# ALADAN at IWSLT25 Low-resource Arabic Dialectal Speech Translation Task

Josef Jon[2,4], Waad Ben Kheder[1], André Beyer[3], Claude Barras[1], and Jean-Luc Gauvain[1]

[1]Vocapia Research, France
[2]Lingea, Czechia
[3]Crowdee, Germany
[4]Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Czechia

## Abstract

We present our IWSLT 2025 submission for the low-resource track on North Levantine Arabic to English speech translation, building on our IWSLT 2024 efforts. We retain last year's cascade ASR architecture that combines a TDNN-F model and a Zipformer for the ASR step. We upgrade the Zipformer to the Zipformer-Large variant (253 M parameters vs. 66 M) to capture richer acoustic representations. For the MT part, to further alleviate data sparsity, we created a crowd-sourced parallel corpus covering five major Arabic dialects (Tunisian, Levantine, Moroccan, Algerian, Egyptian) curated via rigorous qualification and filtering. We show that using crowd-sourced data is feasible in low-resource scenarios as we observe improved automatic evaluation metrics across all dialects. We also experimented with the dataset under a high-resource scenario, where we had access to a large, high-quality Levantine Arabic corpus from LDC. In this setting, adding the crowd-sourced data does not improve the scores on the official validation set anymore. Our final submission scores 20.0 BLEU on the official test set.

## 1 Introduction

Dialectal Arabic speech translation (ST) remains one of the most challenging tasks in spoken language processing due to (i) the scarcity of high-quality, parallel speech–text resources for non-standardized varieties, (ii) high phonetic and orthographic variability among dialects, and (iii) domain mismatches between available corpora (e.g., broadcasts in Modern Standard Arabic) and conversational speech. Although end-to-end models and pre-trained encoders have advanced general ASR and NMT, most publicly available data still target Modern Standard Arabic (Al-Fetyani et al., 2023; Ali et al., 2016), leaving dialectal variants under-resourced. Previous IWSLT evaluations (Yan et al., 2022; Anastasopoulos et al., 2022; Agarwal et al., 2023; Hussein et al., 2023; Boito et al., 2022; Ahmad et al., 2024) have tackled these issues using transfer learning and fine-tuning strategies, yet a comprehensive solution for multiple dialects is still lacking.

In our IWSLT 2024 submission, ALADAN achieved first place in the Levantine Arabic task by combining a cascade ASR pipeline (TDNN-F + Zipformer) with fine-tuned NLLB and prompt-driven LLMs (Command-R), leveraging a crowd-sourced parallel corpus for Tunisian and Levantine Arabic. We also demonstrated that prudent data normalization and a hybrid system combination (ROVER) yield substantial WER and BLEU improvements. Building on this success, our IWSLT 2025 system introduces two key innovations:

- **Multi-Dialect ASR with Zipformer-Large:** We replace last year's Zipformer (66M parameters) with the 253M-parameter Zipformer-Large to better model long-range dependencies and acoustic nuances in dialectal speech. We also train a single multi-dialect model instead of deriving dialect-specific ASR models via fine-tuning.
- **Expanded multi-dialect crowd-sourcing:** We extend our crowd-sourced collection beyond Tunisian and Levantine to include Moroccan, Algerian, and Egyptian dialects, yielding more than 160k new parallel sentences after rigorous quality control. These data are used to fine-tune the NLLB-200 and Cohere Command-R models under QLoRA, enhancing cross-dialect robustness.

Our paper is organized as follows. Section 2 details the data collection and normalization procedures. Section 3 presents our ASR and ST models, detailing their architecture, training, fine-tuning, and performance on Levantine datasets and internal tests. In Section 4, we conclude with a discussion of future directions for low-resource dialect

*Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 252–259
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

translation.

## 2 Methods

### 2.1 Text normalization

The absence of standardized conventions across different Arabic dialects requires the development of robust text normalization procedures to reduce ambiguity. In this work, we adopt the same text normalization methodology used in Ben Kheder et al. (2024). Our normalization process operates on the character- and word-level. Character-level normalization promotes uniformity in the orthographic representation of various dialects, improving consistency across datasets. Table 1 summarizes the rules used in our experiments.

| **Dialect** | **Normalizations** |
|---|---|
| All dialects | ژ => ر / پ => ب |
| apc/arz/ary | ڤ or ڥ => ف |
| aeb/arq | ڤ => ق / ڥ => ف |
| ary | ڭ => ق / گ => ق |

Table 1: Characters normalization rules for different Arabic dialects.

