Curriculum Learning and Pseudo-Labeling Improve the Generalization of Multi-Label Arabic Dialect Identification Models

Ali Mekky* **Mohamed El Zeftawy*** **Lara Hassan*** **Amr Keleg** **Preslav Nakov**
Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
{firstname.lastname}@mbzuai.ac.ae

Abstract

Being modeled as a single-label classification task for a long time, recent work has argued that Arabic Dialect Identification (ADI) should be framed as a multi-label classification task. However, ADI remains constrained by the availability of single-label datasets, with no large-scale multi-label resources available for training. By analyzing models trained on single-label ADI data, we show that the main difficulty in repurposing such datasets for Multi-Label Arabic Dialect Identification (MLADI) lies in the selection of negative samples, as many sentences treated as negative could be acceptable in multiple dialects. To address these issues, we construct a multi-label dataset by generating automatic multi-label annotations using GPT-4o and binary dialect acceptability classifiers, with aggregation guided by the Arabic Level of Dialectness (ALDi). Afterward, we train a BERT-based multi-label classifier using curriculum learning strategies aligned with dialectal complexity and label cardinality. On the MLADI leaderboard, our best-performing LahjatBERT model achieves a macro F1 of 0.69, compared to 0.55 for the strongest previously reported system. Code and data are available at `https://mohamedalaa9.github.io/lahjatbert/`.

1 Introduction

Arabic has a wide range of diverse dialects spoken across the Arab World. While Modern Standard Arabic (MSA) is used in official communication, education, and media, everyday conversations typically happen in local dialects (Habash, 2010). Dialects vary between countries, and can even vary within same-country cities and local communities, creating a complex linguistic landscape (Zaidan and Callison-Burch, 2014; Althobaiti, 2020).

Arabic Dialect Identification (ADI) is the task that aims to identify the dialect of a sentence. ADI has been modeled as a single-label classification problem, with systems trained and evaluated on resources such as AOC (Zaidan and Callison-Burch, 2011), MADAR (Bouamor et al., 2019), QADI (Abdelali et al., 2021), and NADI benchmarks (Abdul-Mageed et al., 2020, 2021b, 2022, 2023). However, as illustrated in Table 1, a single utterance may simultaneously sound natural to speakers from multiple countries, making it inherently multi-dialectal and difficult for traditional single-label classification methods to produce accurate results (Keleg and Magdy, 2023; Olsen et al., 2023). Even moderately long sentences can be acceptable in multiple dialects (Keleg et al., 2025).

Dialects	**Sentence**
Jordan, Palestine	اتفقنا نظل جنب بعض ع الحلوه والمره (We agreed to stick by each other through thick and thin.)
Algeria, Egypt, Jordan, Palestine, Sudan, Syria, Tunisia, Yemen	احسبو حسابي معاكم (Make sure to count me in with you.)
Algeria, Palestine, Sudan, Yemen	بسمتك يا زين تسوي ألف بسمه (Your smile, oh beautiful one, is worth a thousand other smiles.)
All Arabic Dialects	اللهم الجنه لمن ذهبت أرواحهم إليك (O Allah, grant Paradise to those whose souls have returned to You.)

Table 1: Examples illustrating dialect overlap in Arabic, sampled from the manually annotated NADI2024 Subtask 1 development set. **Note:** the dataset has labels for eight country-level dialects only.

To better capture this phenomenon, the field has begun transitioning to *Multi-Label Arabic Dialect Identification* (MLADI), which allows an utterance to be tagged as acceptable in multiple dialects.[1] This transition has been driven by the evolution of the NADI shared tasks, a recurring benchmark se-

* Equal contribution.

[1] Multi-label Dialect Identification is being explored for other languages such as French and Spanish (Bernier-colborne et al., 2023; Zampieri et al., 2024; Lopetegui et al., 2025).

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 261–274

ries for Arabic Dialect Identification. While earlier editions of NADI (2020–2023) framed ADI as a single-label classification problem, assigning each sentence to a single country-level dialect (Abdul-Mageed et al., 2020, 2021b, 2023), the `NADI2024` shared task introduced multi-label annotations to explicitly account for dialectal overlap (Abdul-Mageed et al., 2024). However, in `NADI2024`, multi-label annotations are provided only for the development and test sets, while the training data remains single-labeled, creating a mismatch between the nature of the task and the structure of the training data. As a result, MLADI introduces a weakly-supervised classification problem, where models must learn to predict multiple dialects based on datasets that provide only one ground-truth label per instance.

