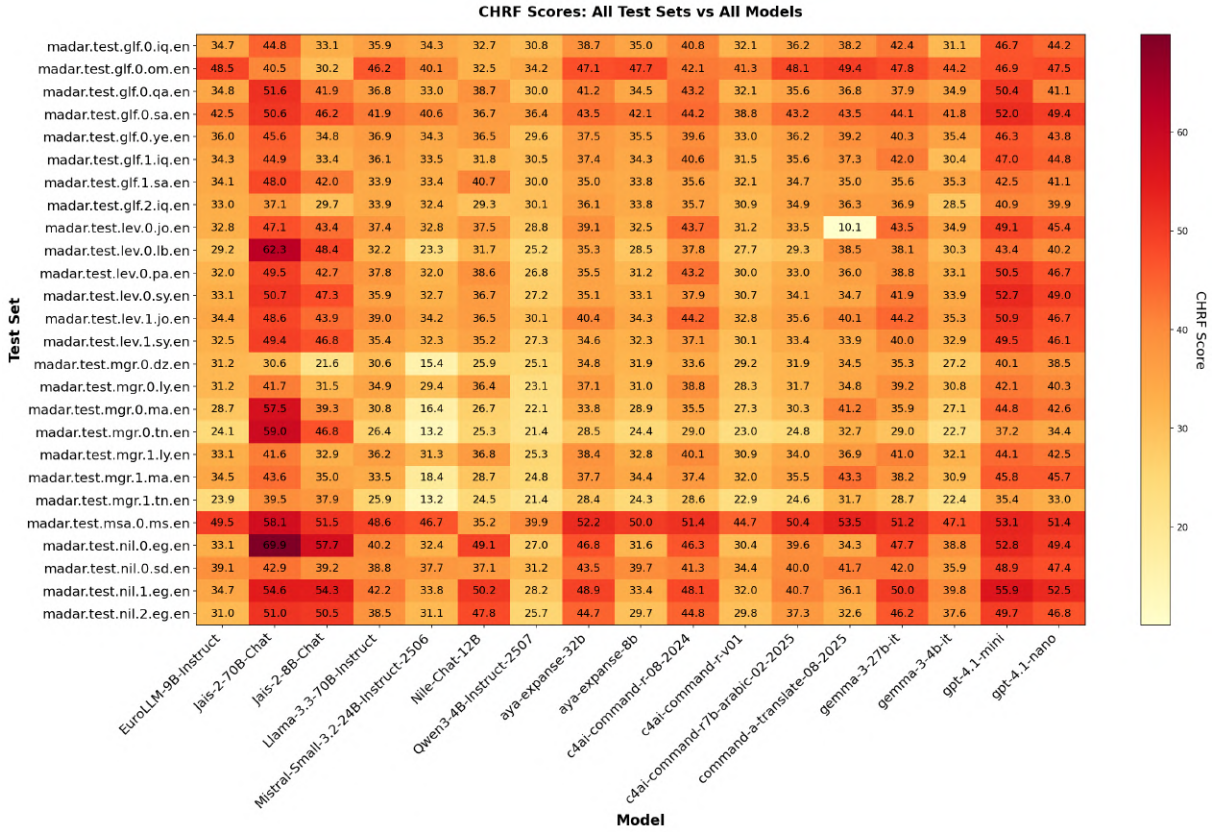

Figure 4: ChrF scores for individual test sets for all models in English to Arabic direction. Translation produced using dialect-specific prompt.

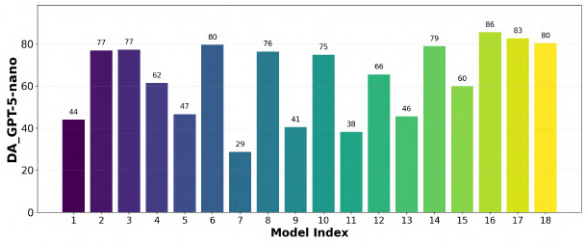

Figure 5: GEMBA GPT-5-nano DA evaluation scores for all models on madar.test.nil.1.eg, using the dialect-specific prompt. See Figure 2 for the mapping of the indices on the x axis to model names.

5.1 GEMBA

We also used an LLM-as-a-judge approach (GEMBA) to score the translations. We simulated a direct assessment setting, with a 0-100 scale, up 50 points being awarded for the translation accuracy and 50 points being reserved for the dialect appropriateness. We used `gpt-5-nano` as the judge. Due to budget constraints, we have only run the scoring on the same test set we have performed our human evaluation on, `madar.test.nil.1.eg`. The results are shown in Figure 5. There are multiple well-performing models, and the results among the top models are more even than according to ChrF and BLEU scores.

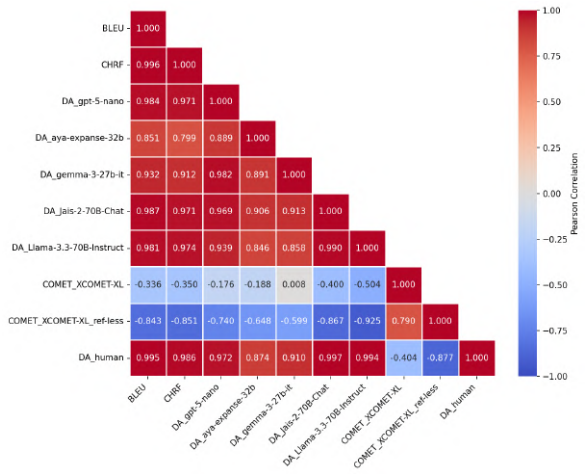

Figure 6: Pearson's correlation of system-level automatic metrics and human DA on the first 100 sentences of `madar.test.nil.1.eg` for translation created with the dialect-specific prompt

5.2 Manual analysis

The manual error analysis followed a two-dimensional annotation scheme consisting of translation **accuracy** and **dialectness**, which were assessed independently for each generated output.

Accuracy annotations capture semantic adequacy errors, indicating whether the meaning of the English source sentence is correctly conveyed

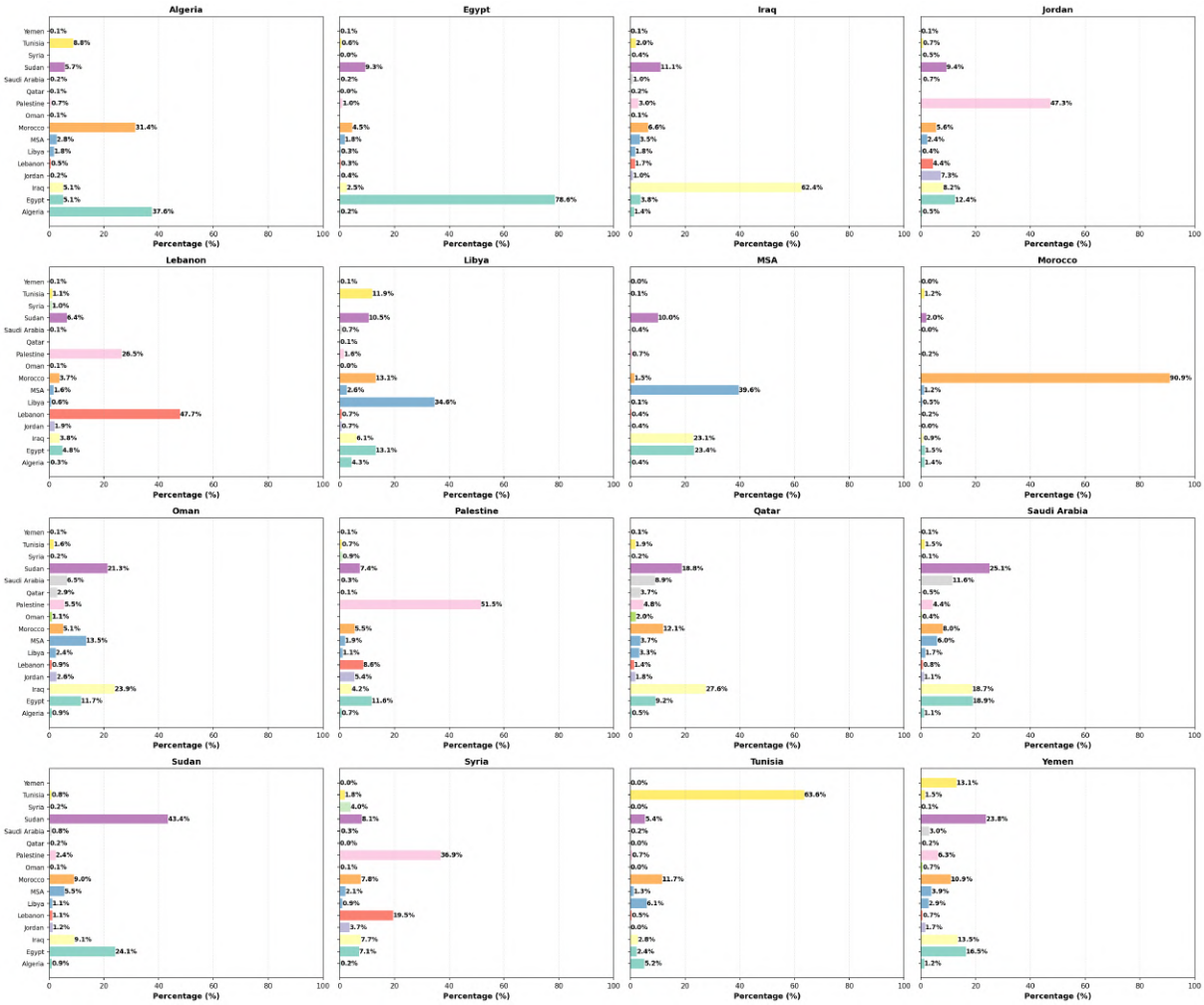

Figure 7: Distribution of dialects in reference translations. The dialects are classified on the sentence level, using `Jais-2-70B-Chat`.

in the translation. Errors under this dimension include omissions, additions, mistranslations, and incorrect semantic relations.

Dialectness annotations target the dialectal realization, reflecting the extent to which the generated translation conforms to the lexical, morphological, and syntactic conventions of the target dialect. Errors in this category include the use of MSA or other non-target dialect forms, unnatural code-switching, and dialect-inappropriate constructions.

Crucially, the two annotation dimensions are orthogonal: a translation may be semantically accurate while failing to realize the target dialect, or conversely, may exhibit strong dialectal features despite conveying an incorrect or incomplete meaning. We performed a small-scale manual analysis on a sample of 100 sentences from `madar.test.nil.1.eg` test set. `Jais-2-70B-Chat` and `gpt-4.1-mini` were selected as the two best-performing systems according to automatic evaluation scores, and `EuroLLM-9B-Instruct` and `C4ai-command-r-08-2024`, which achieved comparatively lower automatic scores. Including both high- and lower-performing models allowed us to examine qualitative differences in error patterns and dialectal realization across performance tiers. Figure 6 shows Pearson's system-level correlation between automatic metrics and human assessment (the mean of dialectalness and accuracy DA scores). The scores suggest the suitability of BLEU, ChrF and LLM-produced DA scoring, which has the advantage of not needing a reference translation. It also further demonstrantes the unsuitability of COMET scores for our use-case, already discussed earlier. However, due to the small-scale nature of our human evaluation, further data is needed to confirm these observations.

5.3 Dialect classification

We employed LLMs to identify the dialect of a text. First, we classified the reference translations to ver-

Model	J70B	G-mini	ELLM-9B	CR-824
Accuracy	94.6	96.3	93.0	93.4
Dialect	98.5	96.0	36.4	86.8
Mean Score	96.5	96.2	64.7	90.1

Table 4: Manual evaluation scores for accuracy and dialectalness on the translation of the first 100 sentences of `madar.test.nil.1.eg` produced with the dialect-specific prompt by the following models: Jais-2-70B-Chat (J70B), GPT-4.1-mini (G-mini), EuroLLM-9B-Instruct (ELLM-9B), Command-r-08-2024 (CR-824).

ify the ability of the models to correctly assess the dialect used in a text. We assume that in a test set for a given dialect, a large part of sentences posses the features of that dialect. We show the classification results produced by the `Jais-2-70B-Chat` model for the reference human translations in Figure 7. Plots for translations produced by the LLMs are shown in Appendix A.5. We observe that in general, the dialectical human translation exhibit higher percentage of sentences classified as belonging to the given dialect, or a related one (e.g. Egypt and Sudan in the Sudan test set).

6 Discussion

When considered jointly, automatic and manual evaluations reveal complementary strengths and limitations of current LLMs for English-to-Dialectal Arabic MT. Models trained with substantial dialectal Arabic data, such as Jais-2-70B-Chat, outperform general-purpose systems, particularly in producing outputs with high dialectness. By contrast, models optimized for broad multilingual coverage, such as EuroLLM-9B-Instruct, tend to generate more standardized or mixed varieties.

Although automatic metrics like BLEU and ChrF correlate well with human judgments overall on our small sample, manual evaluation provides finer-grained insights into dialectal realization.

Human evaluation further reveals a distinct pattern in the `c4ai-command-r-08-2024` model's handling of gender-ambiguous inputs: unlike other models, which default to masculine morphology, it frequently resolves ambiguity using feminine forms. This behavior suggests an implicit mitigation strategy that is not captured by automatic metrics and raises important questions about bias and sociolinguistic alignment in dialectal MT.

7 Conclusion

We presented a large-scale evaluation of 16 LLMs on English to Arabic MT across 16 Arabic dialects, combining BLEU/ChrF/COMET, LLM-as-a-judge and a human assessment study. Across metrics and test sets, performance is highly uneven: most models translate well into MSA, but many struggle to reliably produce the requested dialect, with particularly low scores common for several dialects.

Arabic-specialized and strong commercial models are the most consistent for dialectal generation. Dialect-specific prompting generally improves dialectal outputs substantially, confirming that some LLMs can condition on dialect instructions. GPT-4.1-mini and Jais-2-70B dominate system-level rankings, while several other multilingual models frequently default to MSA, even when prompted for a dialectal translation.

Our analyses also show limitations of current automatic MT evaluation for dialectal Arabic. Reference-based metrics can substantially undervalue valid dialectal variants due to non-standard orthography and high surface-form variability; this is supported by our manual evaluation. COMET score proved poorly suited to this setting because it cannot be constrained to penalize "wrong-dialect" outputs and tends to favor MSA-like texts.

Our comparison is intended to support researchers and practitioners in selecting suitable models for deployment in machine translation applications. We acknowledge that the test set may have been included in the models' training data, and therefore, the high scores should be interpreted with caution. Nevertheless, low scores are still informative, as they indicate models that are unable to generate the target dialects, and this insight remains practically useful.