Word-level normalization, on the other hand, addresses orthographic variability in dialectal Arabic and foreign words. This step employs rules derived from a combination of a Word2vec model and a weighted Levenshtein distance to identify orthographically similar words appearing in comparable contexts. This process helps normalize clusters of words such as:

- The Tunisian word for "anyway": حاسيلو، حاصلو، حاسيول، حاصولو، حاصيله، حاصلو، حاصيلو.
- The Syrian word for "the computer": الكويبوتير، الكمبيوتر.

For further details on the methodology, we refer readers to (Ben Kheder et al., 2024).

### 2.2 Crowd-sourcing for parallel data collection

We collaborate with Crowdee[1] crowd-sourcing platform to create a parallel dataset. The goal was to generate high-quality translations while addressing the challenges posed by dialectal variations in Arabic. In these tasks, transcripts from CTS/YouTube datasets (described in Ben Kheder et al. (2024)) are used as input.

#### 2.2.1 Crowd worker filtering

We designed linguistic assessment comprising 40 questions for each of the five dialects. The test evaluates linguistic competencies, including grammar proficiency and the ability to translate between the respective dialects and English using multiple choice exercises. Only workers who demonstrated sufficient linguistic skills were allowed to contribute to the dataset.

#### 2.2.2 Translation guidelines

To ensure the quality and completeness of translations, crowd workers receive detailed instructions:

- **Providing a translation:** Translate a single sentence from a short conversational excerpt (6 consecutive sentences corresponding to 3 speaking turns between 2 speakers, extracted from our internal conversational telephone data).
- **Ranking confidence:** Workers were asked to rate their confidence in their translation as "correct", "unsure", or "incorrect".
- **Suggesting alternatives:** Workers were encouraged to offer alternative translations if possible.
- **Adding comments:** Additional comments were invited to clarify translation choices or highlight ambiguities.

#### 2.2.3 Data filtering

Following the translation phase, a data cleaning procedure was implemented to improve the quality of the dataset. This included:

- **Removing machine-like translations:** Sentences with patterns indicative of machine-generated translations were excluded.
- **Language filtering:** Sentences that were in languages other than English (e.g., French or Arabic) were removed.
- **Word count discrepancy:** Examples with significant discrepancies in word count between the source and target were filtered out.
- **Perplexity-based filtering:** the GPT-2 model was used to compute the perplexity of each translated sentence. We removed all sentences that exceed 10 words with perplexity greater than 100, as these likely indicated low-quality translations.

[1] Crowdee—https://www.crowdee.de/

|  | **IWSLT24** | **IWSLT22** |  | **Internal devs (CTS)** |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | *valid apc* | *dev aeb* | *test$_1$ aeb* | *apc* | *arz* | *arq* | *ary* | *aeb* |
| **TDNN-F** | 26.5 | 39.9 | 40.8 | 19.8 | 26.4 | 28.7 | 30.7 | 27.6 |
| **Zipformer-Large** | 21.8 | 31.7 | 32.7 | 14.5 | 20.8 | 23.7 | 23.9 | 22.3 |
| **Both** | **19.9** | **30.6** | **31.8** | **14.0** | **19.3** | **22.1** | **22.8** | **21.6** |

Table 2: WER (%) of ASR models on IWSLT24 Levantine Arabic (apc) validation, IWSLT22 Tunisian Arabic (aeb) dev/test sets and 5 internal devs (apc, arz, arq, ary, aeb). The Levantine "apc", Egyptian "arz", Algerian "arq" internal devs correspond to telephone speech (CTS) while the ones for Moroccan "ary" and Tunisian Arabic "aeb" correspond to YouTube data (radio).

# 3 Experiments

This section describes our experimental settings, used data and results.

## 3.1 Data

In this subsection, we list the datasets we used for training and evaluating our systems.

### 3.1.1 ASR data

- **Training:** We used 4200h of multi-dialect multi-domain data to train our ASR models. For more details, readers may refer to (Ben Kheder et al., 2024).

- **Evaluation:** The models are evaluated on the dev sets from IWSLT22 (aeb) and IWSLT24 (apc). We conduct additional tests on internal devs corresponding to conversational telephone speech ("arq", "arz" and "apc" devs) and YouTube data ("aeb" and "arz").