In this work, we examine the mismatch between the multi-label nature of Arabic dialect usage and the single-label structure of existing ADI datasets by analyzing the behavior of binary dialect-specific acceptability classifiers trained on single-label NADI datasets (Abdul-Mageed et al., 2020, 2021b, 2023). Using the classifiers' training dynamics (Swayamdipta et al., 2020), we show that a substantial portion of samples treated as negative supervision is in fact judged acceptable by native speakers. This observation highlights the difficulty of accurately selecting negative samples when repurposing single-label, geo-located datasets for multi-label dialect identification and motivates explicit multi-label supervision.

Building on these insights, we construct a multi-label training set by aggregating pseudo-labels from two heterogeneous sources: GPT-4o and a set of 18 binary dialect acceptability classifiers. Using this dataset, we train a BERT-based multi-label classifier, which we refer to as LAHJATBERT. We train it under three settings: without curriculum learning, with an ALDi-aware curriculum, and with a label-cardinality-based curriculum (Bengio et al., 2009). The two curriculum strategies progressively expose the model to sentences of increasing dialectal ambiguity, allowing it to learn from simpler instances before confronting harder ones.

Our contributions are threefold:

1. We provide an in-depth analysis of the limitations of reusing single-label ADI datasets for multi-label dialect acceptability.
2. We construct a pseudo-labeled multi-label dataset for MLADI by combining the predictions of GPT-4o and binary dialect classifiers.[2]
3. We introduce LAHJATBERT, a family of multi-label BERT-based models trained on the constructed dataset, and investigate curriculum-based training variants aligned with the multi-label structure of MLADI. The best-performing variant achieves 69.04% macro-F1 on the MLADI leaderboard (Keleg et al., 2025), surpassing the top `NADI2024` system and outperforming larger multilingual and Arabic-specific language models.

2 MLADI Task's Setup and Previous Attempts

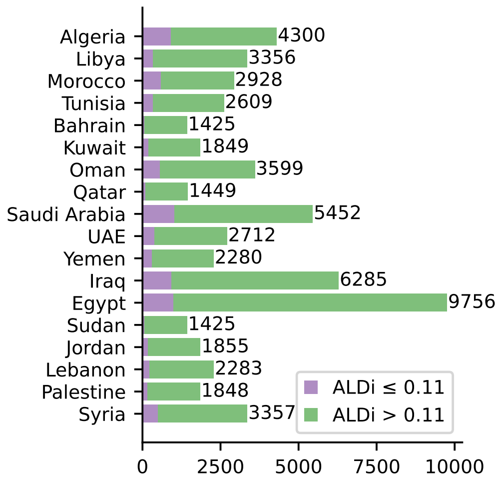

Figure 1: Number of samples in each dialect after combining the NADI 2020, 2021, 2023 datasets. Samples with automatically estimated *Arabic Level of Dialectness* (ALDi; Keleg et al., 2023) ≤ 0.11 are expected to be in MSA. The majority of the MSA samples are expected to be acceptable in all dialects.

The MLADI dataset only provides a development set of 120 samples and a test set of 1000 samples. Each sample of the development set is manually labeled by annotators from 8 countries, while the test set has 11 country-level acceptability labels. To build multi-label ADI systems, the task requires using the following three single-labeled NADI datasets: NADI 2020 (Abdul-Mageed et al., 2020), NADI 2021 (Abdul-Mageed et al., 2021b), and NADI 2023 (Abdul-Mageed et al., 2023). All these datasets provide tweets annotated with their estimated geo-location country label covering 18 countries: Algeria, Bahrain, Egypt, Iraq, Jordan,

[2]Following the NADI shared-task license, we release only tweet IDs and derived labels, not the underlying tweet text.

Kuwait, Lebanon, Libya, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Sudan, Syria, Tunisia, the UAE, and Yemen. Since these datasets rely on the user geo-location rather than manual linguistic annotation, posts authored by users whose geo-location differs from their country of origin can be mislabeled (Abdul-Mageed et al., 2024). NADI 2023 contains 1,000 tweets for each of 18 Arabic country-level dialects. By contrast, NADI 2020 and NADI 2021 are more imbalanced, with dialects such as Bahraini and Qatari underrepresented relative to more frequent varieties like Egyptian and Iraqi. The combined distribution of samples across dialects is shown in Figure 1.