Limitations

The main limitation of our work is the fact that for many of the models, we do not have access to the training data. It is possible that the MADAR test set we use for evaluation was seen by the models. Another limitation is that we did not optimize the translation prompt for specific models, which could lead to a lower translation quality for some of the systems. We also note that using a country name as a specification of a dialect is not optimal, as some dialects span multiple countries, or there are multiple dialects used within a single country. Another limitation is that we performed only a small-scale

human evaluation and mostly rely on automated metrics, which have many known issues.

Acknowledgments

This work was partially supported by SVV project number 260 821, by Czech Ministry of Education, Youth and Sports (grant MŠMT OP JAK Mezisektorová spolupráce CZ.02.01.01/00/23_020/0008518) and by EU EDF project ALADAN, Grant Agreement No 101102545.

It has been using language resources and tools developed and/or stored and/or distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2023062).

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Mark Fishel, Marco Gaido, Dávid Javorský, Marek Kasztelnik, and 33 others. 2025. Findings of the IWSLT 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 412–481, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Ibrahim Said Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, Barry Haddow, Dávid Javorský, Mateusz Krubiński, Tsz Kin Lam, Xutai Ma, Prashant Mathur, Evgeny Matusov, Chandresh Maurya, John P. McCrae, and 25 others. 2024. FINDINGS OF THE IWSLT 2024 EVALUATION CAMPAIGN. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Amjad Almahairi, Kyunghyun Cho, Nizar Habash, and Aaron Courville. 2016. First result on arabic neural machine translation. *Preprint*, arXiv:1606.02680.

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.

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

Laith H. Baniata, Se-Young Park, and Seong-Bae Park. 2018. A neural machine translation model for arabic dialects that utilises multitask learning (mtl). *Computational Intelligence and Neuroscience*, 2018.

Waad Ben Kheder, Josef Jon, André Beyer, Abdel Messaoudi, Rabea Affan, Claude Barras, Maxim Tychonov, and Jean-Luc Gauvain. 2024. ALADAN at IWSLT24 low-resource Arabic dialectal speech translation task. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 192–202, Bangkok, Thailand (in-person and online). 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).

Cohere Labs. 2024a. c4ai-command-r-08-2024.

Cohere Labs. 2024b. c4ai-command-r-v01.

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.

Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and Stephan Vogel. 2016. QCRI's machine translation systems for IWSLT'16. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Wael Farhan, Bashar Talafha, Analle Abuammar, Ruba Jaikat, Mahmoud Al-Ayyoub, {Ahmad Bisher} Tarakji, and Anas Toma. 2020. Unsupervised dialectal neural machine translation. *Information Processing and Management*, 57(3). Publisher Copyright: © 2019 Elsevier Ltd.

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

Nizar Habash and Jun Hu. 2009. Improving Arabic-Chinese statistical machine translation using English

as pivot language. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 173–181, Athens, Greece. Association for Computational Linguistics.

N.Y. Habash. 2010. *Introduction to Arabic Natural Language Processing*. Synthesis digital library of engineering and computer science. Morgan & Claypool Publishers.

Ahmed Heakl, Youssef Zaghloul, Mennatullah Ali, Rania Hossam, and Walid Gomaa. 2024. Arzen-llm: Code-switched egyptian arabic-english translation and speech recognition using llms. *Procedia Computer Science*, 244:113–120. 6th International Conference on AI in 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. 2025a. 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.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025b. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report. *Preprint*, arXiv:2506.04079.

Mistral-Team. 2025. Introducing mistral 3.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2021. Investigating code-mixed Modern Standard Arabic-Egyptian to English machine translation. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 56–64, Online. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022a. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022b. 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.

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

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. 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.

Aziz Mohammed Abdo Saeed. 2025. Machine translation evaluation between arabic and english during

2020 to 2024: A review study. *Arts for Linguistic & Literary Studies*, 7(2):665–678.

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.

Hassan Sajjad, Kareem Darwish, and Yonatan Belinkov. 2013. Translating dialectal Arabic to English. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6, Sofia, Bulgaria. Association for Computational Linguistics.

Wael Salloum and Nizar Habash. 2013. Dialectal Arabic to English machine translation: Pivoting through Modern Standard Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 348–358, Atlanta, Georgia. Association for Computational Linguistics.

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

Gemma Team. 2025. Gemma 3.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. 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.

A Additional results

A.1 English to Arabic

The total BLEU and ChrF scores, computed on the concatenation of all test sets, are presented in Table 5. These scores again show the leading performance of Jais-2 and gpt-4.1 models, followed by Command-R and Gemma 3.

Model	Size	BLEU	CHRF
Arabic-Specialized			
Jais-2-70B-Chat	70B	25.4	51.8
Jais-2-8B-Chat	8B	14.5	41.6
Nile-Chat-12B	12B	7.3	34.1
c4ai-command-r7b-arabic-02-2025	7B	8.1	34.5
Multilingual			
EuroLLM-9B-Instruct	9B	7.0	33.3
Llama-3.3-70B-Instruct	70B	8.6	35.3
Mistral-Small-3.2-24B-Instruct-2506	24B	6.0	28.4
Qwen3-4B-Instruct-2507	4B	4.6	27.8
aya-expanse-32b	32B	9.9	38.0
aya-expanse-8b	8B	6.5	33.2
c4ai-command-r-08-2024	32B	11.1	39.1
c4ai-command-r-v01	35B	5.4	30.2
command-a-translate-08-2025	111B	8.7	37.7
gemma-3-27b-it	27B	10.5	39.2
gemma-3-4b-it	4B	7.1	32.7
Commercial API			
gpt-4.1-mini	N/A	15.0	44.8
gpt-4.1-nano	N/A	12.4	42.1

Table 5: Overall BLEU and ChrF scores computed on the concatenation of all test sets translated using the dialect-specific prompts.

Figure 9 shows the BLEU scores on single test sets for translations produced using the dialect-specific prompt. For comparison, Figure 10 shows BLEU scores for translations produced with general Arabic prompt. Figure 11 shows the ChrF scores for translations with the general prompt and can be compared to the dialect-specific prompt scores in Figure 4. Comparison of plots allows to distinguish which models are sensitive to dialect-specific prompts and able to produce adequate dialectal realization of the translation.

A.2 Arabic to English

Figure 12 shows ChrF scores for Arabic to English translations. The scores are comparatively higher compared to English to Arabic scores (Figure 8), which may reflect both better translation quality and lower morphological complexity of English.

A.3 COMET scores

Figures 13, 14, 15 and 16 show reference and reference-less versions of COMET-XL on translation using the dialect-specific prompt (Figures 13 and 14) and the general Arabic prompt (Figures 15 and 16). In the comparison with other

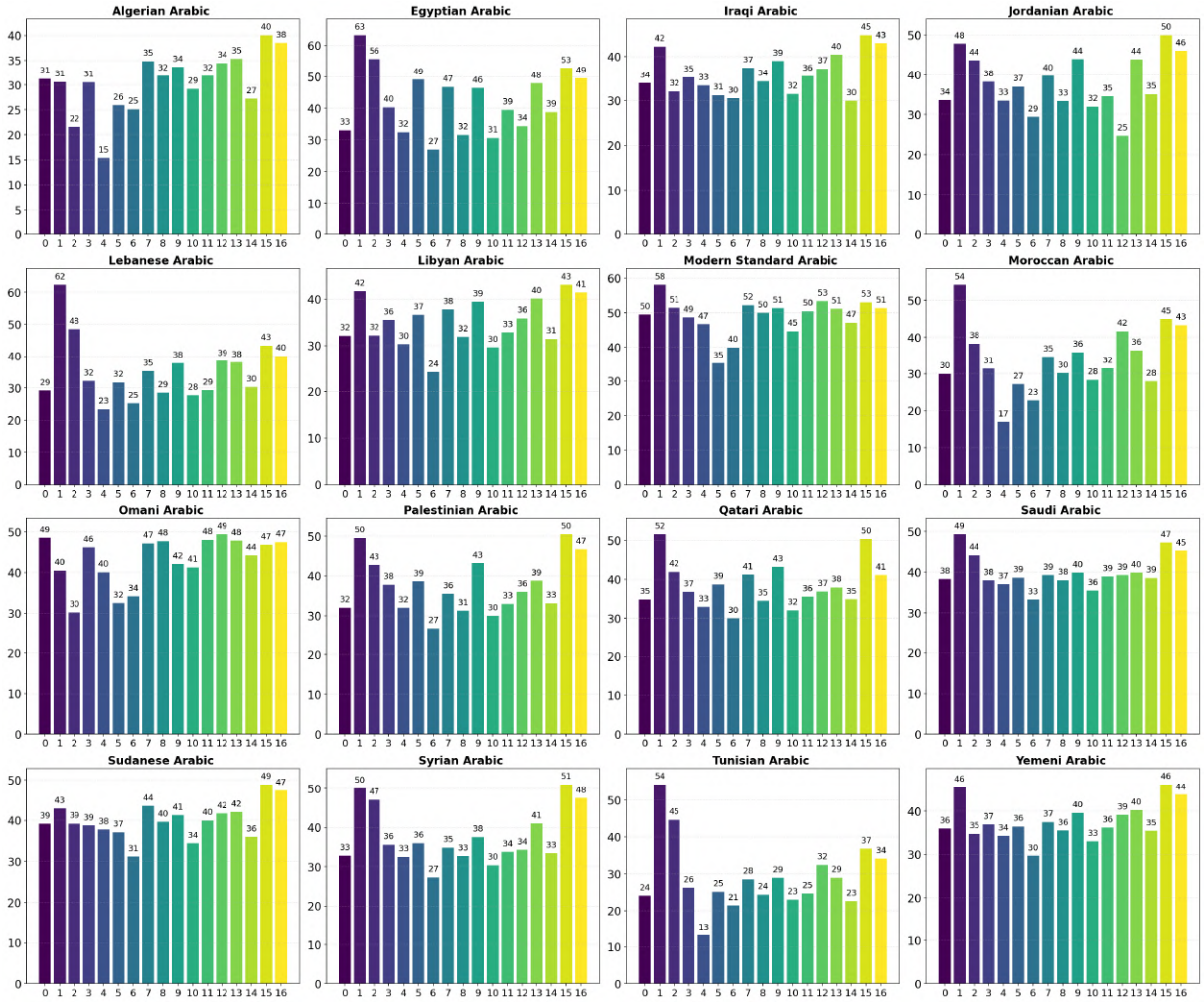

Figure 8: ChrF scores for merged dialect test sets for all models in English to Arabic, using the dialect-specific prompt. See Figure 2 for the mapping of the indices on the x axis to model names.

metrics and our human evaluation, this shows that COMET scores actually penalize translations that are correctly realized in dialectal Arabic, due to the preference of the COMET model for MSA. Using references with COMET only partially mitigates this issue, the MSA translations are still preferred. See for example the correlations in Figure 6 for further evidence. For the Arabic to English direction, we show the reference-less and reference-based scores in Figures 17 and 18, respectively.

A.4 Manual Error Analysis

The errors observed throughout the human annotation process fell into one of these four categories:

- **Adequacy**: This category included instances of mistranslation, partial loss of source meaning, literal translation, and severely inadequate or irrelevant outputs, where the meaning of the source sentence was partially or completely distorted.

An example of a literal translation error is the word "around" in the phrase "around here," which was translated as حـوَالي meaning "around" in the sense of an approximate quantity rather than location.

- **Fluency**: Fluency errors included orthographic errors, grammatical (morphosyntactic) errors such as agreement and inflection mistakes, for example, the generated translation for How much is the breakfast? is الفطَار كَام (How much breakfast) instead of الفطَار

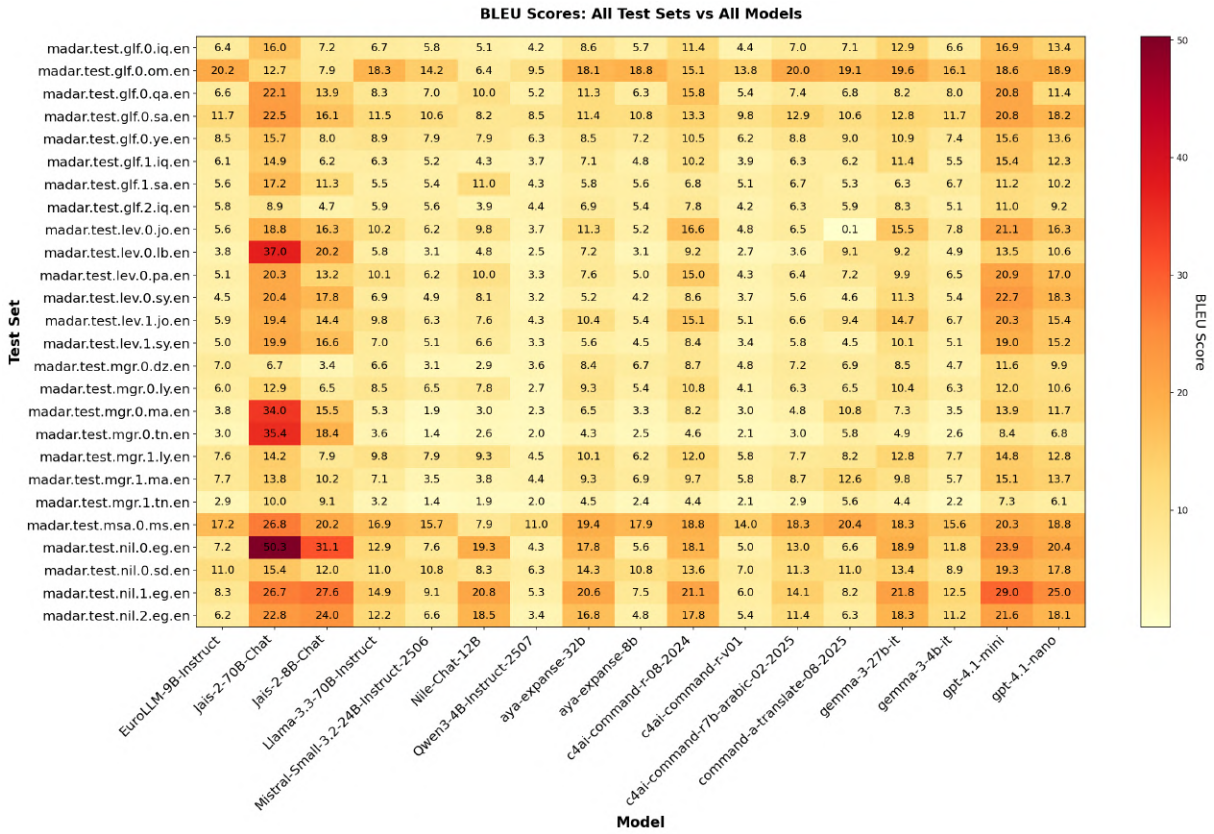

Figure 9: BLEU scores for individual test sets for all models in English to Arabic direction. Translation produced using a dialect-specific prompt.