### 3.1.2 NMT data

For the crowd-sourcing experiments, used the crowd-sourced datasets to finetune the NMT models and LLMs. The sizes of the datasets are listed in Table 3. For evaluation, we used a held-out part of the crowd-sourcing datasets, parts of the AraBench dataset (Sajjad et al., 2020) and the IWSLT 2024 test set from the dialectal Arabic shared task. For the final submission, we used the same datasets as our last year's submission (Ben Kheder et al., 2024), i.e. LDC2012T09, PADIC, MADAR, GlobalVoices, smaller crowd-sourced data, IWSLT22 Tunisian Arabic and the official training dataset for this task, provided by the organizers.

| **Dialect** | **Sentences (k)** |
|---|---|
| arq (Algerian) | 51.9 |
| arz (Egyptian) | 52.8 |
| ary (Moroccan) | 19.1 |
| apc (Levantine) | 14.8 |
| aeb (Tunisian) | 22.7 |

Table 3: NMT crowd-sourced dataset sizes.

## 3.2 Metrics

We score ASR using word error rate (WER). To measure the quality of the MT, we use 3 metrics: BLEU (Papineni et al., 2002), ChrF (Popović, 2015) and XCOMET-XL (Guerreiro et al., 2024).

## 3.3 ASR

### 3.3.1 ASR models

Our ASR front-end follows the cascade design of (Ben Kheder et al., 2024), combining a conventional TDNN-F model with an end-to-end Zipformer. The key innovation in this year's submission is the replacement of last year's 66 M-parameter Zipformer-M with a much larger, 253 M-parameter "Zipformer-Large" and the design of a single multi-dialect model (instead of deriving dialect-specific models via fine-tuning).

1. **TDNN-F model:** 15 layers of factorized TDNN with ReLU activations (layer dimension 1920) and linear bottlenecks (dimensions 320, 240) trained using the LF-MMI objective.

2. **Zipformer-Large:** The base design follows the "Zipformer-L" configuration of (Yao et al., 2023), modified as follows:

|  | **Configuration** |
|---|---|
| CNN kernel sizes | {63, 31, 15, 15, 31, 6} |
| Encoder hidden dim. | {192, 512, 1024, 1536, 512, 256} |
| Feed-forward dim. | {512, 768, 1024, 2048, 1024, 768} |

Table 4: Configuration of Zipformer-Large.

The output of the two models are combined using the ROVER algorithm.

### 3.3.2 Training procedure

We train multi-dialect models using all available data to take advantage of the acoustic and linguistic similarities between different Arabic dialects. The TDNN-F model is trained for 20 epochs (on all data) using lr=1e-3 and the Zipformer model is trained for 80 epochs using lr=4e-3.

### 3.3.3 Results

Table 2 shows the WERs of our ASR systems after applying the normalization procedure. This normalization significantly improved the WERs for "apc" and "aeb" by 10% and 18%, respectively. The combined model achieved even greater improvements, demonstrating the complementarity of the two models and outperforming all WERs reported in (Agarwal et al., 2023) for "aeb".

## 3.4 Speech translation

For speech translation (ST), we apply the cascaded approach: we use the ASR to obtain transcriptions and then we translate them using an NMT.

### 3.4.1 MT models

We finetune one pretrained NMT model (*NLLB-1.3B*) and 3 LLMs: Command-R V0.1 (4-bit quantized, CohereForAI/c4ai-command-r-v01-4bit), Aya Expanse 8B and EuroLLM 9B Instruct. We use QLoRA finetuning, using the transformers, peft and trl libraries. We set the LoRA rank size to $r = 32$ and $\alpha = 16$. We finetuned the models by AdamW optimizer, with warmup ratio of 0.03. We ran multiple training runs with learning rates $lr = \{2e-4, 1e-4, 5e-5, 1e-5\}$.

### 3.4.2 MT results

First, we compare the base and the finetuned models on the crowd-sourced test set (with reference transcriptions on the source side) and on the apc test set from IWSLT 2024 low-resource Arabic Dialectal Speech translation task. The results are presented in Table 5. We see that for all dialects, the evaluation scores improve significantly for all models. The best scoring finetuned model across all the dialects is the Command-R model, while all the other models are competitive. Of the base models, without finetuning, Command-R and Aya-expanse-8B provide the best scores. In particular, for the IWSLT test set (*apc/iwslt*), we obtain a large improvement in automated scores even though it is a different domain (interviews with refugees) from our crowd-sourced training data (telephone conversations).