Previous MLADI Attempts. Kanjirangat et al. (2024) used a nearest-neighbour approach to predict multiple labels for each sample, by encoding the training data samples and the test set into the same embedding space. Karoui et al. (2024) tackled the weak supervision problem by applying a similarity-based label expansion strategy. Their method, SIMMT, heuristically assigns additional labels to each sample based on vocabulary similarity between dialects, followed by multi-label fine-tuning of transformer models. This approach demonstrated that moving from single-label to multi-label supervision can improve performance on the MLADI task. Our work replaces heuristic label expansion with a pseudo-labeling framework that integrates complementary signals from multiple models to produce multi-label annotations. This yields a richer supervision signal than surface-level similarity measures.

3 Difficulties of Using Existing Datasets for Dialect Acceptability Classification

An intuitive idea to build multi-label ADI systems is to repurpose single-label ADI datasets for training multiple independent binary classifiers, each of which assesses the acceptability of a sentence in a specific dialect. For a specific country-level dialect (*Cntry*), one might take samples geolocated to *Cntry* as **positive** (i.e., *acceptable*) samples for the country's classifier, and samples geolocated to other countries as **negatives**. Previous attempts have shown that this technique does not achieve the best classification performance (Karoui et al., 2024; Kanjirangat et al., 2024). However, they have not analyzed the reasons for the failure of this technique, which we investigate here.

3.1 Other Countries' Samples Are Not Always Negative Samples

Definition For a dialect acceptability model, the negative class represents *sentences with **any linguistic feature** (e.g., a morpheme or a lexical item) that is **not acceptable** in the considered dialect*. The positive class represents the remaining sentences acceptable in this dialect or in MSA.

The majority of the samples geo-located to the considered dialect are expected to be positive samples. The fact that some of these samples could also be acceptable in other dialects does not impact their categorization as positive samples, according to the aforementioned definition. To the contrary, considering samples geo-located to other dialects as negative samples is problematic. More specifically, a subset of these samples is expected to also be acceptable in the dialect considered. Consequently, this subset of negative sentences should be reassigned to the positive class. We next show how model training dynamics could guide the identification of wrongly assigned negative samples.[3]

3.2 Training Dynamics and Multi-label Samples

Swayamdipta et al. (2020) used three metrics tracked during the model training process to categorize each sample's difficulty to be learned. For a training sample $\mathrm{x_i}$ with a label $\mathrm{y_i}$, the confidence that the model assigns to the target $\mathrm{y_i}$ is tracked at the end of each training epoch (or every number '*N*' of training batches). The mean and standard deviation of the tracked confidence scores assigned to the sample's target label are termed the ***Confidence*** and ***Variability*** metrics, respectively. ***Correctness*** represents the percentage of epochs (or *N* batches) for which the model assigns a higher probability to the sample's target label than all the other labels. Based on these metrics, the training dataset is split into three main categories—*(1) easy to learn* samples: high-*Confidence* and low-*Variability*, *(2) ambiguous* samples: moderate-*Confidence* and high-*Variability*, and *(3) hard to learn*: low-*Confidence* and low-*Variability*.

Intuition A dialect acceptability classifier would struggle to learn negative samples that are also acceptable in the considered dialect, as these samples

[3] Abdul-Mageed et al. (2024) found that NADI's geo-location method has a moderate to high accuracy of ensuring a sentence's acceptability in a specific country-level dialect.