بِكَام with the correct proposition and meaning in addition to awkward or non-idiomatic phrasing, where the sentence was semantically correct but sounded unnatural to a native speaker.

- **Dialectal appropriateness**: This category captured cases where the output did not conform to the target dialect. Errors ranged from minor cases, where only a small portion of the output was in Modern Standard Arabic (MSA), to major cases, where the output was predominantly or entirely in MSA or contained incorrect dialectal lexical choices.
- **Completeness**: This category included omission errors (truncated or incomplete outputs), insertion errors (addition of words not present in the source), and copying errors, such as untranslated or transliterated words.

Table 6 shows the detailed error analysis for each model.

A.5 Dialect classification

Figures 19, 20, 21 and 22 show dialect classification performed by the `Jais-2-70B-Chat` for test sets translated using the dialect specific prompt by `Jais-2-70B-Chat`, `gpt-4.1-mini`, `EuroLLM-9B-Instruct` and `Nile-Chat-12B`, respectively. We can see a number of interesting trends. First of all, we see that according to the assessment by the classification model, Jais-2-70B-Chat and gpt-4.1-mini produce outputs that have similar distribution dialects as the reference. It is apparent that they reflect the dialect name in the prompt, for comparison, see Figures 23, 24, 25, 26, which present the classification results on translation produced by the same models general Arabic prompt instead of dialect-specific one. The appropriate dialects are notably more represented in the translations with the dialect-specific prompt. On the other hand, the EuroLLM model is not producing the texts in the correct dialects, using mostly MSA instead. This tendency was also shown by our manual evaluation. Another observation is related to the Nile-Chat-12B model, which, regardless of dialect specification, produces texts in the Egyptian and Sudanese dialects. This is intended by the model authors and our evaluation thus confirms that the model is Nile-dialects specific.

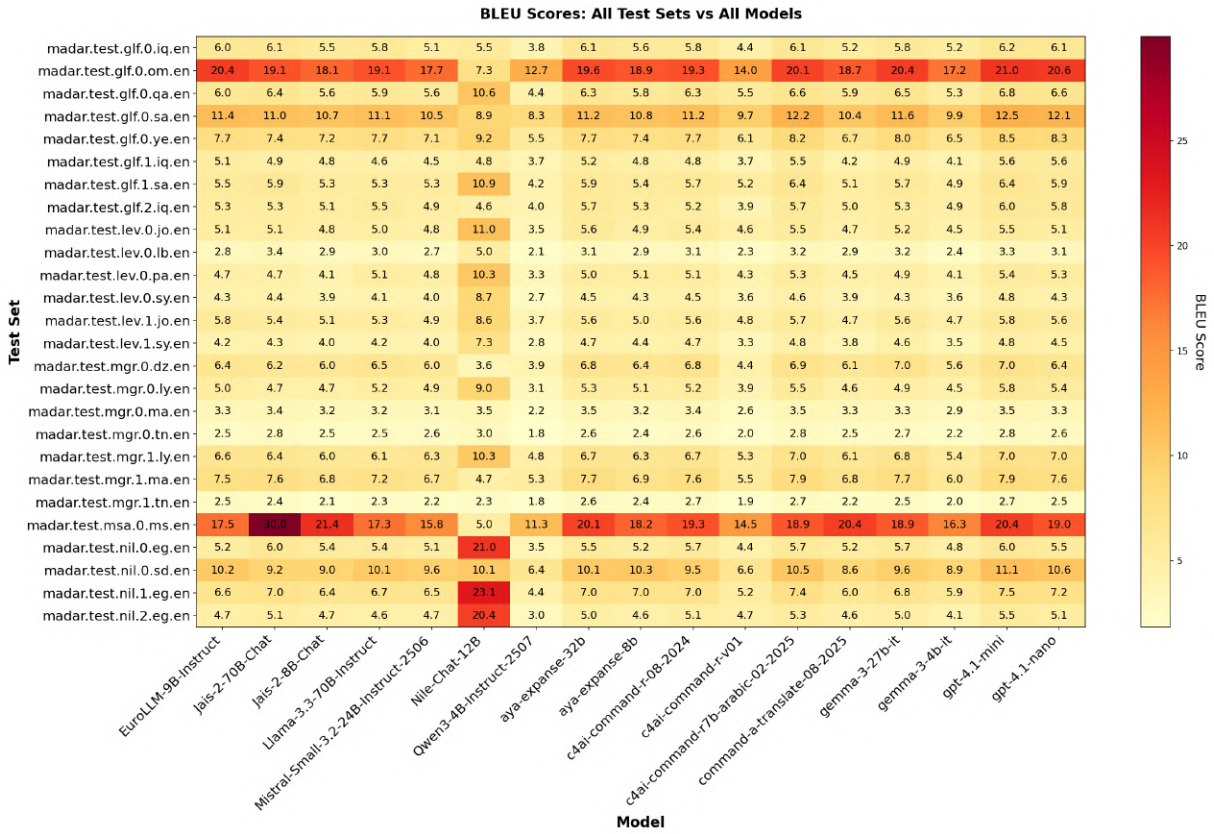

Figure 10: BLEU scores for individual test sets for all models in English to Arabic direction. Translation produced using a general Arabic prompt.

Error Category	**J70B**	**G-mini**	**ELLM-9B**	**CR-824**
Adequacy				
Mistranslation	10 (66.7%)	6 (21.4%)	13 (13.8%)	24 (34.8%)
Hallucinated translation	1 (6.7%)	0 (0.0%)	0 (0.0%)	1 (1.5%)
Literal Translation	0 (0.0%)	11 (39.3%)	0 (0.0%)	1 (1.5%)
Fluency				
Orthographical Error	1 (6.7%)	1 (3.6%)	6 (6.4%)	0 (0.0%)
Morphosyntactic Error	1 (6.7%)	2 (7.1%)	5 (5.3%)	4 (5.8%)
Non-idiomatic phrasing	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (1.5%)
Dialectal appropriateness				
Dialect Mismatch (Major)	0 (0.0%)	0 (0.0%)	59 (62.8%)	7 (10.1%)
Dialect Mismatch	0 (0.0%)	7 (25.0%)	10 (10.6%)	30 (43.5%)
Completeness				
Copying	1 (6.7%)	0 (0.0%)	1 (1.1%)	1 (1.5%)
Omission	1 (6.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Insertion	0 (0.0%)	1 (3.6%)	0 (0.0%)	0 (0.0%)
Total Errors	15	28	94	69

Table 6: Detailed Error Analysis of Egyptian Arabic Translations generated by: Jais-2-70B- Chat (J70B), GPT-4.1-mini (G-mini), EuroLLM-9B- Instruct (ELLM-9B), Command-r-08-2024 (CR-824).

Figures 28, 29, 30 and 31 present the distributions of dialects in the translations performed by the same set of models as above, using the dialect-specific prompt, but classified us-

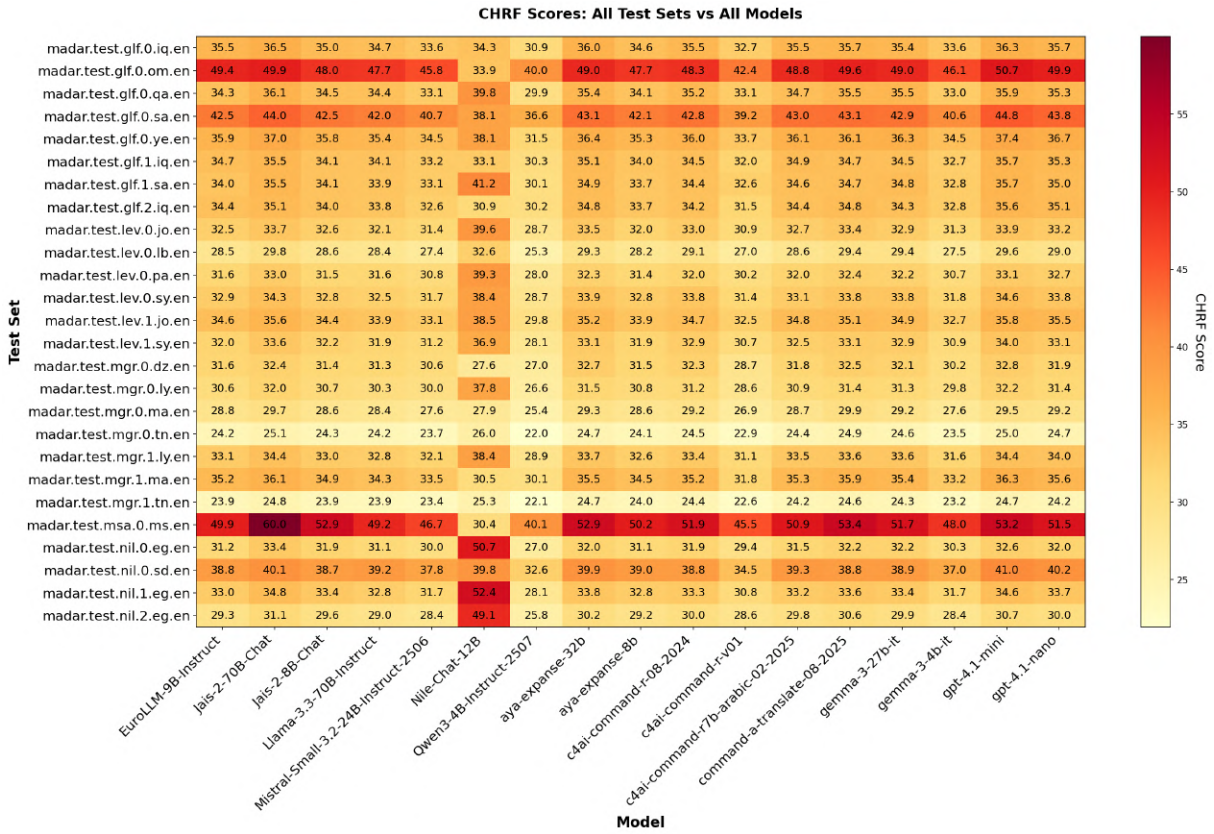

Figure 11: ChrF scores for individual test sets for all models in English to Arabic direction. Translation produced using a general Arabic prompt.

ing `gemma-27b-it` instead of `Jais-2-70B-Chat`. Figure 27 shows the dialect distribution in the reference translations, also assessed by the `gemma-27b-it` model. It is apparent that different LLMs have different priors about the dialect distributions, e.g. `gemma-27b-it` classifies noticeably larger portion of the texts as Lebanese, even for test sets from other dialects.

B Prompts

B.1 Translation Prompt

System prompt

```
You are a professional, very precise translator and a native speaker. Translate inputs based on the instructions and always print out only the text of the best possible translation, with no explanations.
Keep the same formatting (e.g. markup, lines, spacing) as the original. Do not translate untranslatable parts of the input (URLs, code, and similar).
```

Instruction prompt

```
Translate the following text into {lang}, only print out the translation, not add any explanations: {line}
```

B.2 GEMBA Translation Evaluation Prompt

```
Score the following translation from {source_lang} to {target_lang} on a continuous scale from 0 to 100, where a score of zero means “no meaning preserved” and a score of one hundred means “perfect meaning and grammar”.
Be very strict with checking the target language or dialect. For example, if an Arabic dialect is specified and the translation is in Modern Standard Arabic or another dialect, penalize harshly by subtracting 50 points from the score you would otherwise give, if the dialect is completely incorrect.
{source_lang} source: “{source_seg}”
{target_lang} translation: “{target_seg}”
Score:
```

B.3 Dialect Classification Prompt

```
You are an expert in Arabic dialects. Classify the following Arabic text by determining which country or region it is from.
Choose the most appropriate country from this list: {countries_str}, or “Modern Standard Arabic” if it is formal standard Arabic, or “Other” if none of the above apply.
```

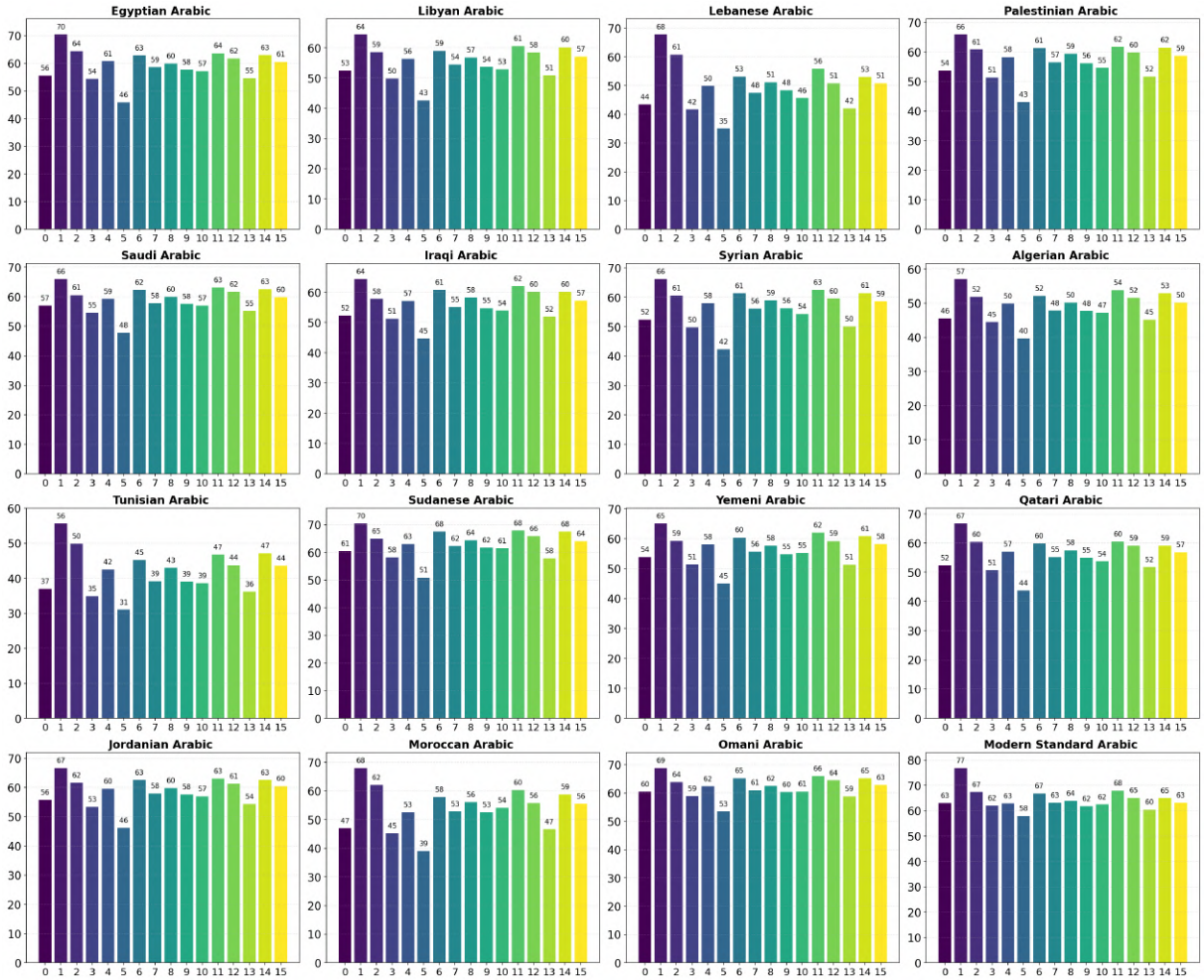

Figure 12: ChrF scores for merged dialect test sets for all models in Arabic to English direction. See Figure 2 for the mapping of the indices on the x axis to model names.

```
Arabic text: “{text}”
Respond with ONLY the country name from the list
above, nothing else.
```