We also compare the base and the finetuned Command-R and NLLB models on the part of the test sets in the AraBench dataset. The comparison is shown in Table 10 in Appendix B. Here, base NLLB performs the best on these test sets and finetuning decreases the performance for NLLB, but improves it for Command-R.

### 3.4.3 ST results

We evaluated both base and finetuned NLLB and Command-R models on ASR outputs from the crowd-sourced and IWSLT test sets (Tables 6 and 7). As with reference transcriptions, finetuning with crowd-sourced data significantly improves performance across dialects. Although the BLEU and ChrF scores are similar, Command-R consistently outperforms NLLB in XCOMET-XL. Our finetuned model scores 23.7 BLEU on IWSLT, compared to 28.7 for the top shared task system and 20.9 for the runner-up, despite those systems using much more fine-tuning data, including in-domain training set. This shows that crowd-sourcing is a viable option to improve automated metric scores for dialectal Arabic ST even on out-of-domain test sets.

## 3.5 Final submission

We also finetuned the Command-R model on the same datasets as our submission from last year (Ben Kheder et al., 2024). We note that, as opposed to the previously described experiments, we trained the models for document-level translation, with a maximum context size of 100 lines, same as last year. We experimented with adding the crowd-sourced data described earlier on top of these datasets and the results are shown in Table 8. Adding crowd-sourced data on top of an already large and high-quality dataset does not have any positive effect on BLEU and ChrF scores. For the final submission, we selected 29 checkpoints with best BLEU scores on the validation set, translated the test set with them and ran Minimum Bayes risk decoding using wmt22-comet-da score as the objective function. The final official results from IWSLT 2025 (Abdulmumin et al., 2025) are shown in Table 9.

|  |  | Base model |  |  | Finetuned model |  |  |
|---|---|---|---|---|---|---|---|
| **Model** | **Language** | **BLEU** | **ChrF** | **COMET** | **BLEU** | **ChrF** | **COMET** |
| NLLB | arq | 5.8 | 25.9 | 0.558 | 24.8 | 47.8 | 0.765 |
|  | arz | 14.8 | 36.9 | 0.586 | 31.8 | 52.9 | 0.807 |
|  | apc | 15.4 | 38.0 | 0.657 | 29.5 | 51.0 | 0.836 |
|  | ary | 16.2 | 39.7 | 0.528 | **33.5** | 55.4 | **0.726** |
|  | aeb | 17.3 | 40.2 | 0.549 | 30.9 | 53.2 | 0.741 |
|  | iwslt24/apc | 19.2 | 44.6 | 0.689 |  |  |  |
| Command-R-V0.1-4bit | arq | 12.6 | 34.3 | 0.573 | **28.8** | **49.2** | **0.778** |
|  | arz | 18.0 | 42.1 | 0.624 | **33.2** | **53.6** | **0.820** |
|  | apc | 20.1 | 42.4 | 0.661 | **35.4** | **55.7** | **0.856** |
|  | ary | 20.2 | 44.5 | 0.596 | 33.3 | 54.7 | 0.718 |
|  | aeb | 20.3 | 44.8 | 0.613 | **31.0** | **53.4** | **0.759** |
|  | iwslt24/apc | 19.7 | 45.9 | 0.818 | 28.2 | 53.4 | 0.848 |
| EuroLLM-9B | arq | 9.8 | 31.3 | 0.519 | 27.0 | 48.2 | 0.773 |
|  | arz | 21.1 | 44.5 | 0.611 | 32.3 | 52.7 | 0.805 |
|  | apc | 16.0 | 40.2 | 0.599 | 31.8 | 52.6 | 0.839 |
|  | ary | 19.5 | 45.1 | 0.514 | 31.7 | 53.7 | 0.712 |
|  | aeb | 21.1 | 45.6 | 0.578 | 29.7 | 52.2 | 0.750 |
| Aya-expanse-8b | arq | 13.4 | 34.7 | 0.557 | 26.8 | 47.8 | 0.775 |
|  | arz | 24.8 | 47.3 | 0.632 | 32.8 | 53.1 | 0.815 |
|  | apc | 21.8 | 43.7 | 0.644 | 33.0 | 53.1 | 0.845 |
|  | ary | 24.8 | 48.4 | 0.568 | 31.8 | 53.1 | 0.707 |
|  | aeb | 23.4 | 47.9 | 0.607 | 29.3 | 51.7 | 0.747 |