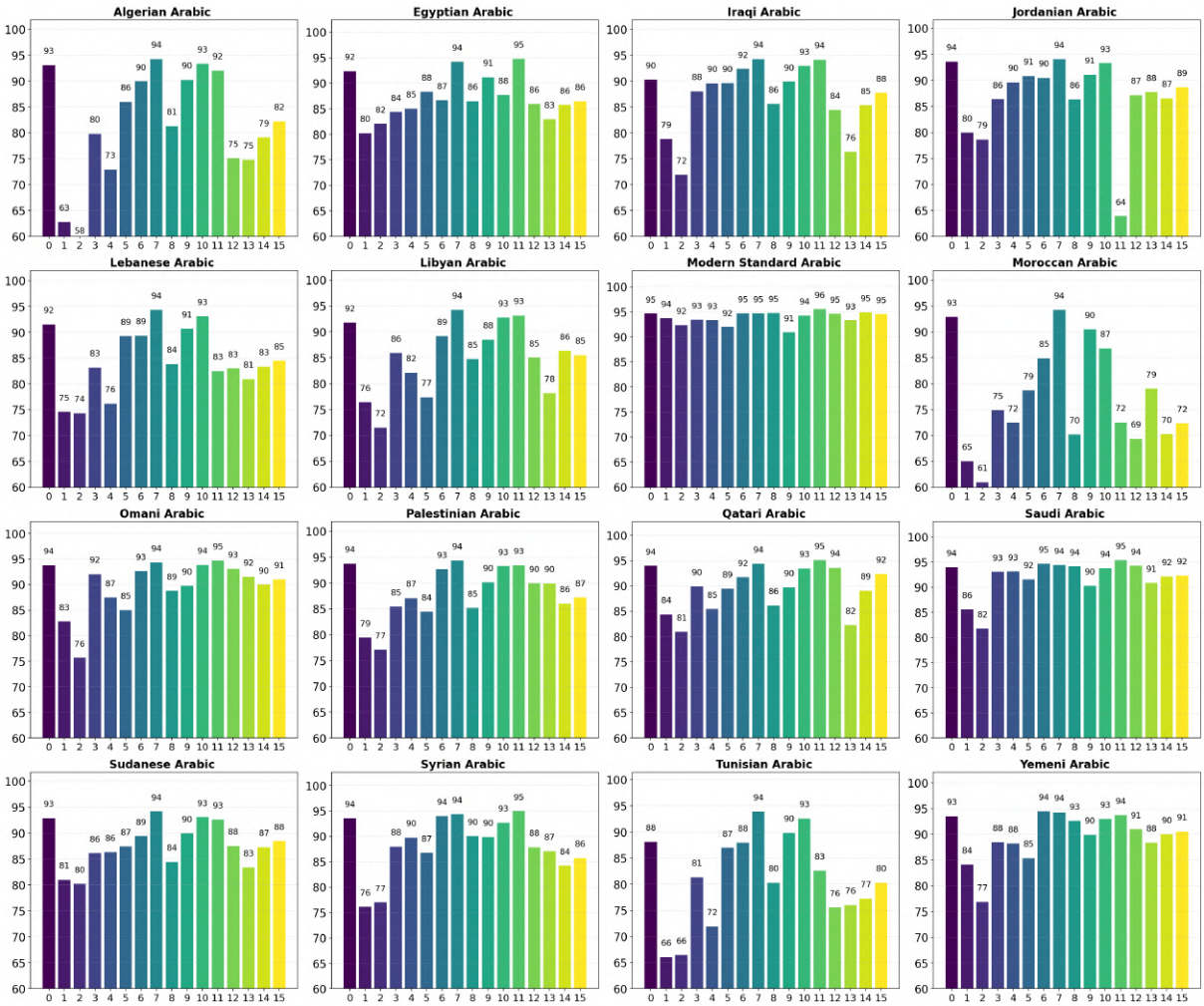

Figure 13: COMET scores **without** using the reference for merged dialect test sets for all models in English to Arabic, using the **dialect-specific prompt**. See Figure 2 for the mapping of the indices on the *x* axis to model names.

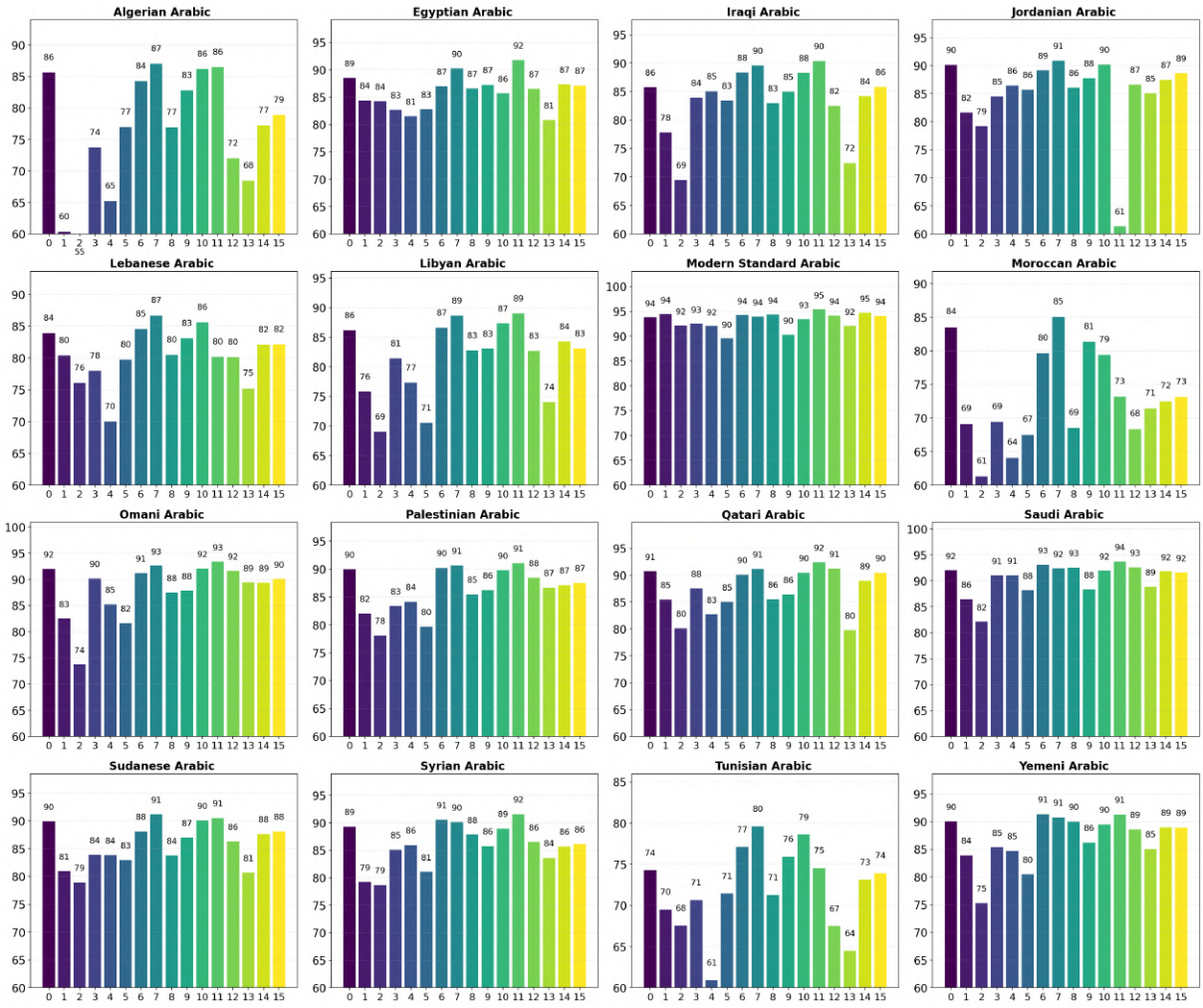

Figure 14: COMET scores **using the reference** for merged dialect test sets for all models in English to Arabic, using the **dialect-specific prompt**. See Figure 2 for the mapping of the indices on the *x* axis to model names.

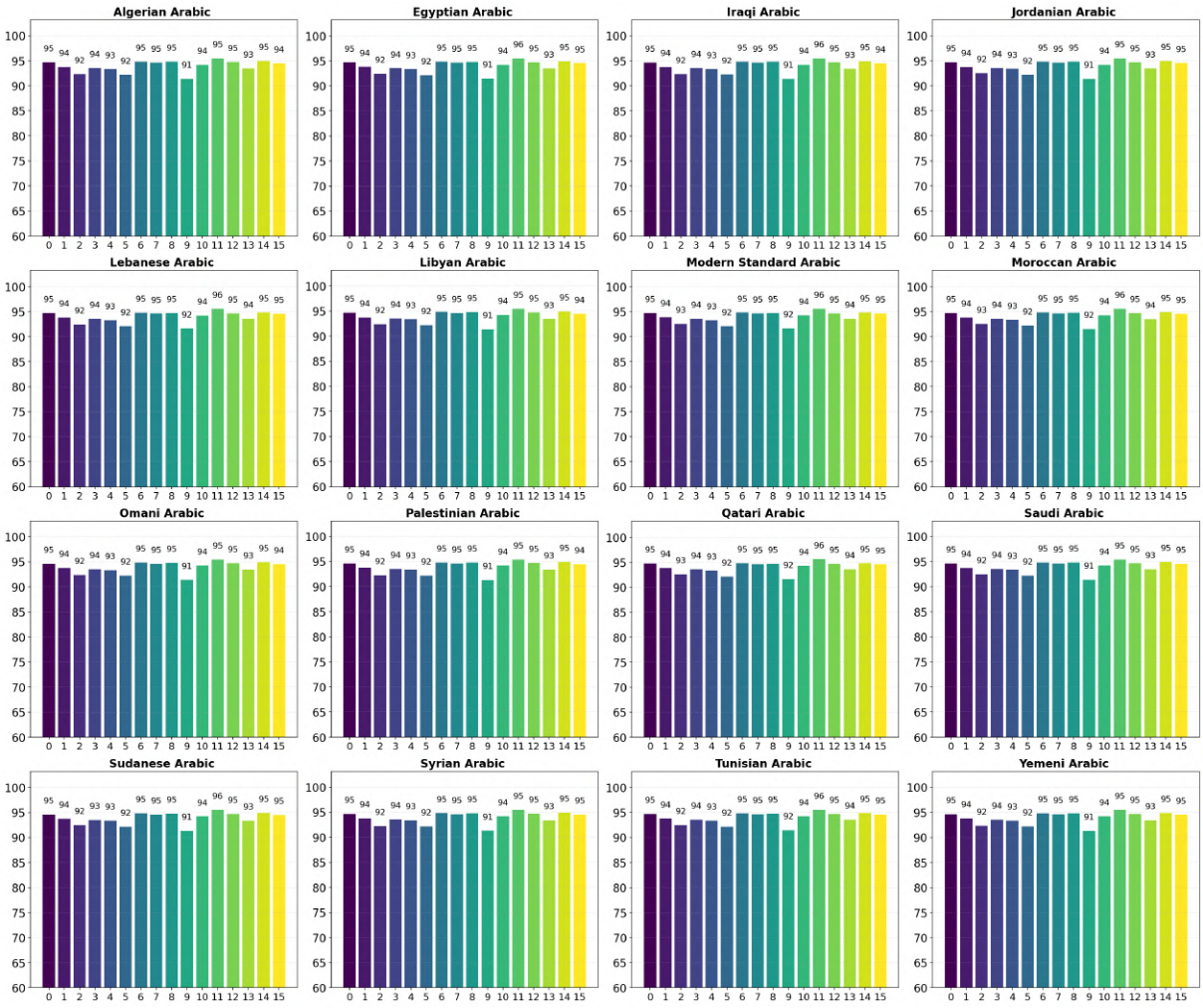

Figure 15: COMET scores without using the reference for merged dialect test sets for all models in English to Arabic, using the **general Arabic prompt** See Figure 2 for the mapping of the indices on the *x* axis to model names.

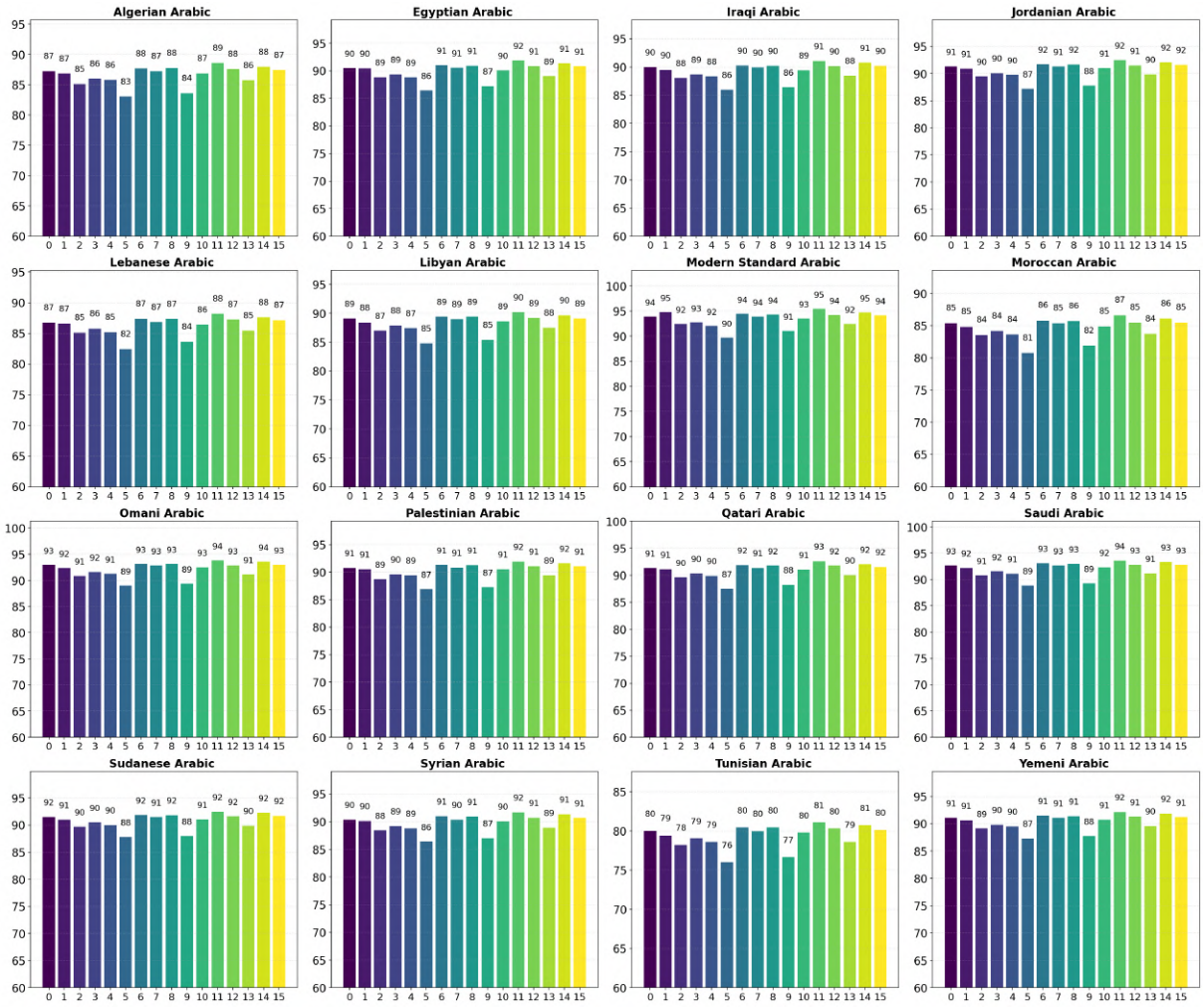

Figure 16: COMET scores **using the reference** for merged dialect test sets for all models in English to Arabic, using the **general Arabic prompt**. See Figure 2 for the mapping of the indices on the *x* axis to model names.

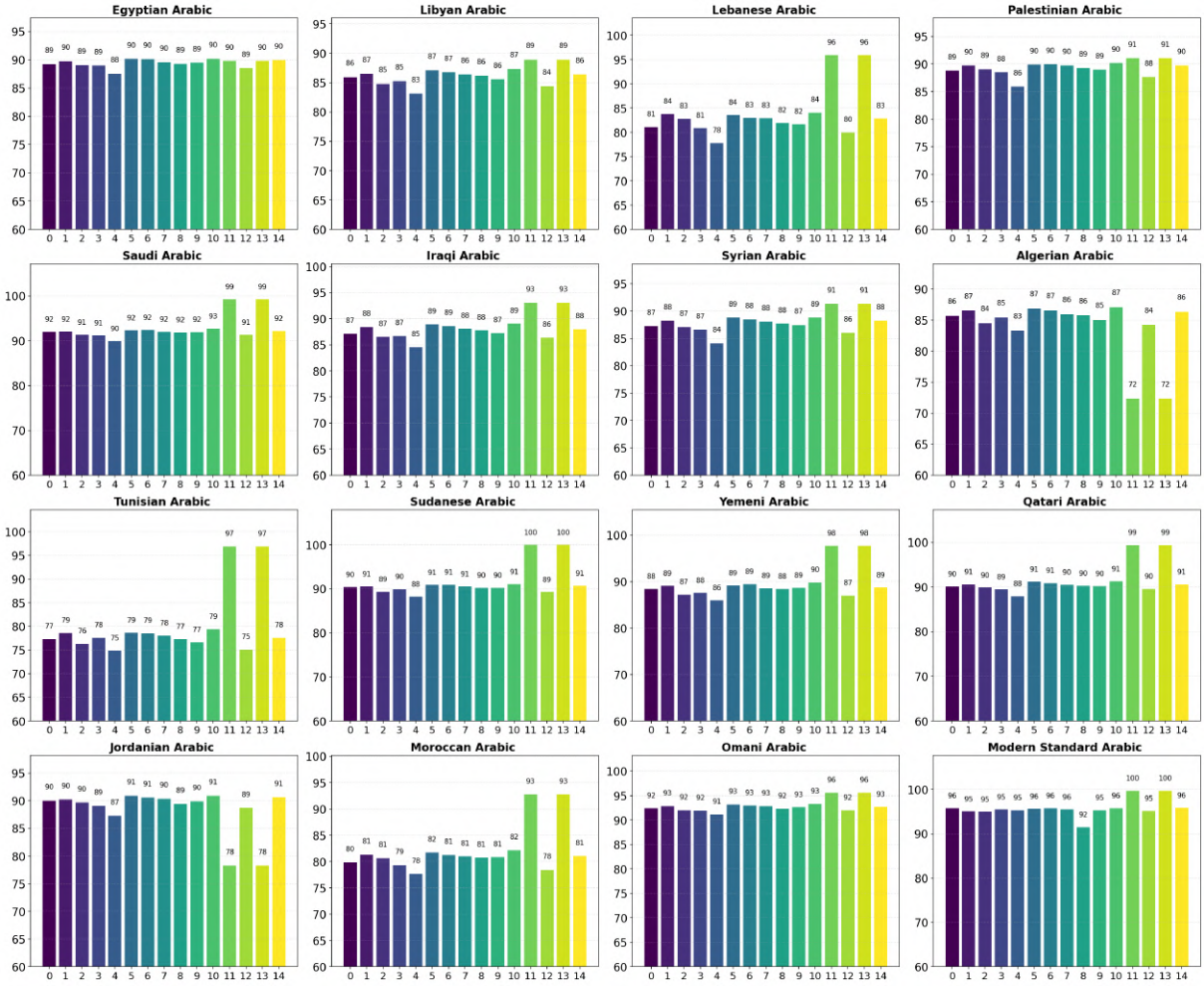

Figure 17: COMET scores without using the reference for merged dialect test sets for all models in Arabic to English. See Figure 2 for the mapping of the indices on the *x* axis to model names.

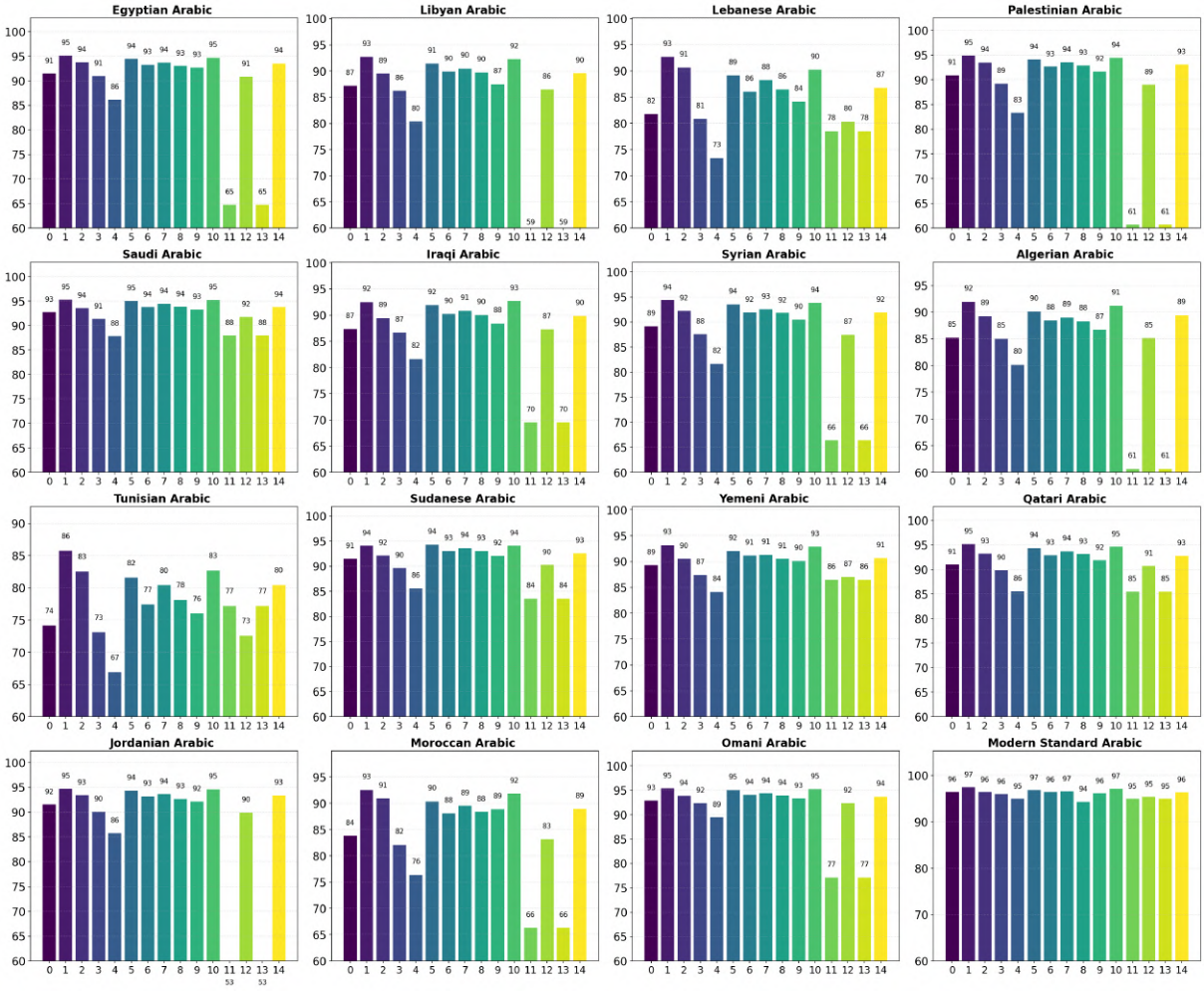

Figure 18: COMET scores **using the reference** for merged dialect test sets for all models in Arabic to English. See Figure 2 for the mapping of the indices on the *x* axis to model names.

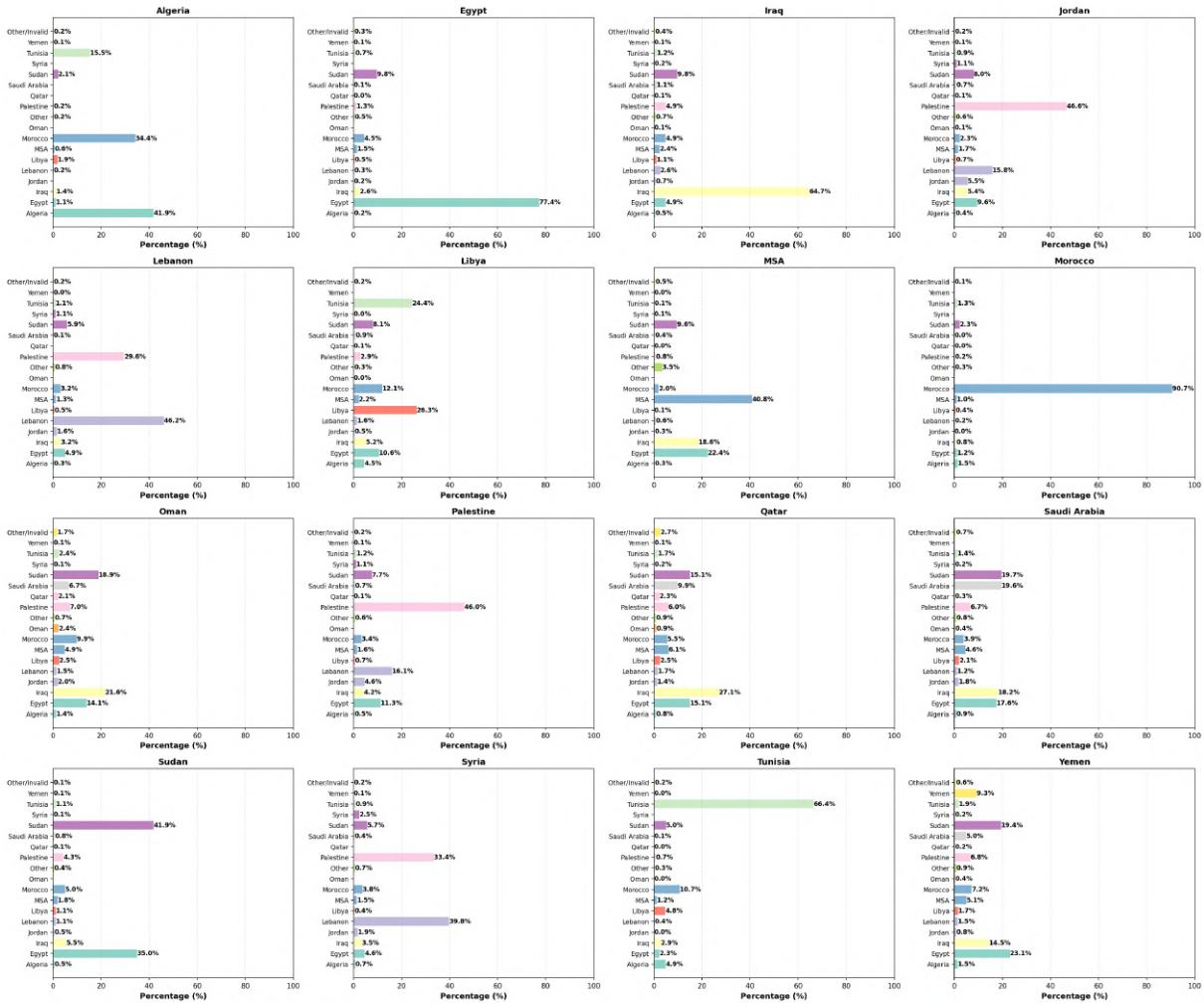

Figure 19: Distribution of dialects in translations produced by `Jais-2-70B-Chat` using the dialect-specific prompt. The dialects are classified on the sentence level, using `Jais-2-70B-Chat`.