Table 5: Results of base and finetuned models on our test sets and IWSLT 2024 test set in text-to-text translation (using reference transcriptions of the source speech as the source for the MT).

|  | **Language** | **BLEU** | **ChrF** | **COMET** |
|---|---|---|---|---|
| **NLLB** | arq | 6.0 | 23.7 | 0.567 |
|  | arz | 13.3 | 34.9 | 0.609 |
|  | apc | 13.6 | 34.3 | 0.663 |
|  | ary | 14.2 | 36.1 | 0.522 |
|  | aeb | 13.6 | 35.2 | 0.504 |
| **Command-R** | arq | 7.0 | 27.0 | 0.569 |
|  | arz | 15.5 | 36.9 | 0.607 |
|  | apc | 17.1 | 38.6 | 0.671 |
|  | ary | 17.3 | 40.4 | 0.563 |
|  | aeb | 18.0 | 40.8 | 0.559 |
|  | iwslt24/apc | 16.5 | 42.1 | 0.766 |

Table 6: Cascaded speech translation scores of **base**, non-finetuned models on our test sets (using our ASR transcriptions of the source speech).

|  | **Language** | **BLEU** | **ChrF** | **COMET** |
|---|---|---|---|---|
| **NLLB** | arq | 21.2 | 40.7 | 0.731 |
|  | arz | 26.2 | 46.3 | 0.757 |
|  | apc | 24.2 | 44.2 | 0.790 |
|  | ary | 25.7 | 47.8 | 0.643 |
|  | aeb | 23.2 | 45.4 | 0.646 |
| **Command-R** | arq | 21.7 | 41.6 | 0.741 |
|  | arz | 26.0 | 46.7 | 0.767 |
|  | apc | 27.2 | 47.6 | 0.805 |
|  | ary | 25.4 | 47.1 | 0.652 |
|  | aeb | 23.2 | 45.4 | 0.673 |
|  | iwslt24/apc | 23.7 | 48.6 | 0.803 |

Table 7: Cascaded speech translation scores of **finetuned** models on our test sets.

## 4 Conclusions

In this work, we demonstrated that carefully engineered data collection and model adaptation can substantially advance low-resource dialectal Arabic speech translation. By expanding our crowd-sourced parallel corpus to five dialects (Tunisian, Levantine, Moroccan, Algerian, Egyptian), including rigorous qualification tests and multi-stage filtering, we provided rich, targeted material for NMT fine-tuning. Upgrading our acoustic front-end to a 253 M-parameter Zipformer-Large and combining it with TDNN-F via ROVER further drove down WER. On the translation side, fine-tuning NLLB-200 and Command-R models, with

|  |  | **Valid 2024** | **Test 2024** |
|---|---|---|---|
| **2024 dataset** | 2024 ASR | 30.3/53.5 | 27.5/50.6 |
|  | 2025 ASR | 31.4/54.7 | 27.4/50.3 |
|  | Human | 33.5/58.5 | - |
| **+new crowd** | 2024 ASR | 29.9/53.3 | 27.4/50.3 |
|  | 2025 ASR | 31.1/54.6 | 27.2/50.2 |
|  | Human | 33.6/58.7 | - |
| **Final MBR** |  | 32.5/55.6 | 28.0/51.7 |

Table 8: BLEU/ChrF scores of document-level models trained on our last year's dataset and after adding the new crowd-sourced dataset described above. We also compared using our last year's ASR model with this year's improved model and to the human reference transcription.

| Submission Name | BLEU | COMET | CHRF |
|---|---|---|---|
| AIB_Marco contrastive1 | 15.82 | 0.6456 | 36.23 |
| AIB_Marco contrastive2 | 10.53 | 0.5727 | 27.69 |
| AIB_Marco contrastive3 | 16.22 | 0.6669 | 37.48 |
| AIB_Marco contrastive4 | 16.47 | 0.683 | 37.96 |
| AIB_Marco primary | 12.01 | 0.6547 | 34.19 |
| **ALADAN primary** | **20.02** | **0.6613** | **39.91** |
| jhu contrastive1 | 15.39 | 0.6569 | 35.91 |
| jhu primary | 14.64 | 0.6493 | 36.23 |
| lia contrastive1 | 21.02 | 0.6983 | 42.92 |
| lia contrastive2 | 21.45 | 0.694 | 43.13 |
| lia primary | 22.56 | 0.7193 | 44.72 |
| kit contrastive1 | 19.11 | 0.6832 | 40.95 |
| kit contrastive2 | 21.93 | 0.6968 | 44.67 |
| kit primary | 23.34 | 0.7043 | 45.09 |

Table 9: The official results of the Levantine Arabic task from IWSLT 2025. Our submission in bold.