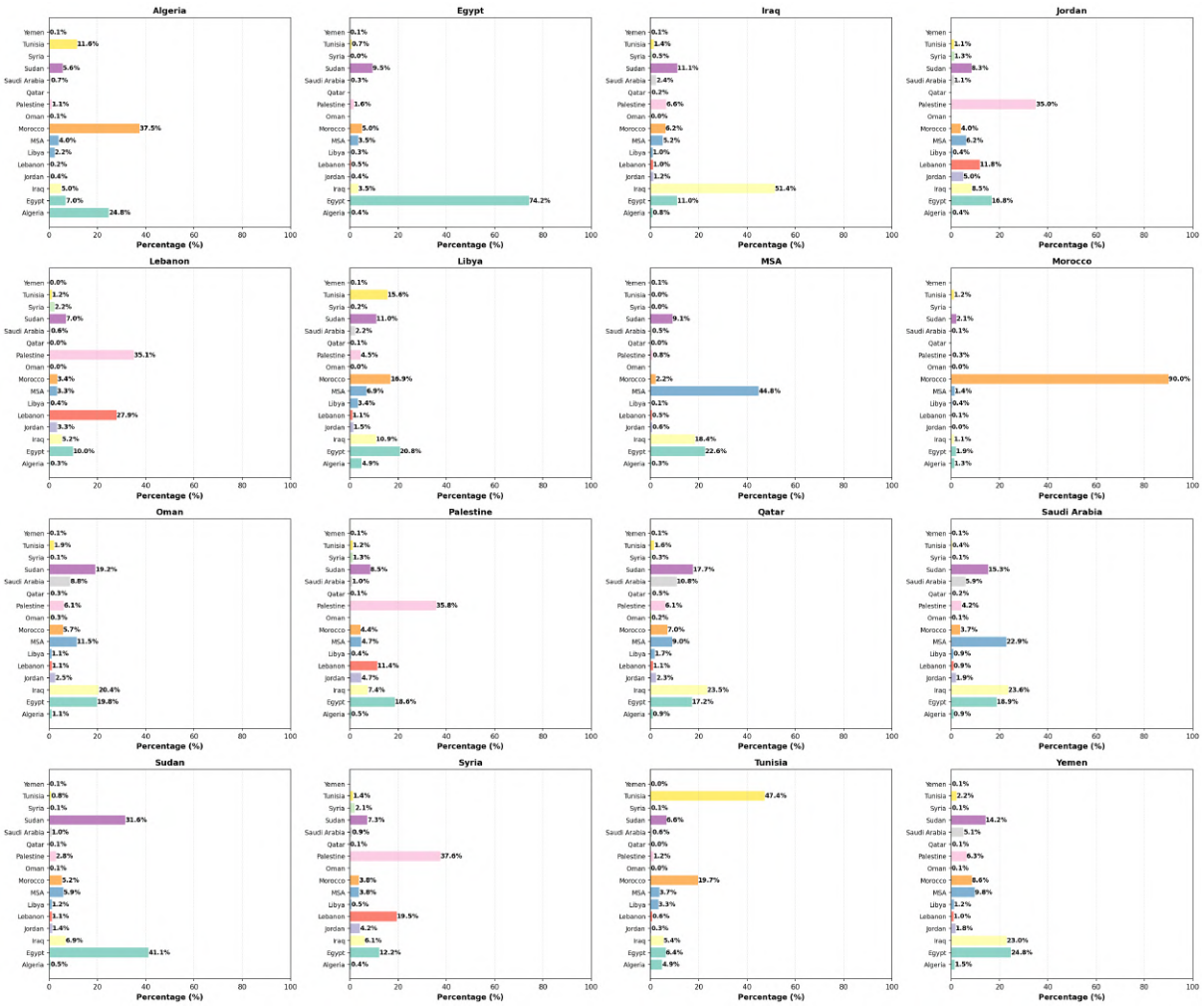

Figure 20: Distribution of dialects in translations produced by `gpt-4.1-mini` using the dialect-specific prompt. The dialects are classified on the sentence level, using `Jais-2-70B-Chat`.

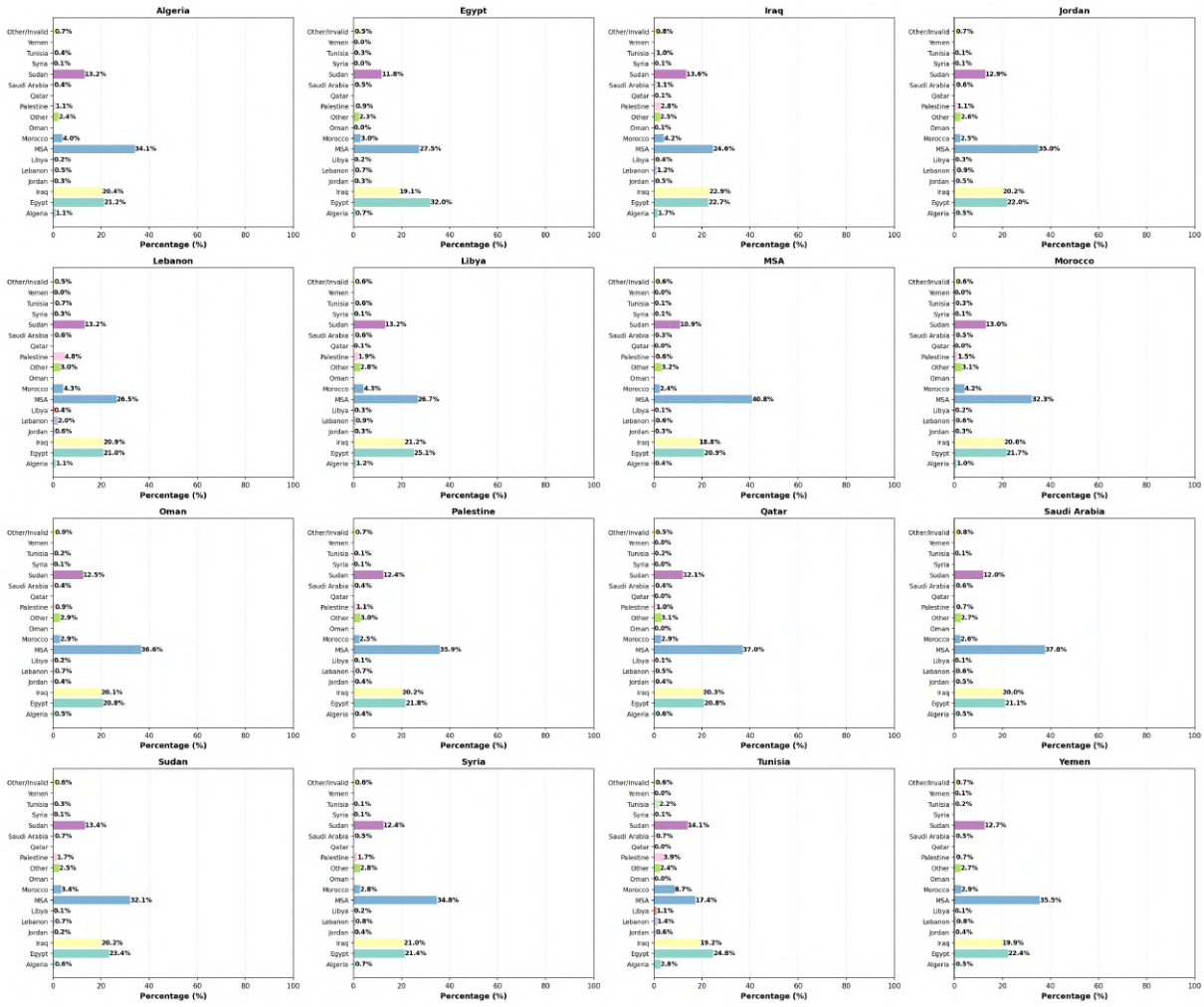

Figure 21: Distribution of dialects in translations produced by `EuroLLM-9B-Instruct` using the dialect-specific prompt. The dialects are classified on the sentence level, using `Jais-2-70B-Chat`.

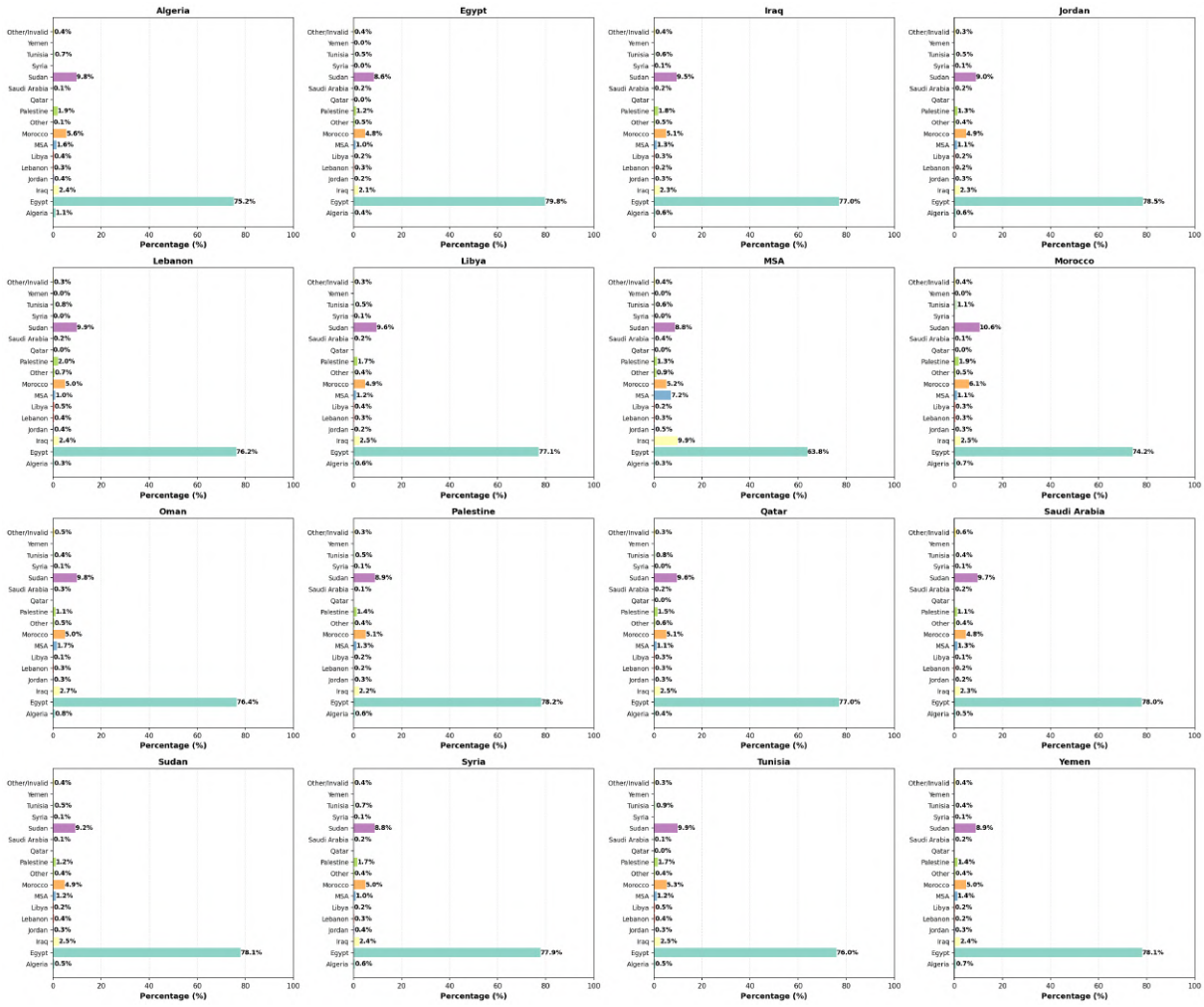

Figure 22: Distribution of dialects in translations produced by `Nile-Chat-12B` using the dialect-specific prompt. The dialects are classified on the sentence level, using `Jais-2-70B-Chat`.

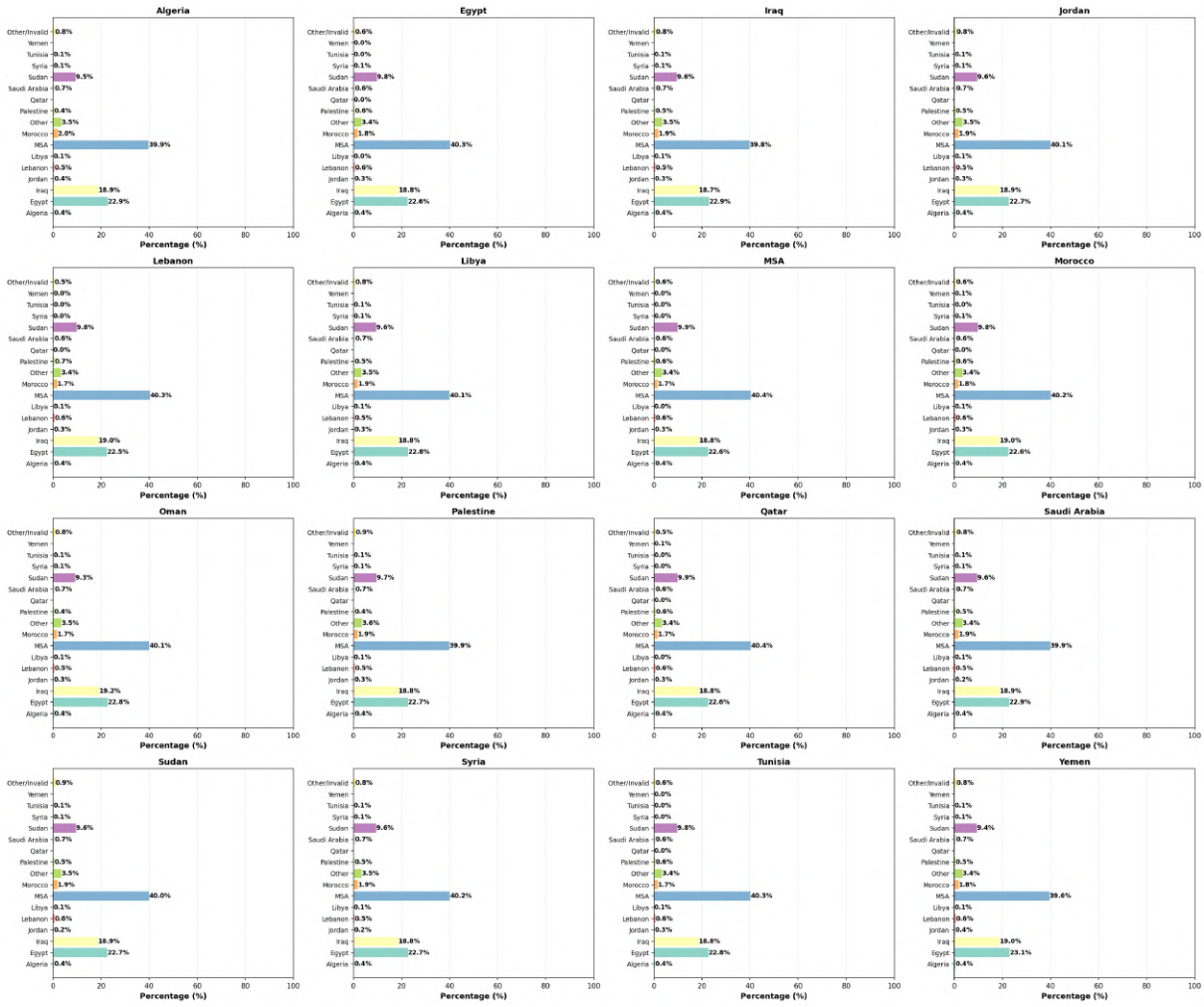

Figure 23: Distribution of dialects in translations produced by `Jais-2-70B-Chat` using the general Arabic prompt. The dialects are classified on the sentence level, using `Jais-2-70B-Chat`.