QLoRA for the latter, on this multi-dialect dataset yielded significant BLEU and COMET gains on our in-domain test sets. These findings confirm that combining expanded crowd-sourcing with unsupervised data augmentation and model scaling is a viable and resource-efficient strategy to boost dialectal Arabic translation, even when faced with new domains. However, our experiments with the final submission show that adding this dataset on top of already extensive, high-quality corpora we used to train our last year's submission does not improve BLEU and ChrF scores on the official validation set. This suggests that the crowdsourcing approach is more viable in low-resource scenarios, as the knowledge provided by the crowdsourced dataset might already be covered in the larger corpora.

## Acknowledgments

This work was funded by the the European Defence Fund (EDF) 2021 project ALADAN (Ai-based LAnguage technology development framework for Defence ApplicatioNs; Grant ID: 101102545). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them. It was also partially supported by the Charles University Grant Agency in Prague (GAUK 244523) and by SVV project number 260 821. We would like to express our gratitude to LDC for kindly providing the data used in this study.

## References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Ashwin, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Marco Gaido, Dávid Javorský, Marek Kasztelnik, Tsz Kin Lam, Danni Liu, Evgeny Matusov, Chandresh Kumar Maurya, John P. McCrae, Salima Mdhaffar, Yasmin Moslem, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Atul Kr. Ojha, John E. Ortega, Sara Papi, Pavel Pecina, Peter Polák, Piotr Połeć, Beatrice Savoldi, Nivedita Sethiya, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Marco Turchi, Alex Waibel, Patrick Wilken, Rodolfo Zevallos, Vilém Zouhar, and Maike Züfle. 2025. Findings of the iwslt 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, Vienna, Austia (in-person and online). Association for Computational Linguistics. To appear.

Milind Agarwal, Sweta Agarwal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, et al. 2023. Findings of the iwslt 2023 evaluation campaign. 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, et al. 2024. Findings of the iwslt 2024 evaluation campaign. *arXiv preprint arXiv:2411.05088*.

Mohammad Al-Fetyani, Muhammad Al-Barham, Gheith Abandah, Adham Alsharkawi, and Maha Dawas. 2023. Masc: Massive arabic speech corpus. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1006--1013. IEEE.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 279--284. IEEE.

Antonios Anastasopoulos, Loc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, et al. 2022. Findings of the iwslt 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98--157. Association for Computational Linguistics.

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.

Marcely Zanon Boito, John Ortega, Hugo Riguidel, Antoine Laurent, Loïc Barrault, Fethi Bougares, Firas Chaabani, Ha Nguyen, Florentin Barbier, Souhir Gahbiche, et al. 2022. On-trac consortium systems for the iwslt 2022 dialect and low-resource speech translation tasks. *arXiv preprint arXiv:2205.01987*.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979--995.

Amir Hussein, Cihan Xiao, Neha Verma, Thomas Thebaud, Matthew Wiesner, and Sanjeev Khudanpur. 2023. Jhu iwslt 2023 dialect speech translation system description. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 283--290.

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.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. Arabench: Benchmarking dialectal arabic-english machine translation. In *GOLING*, pages 123--456.

Brian Yan, Patrick Fernandes, Siddharth Dalmia, Jiatong Shi, Yifan Peng, Dan Berrebbi, Xinyi Wang, Graham Neubig, and Shinji Watanabe. 2022. Cmu's iwslt 2022 dialect speech translation system. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 298--307.

Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey. 2023. Zipformer: A faster and better encoder for automatic speech recognition. *arXiv preprint arXiv:2310.11230*.

# A Document-level translation

We also compared line-by-line translation to translating the whole conversation for the crowd-sourced dataset. We used the same document-level prompt as (Ben Kheder et al., 2024). Surprisingly, in this case translating line by line worked better. We hypothesize that the repetitiveness of the dataset causes this. Many simple utterances (e.g. "Yeah.") are repeated next to each other in the training data, which leads the model to overestimate the probability of repeating the same line in the document-level translation. We leave a better understanding of this issue for future work.