Figure 24: Distribution of dialects in translations produced by `gpt-4.1-mini` using the general Arabic prompt. The dialects are classified on the sentence level, using `Jais-2-70B-Chat`.

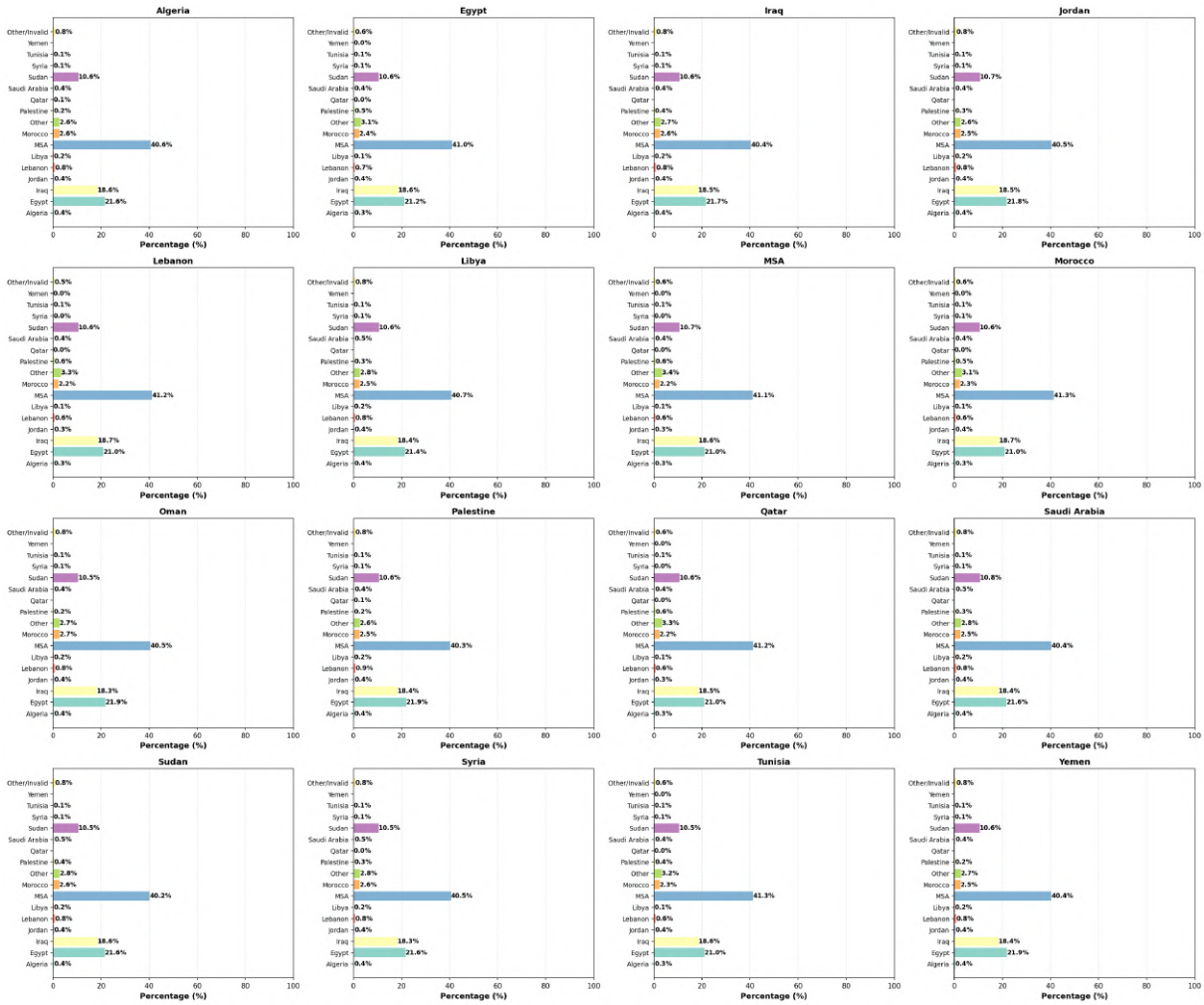

Figure 25: Distribution of dialects in translations produced by `EuroLLM-9B-Instruct` using the general Arabic prompt. The dialects are classified on the sentence level, using `Jais-2-70B-Chat`.

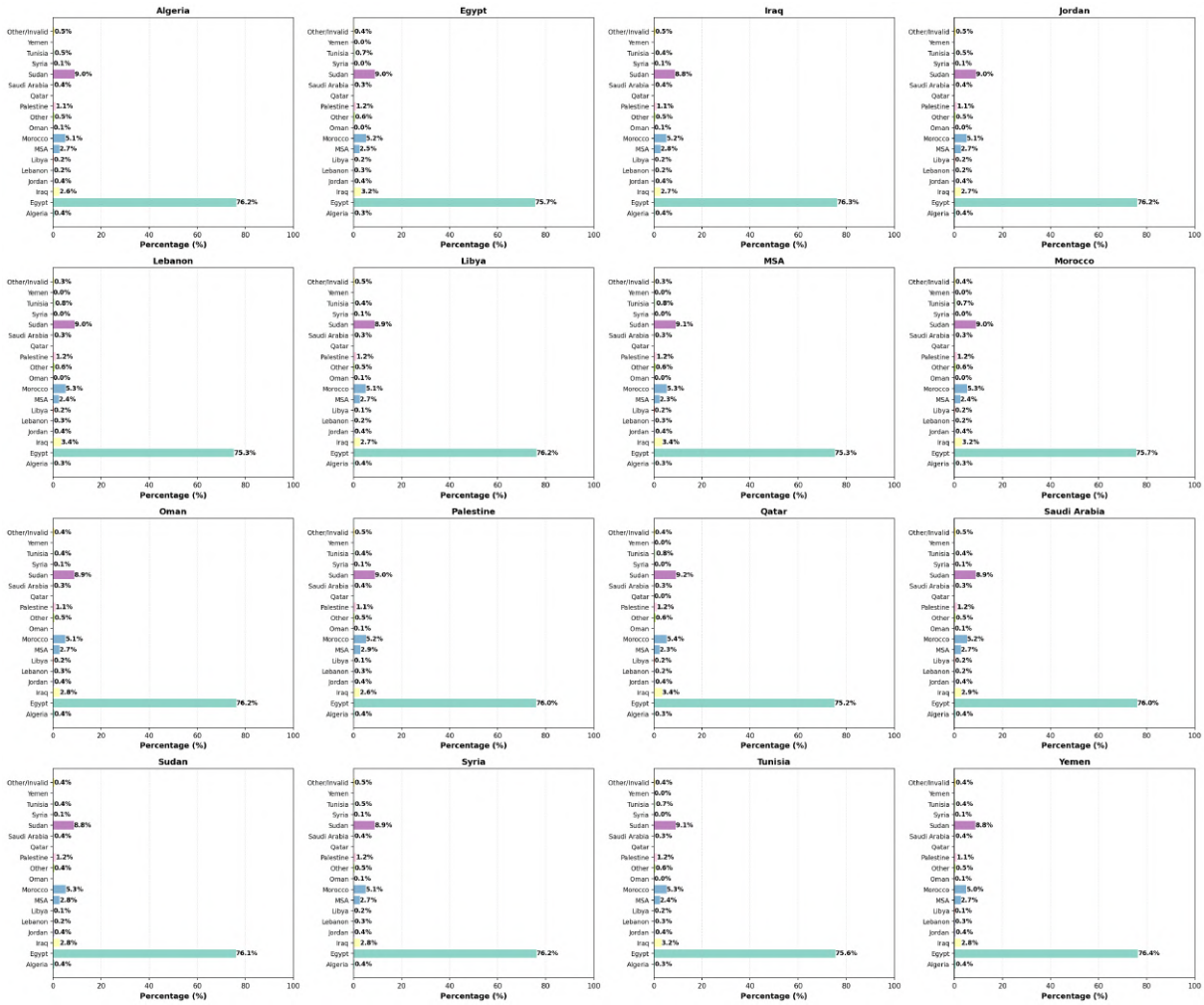

Figure 26: Distribution of dialects in translations produced by `Nile-Chat-12B` using the general Arabic prompt. The dialects are classified on the sentence level, using `Jais-2-70B-Chat`.

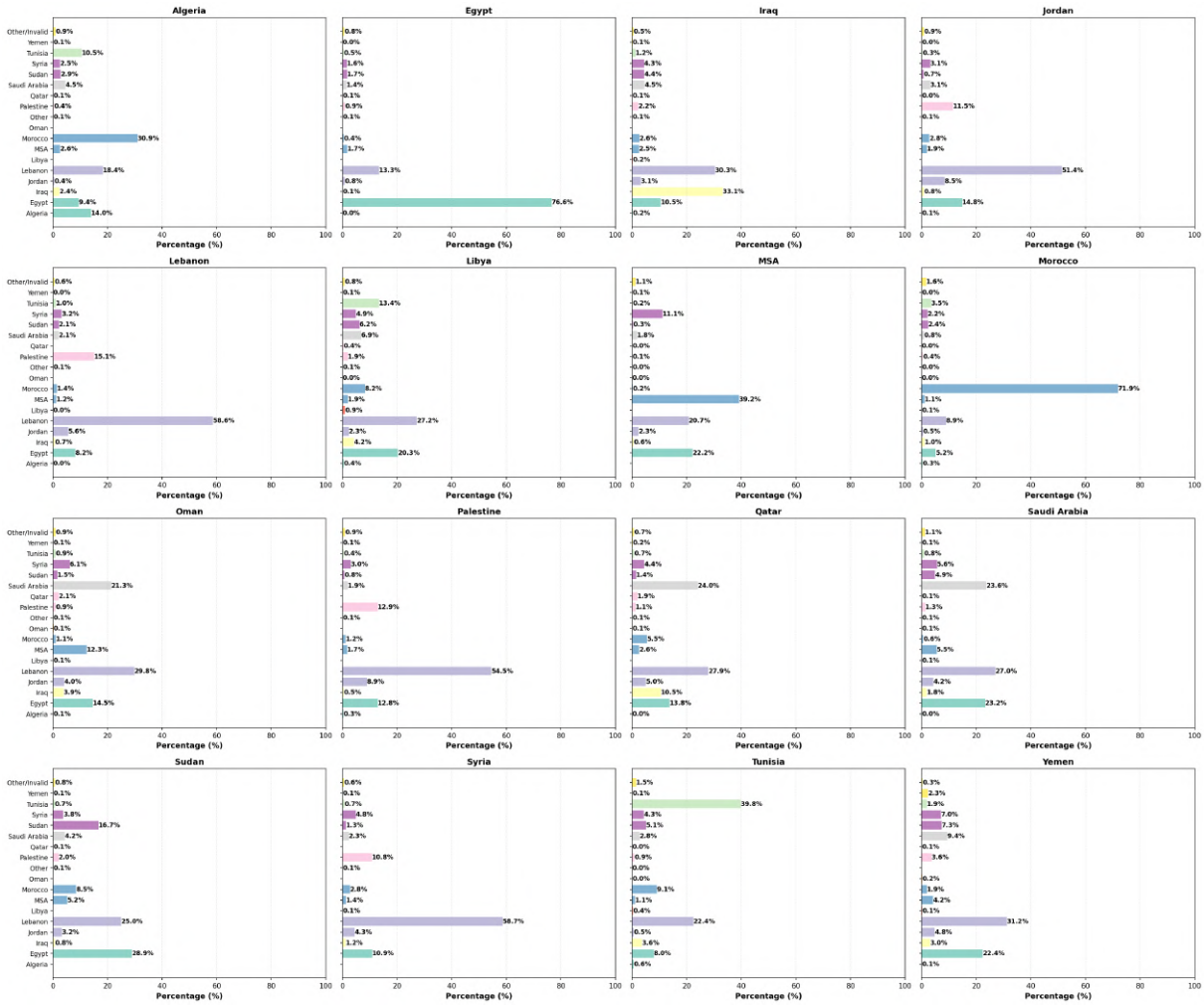

Figure 27: Distribution of dialects in the reference translations. The dialects are classified on the sentence level, using `gemma-3-27b-it`.