# B AraBench test set

We also evaluated our model on test sets from the AraBench (Sajjad et al., 2020) benchmark, specifically the MADAR (Bouamor et al., 2018) test sets for dialects we used during finetuning. Contrary to the results on other test sets, the base NLLB model scores the best, and fine-tuning on our crowd-sourced data hurts the evaluation scores. For Command-R, finetuning improves the scores compared to the Command-R base model, but still does not outperform base NLLB. We hypothesize that this might be caused by presence of the test set in the NLLB's training dataset, or by domain mismatch.

|  |  | Base |  |  | Finetuned |  |  |
|---|---|---|---|---|---|---|---|
|  |  | **BLEU** | **ChrF** | **COMET** | **BLEU** | **ChrF** | **COMET** |
| **NLLB** | madar.test.lev.0.jo.ar | 47.3 | 65.2 | 0.924 | 42.3 | 61.5 | 0.938 |
|  | madar.test.lev.0.lb.ar | 42.1 | 59.5 | 0.863 | 36.3 | 54.9 | 0.884 |
|  | madar.test.lev.0.pa.ar | 45.6 | 62.4 | 0.911 | 41.9 | 59.5 | 0.928 |
|  | madar.test.lev.0.sy.ar | 45.9 | 62.9 | 0.911 | 41.5 | 59.4 | 0.926 |
|  | madar.test.lev.1.jo.ar | 47.4 | 64.4 | 0.876 | 43.7 | 61.4 | 0.930 |
|  | madar.test.lev.1.sy.ar | 47.1 | 63.9 | 0.899 | 41.3 | 59.7 | 0.909 |
|  | madar.test.mgr.0.ma.ar | 41.3 | 59.6 | 0.859 | 37.8 | 56.8 | 0.891 |
|  | madar.test.mgr.0.tn.ar | 36.4 | 54.4 | 0.818 | 31.7 | 50.5 | 0.845 |
|  | madar.test.mgr.1.ma.ar | 47.4 | 65.2 | 0.889 | 44.7 | 62.9 | 0.915 |
|  | madar.test.mgr.1.tn.ar | 30.6 | 49.4 | 0.815 | 27.0 | 46.4 | 0.844 |
|  | madar.test.nil.0.eg.ar | 45.7 | 63.1 | 0.904 | 41.3 | 59.8 | 0.931 |
|  | madar.test.nil.1.eg.ar | 52.8 | 68.8 | 0.926 | 50.2 | 66.5 | 0.944 |
| **Command-R** | madar.test.lev.0.jo.ar | 39.3 | 59.3 | 0.943 | 42.8 | 61.1 | 0.946 |
|  | madar.test.lev.0.lb.ar | 29.9 | 50.6 | 0.876 | 33.4 | 53.0 | 0.892 |
|  | madar.test.lev.0.pa.ar | 39.4 | 58.2 | 0.934 | 43.9 | 60.4 | 0.942 |
|  | madar.test.lev.0.sy.ar | 39.0 | 58.3 | 0.931 | 42.1 | 59.5 | 0.941 |
|  | madar.test.lev.1.jo.ar | 40.5 | 59.5 | 0.929 | 43.8 | 61.6 | 0.938 |
|  | madar.test.lev.1.sy.ar | 39.6 | 58.6 | 0.918 | 42.9 | 60.5 | 0.928 |
|  | madar.test.mgr.0.ma.ar | 33.7 | 54.4 | 0.885 | 37.1 | 56.0 | 0.895 |
|  | madar.test.mgr.0.tn.ar | 22.2 | 42.9 | 0.792 | 29.5 | 48.2 | 0.847 |
|  | madar.test.mgr.1.ma.ar | 38.4 | 59.3 | 0.909 | 42.7 | 60.7 | 0.916 |
|  | madar.test.mgr.1.tn.ar | 20.0 | 40.8 | 0.805 | 25.3 | 44.5 | 0.850 |
|  | madar.test.nil.0.eg.ar | 40.1 | 59.4 | 0.935 | 43.0 | 60.9 | 0.942 |
|  | madar.test.nil.1.eg.ar | 45.5 | 63.5 | 0.943 | 50.2 | 66.5 | 0.952 |

Table 10: Automatic evaluation scores of base and finetuned models on MADAR test sets from the AraBench benchmark.