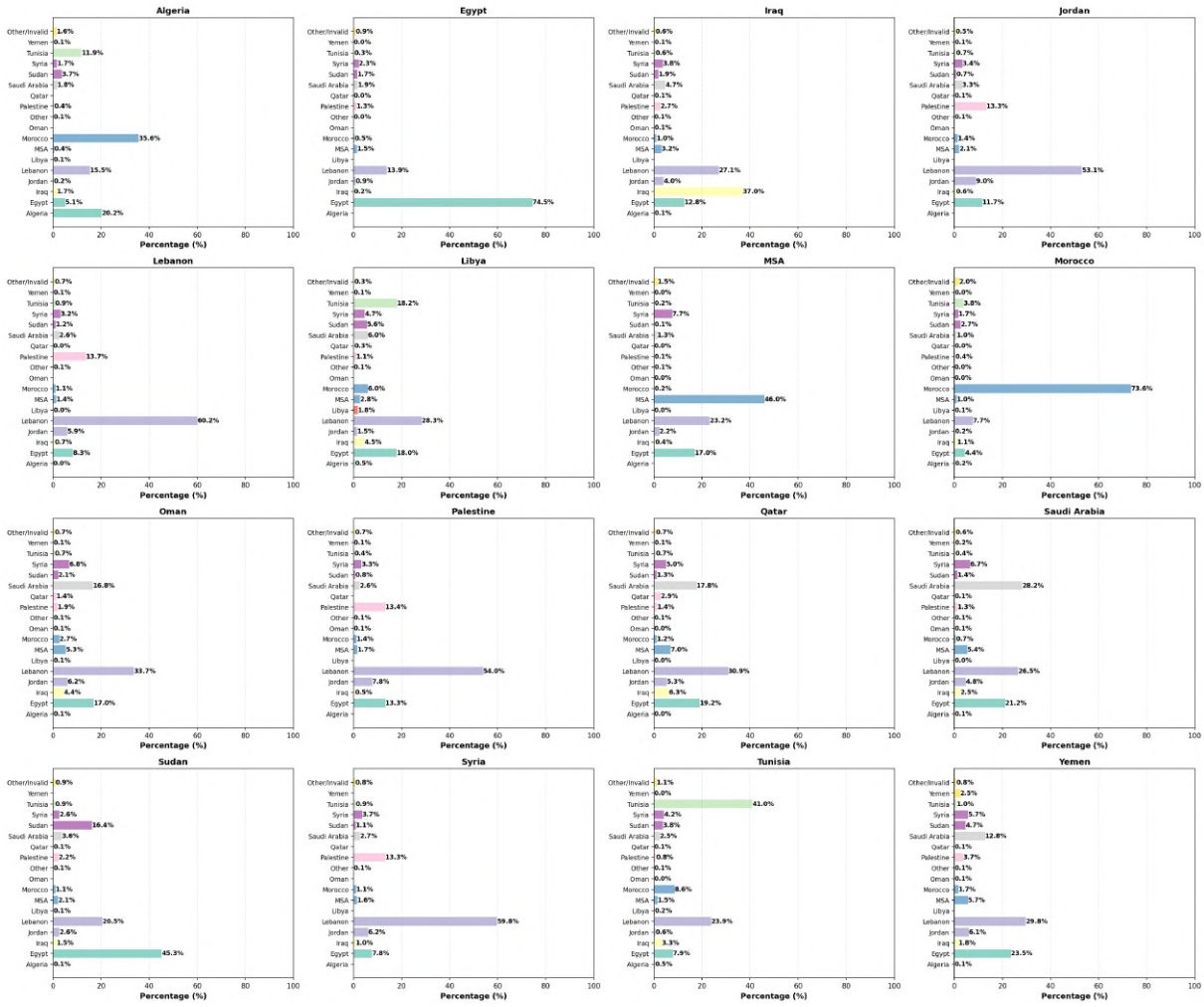

Figure 28: Distribution of dialects in translations produced by `Jais-2-70B-Chat` using the dialect-specific prompt. The dialects are classified on the sentence level, using `gemma-3-27b-it`.

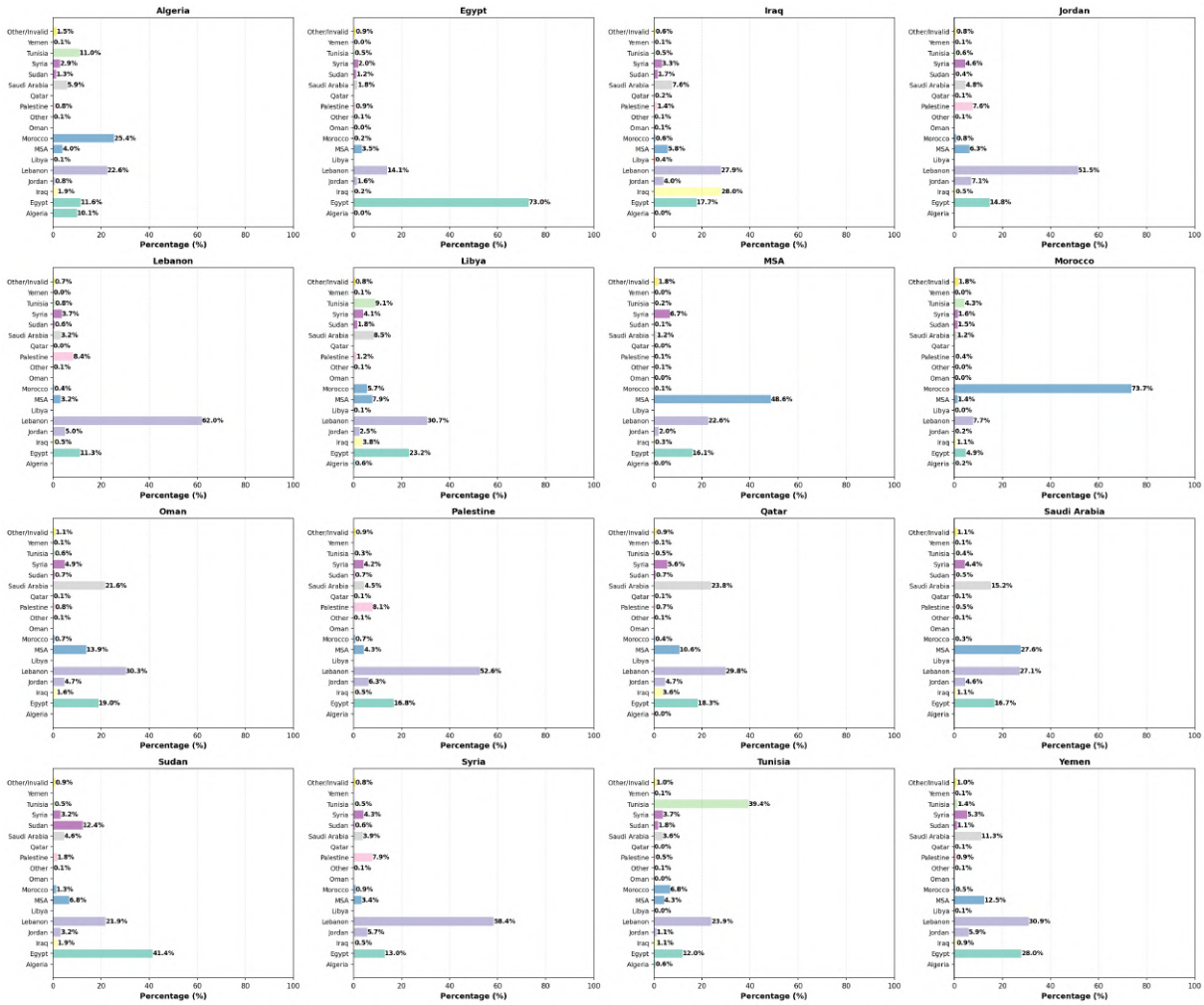

Figure 29: Distribution of dialects in translations produced by `gpt-4.1-mini` using the dialect-specific prompt. The dialects are classified on the sentence level, using `gemma-3-27b-it`.

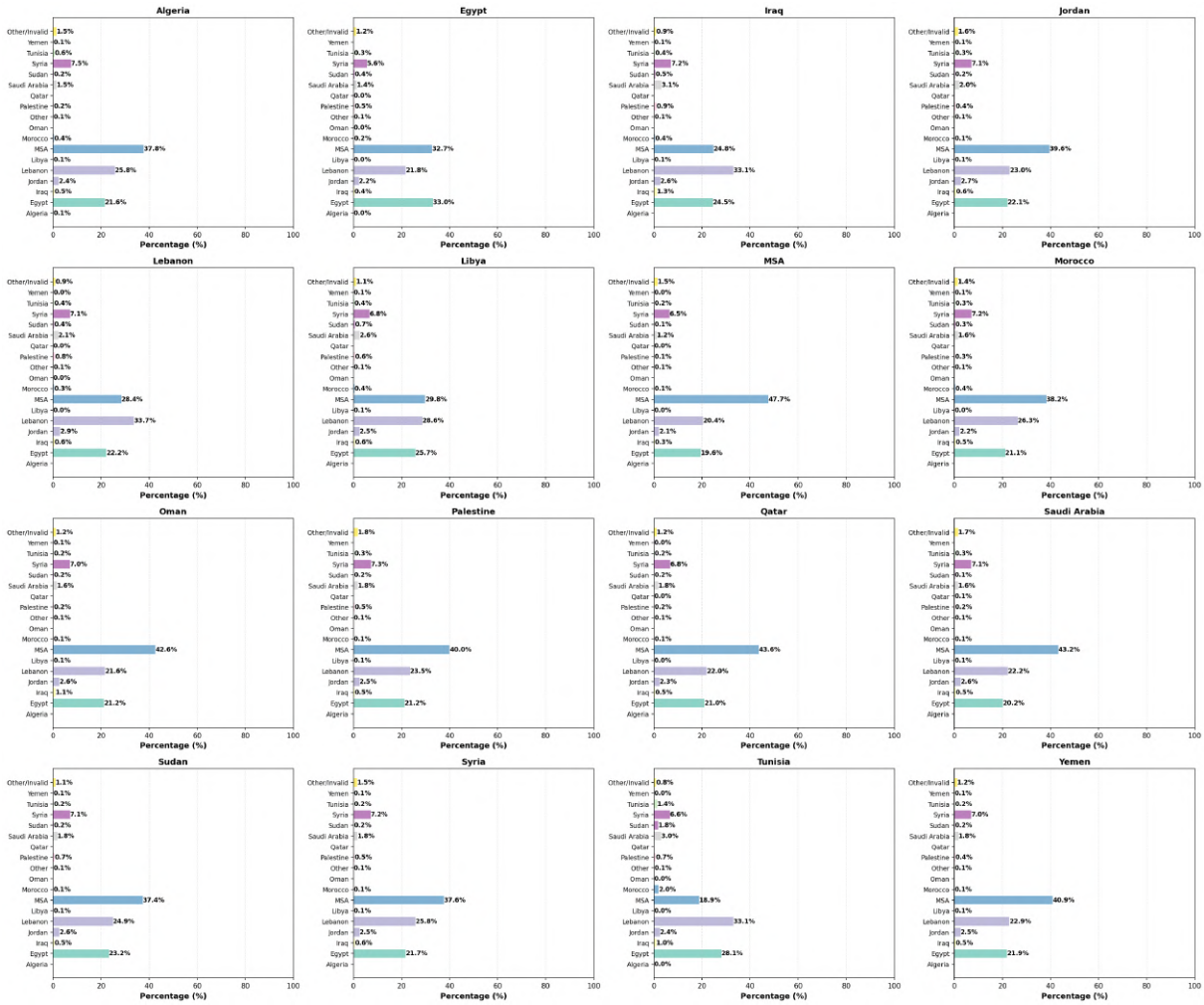

Figure 30: Distribution of dialects in translations produced by `EuroLLM-9B-Instruct` using the dialect-specific prompt. The dialects are classified on the sentence level, using `gemma-3-27b-it`.

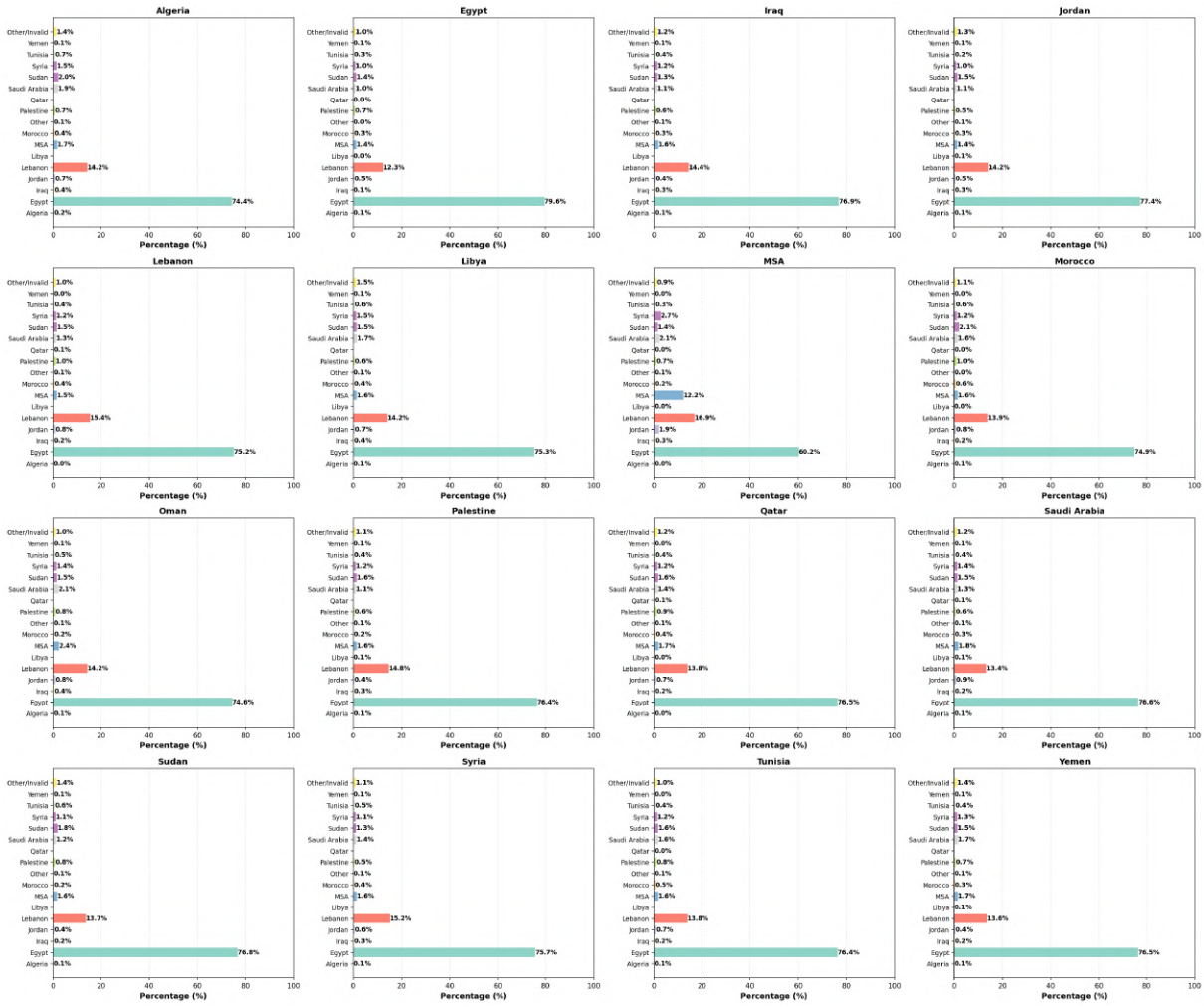

Figure 31: Distribution of dialects in translations produced by `Nile-Chat-12B` using the dialect-specific prompt. The dialects are classified on the sentence level, using `gemma-3-27b-it`.