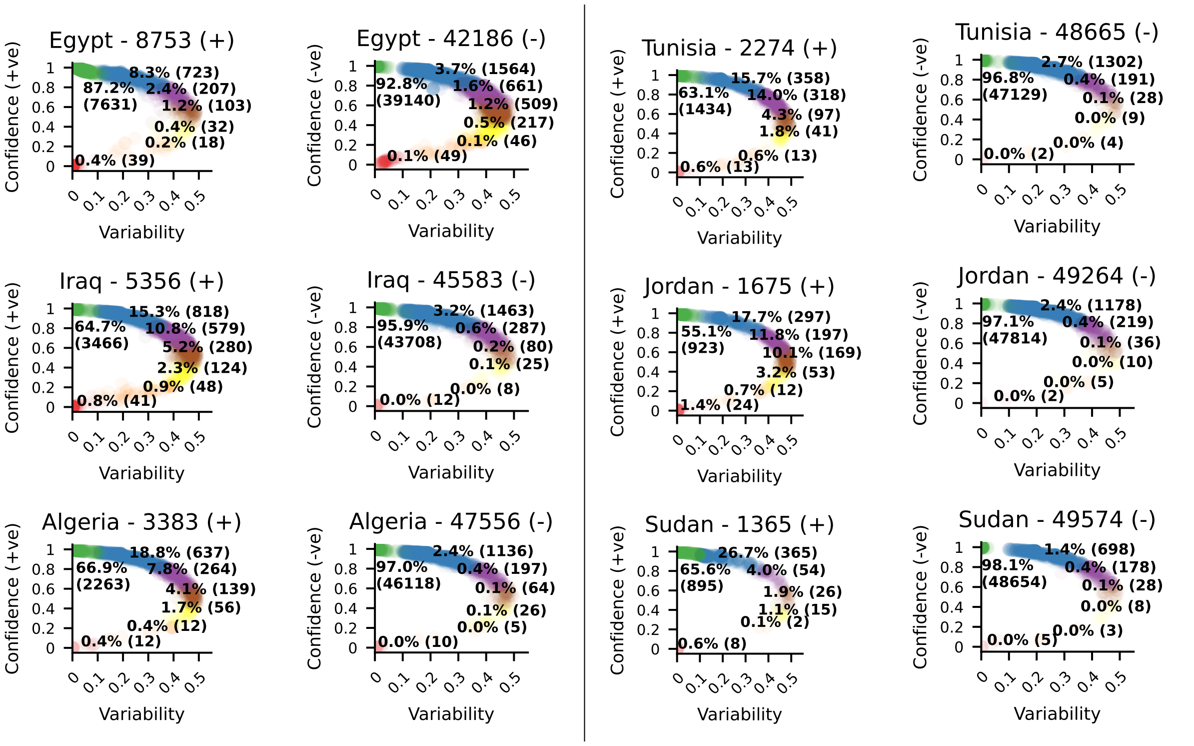

Figure 2: The training dynamics for 6 binary acceptability classifiers, characterized by the mean confidence in the label across different steps/stages of the model training (y-axis), and the standard deviation of these confidence values (x-axis). Each pair shows the training dynamics' metrics for the non-MSA positive (left) and negative (right) samples of a single classifier, with the respective number of samples shown above each subplot. **Note:** Sample's correctness ranges are ■: 0 ■:]0, 0.2[■: [0.2, 0.4[■: [0.4, 0.6[■: [0.6, 0.8[■: [0.8, 1[■: 1

should belong to the positive class. These samples are expected to be *ambiguous* or even *hard to learn*.

Methodology We train 18 dialect acceptability classifiers, one for each country represented in the NADI datasets. For each country's classifier, samples of the NADI datasets geo-located to this country, and MSA samples (ones with ALDi < 0.11) are considered as positive (acceptable) samples, with the remaining samples considered as negative (unacceptable) samples. Confidence scores are tracked every 300 steps ($\approx \frac{1}{5}$th of an epoch) for 5 epochs. However, the first epoch's confidence scores are ignored, as the model's training dynamics could be unstable during the early learning stages (Swayamdipta et al., 2020).

Findings Figure 2 shows the training dynamics for six different acceptability classifiers, for the positive (non-MSA) samples and the negative samples. First, a few samples (<50) for each set have a correctness value of 0. Moreover, the larger the imbalance between the number of positive and negative samples becomes, the smaller the percentage of negative samples with non-perfect correctness is. For instance, 7.2% (n=3,046) of Egypt's negative samples have non-perfect correctness scores compared to only 1.9% (n=920) for Sudan. These negative samples with non-perfect correctness scores are potentially wrongly assigned samples.

3.3 Usability of Training Dynamics in Flagging Wrongly Assigned Samples

To understand the effectiveness of training dynamics in identifying wrongly assigned samples, we engaged native speakers to manually evaluate the acceptability of sentences with varying correctness scores. Specifically, we recruited one annotator from the following countries: Egypt (EG), Iraq (IQ), Algeria (DZ), Tunisia (TN), Jordan (JO), and Sudan (SD).[4]

For each country, we sample 10 examples from each of the seven correctness ranges: *0,]0, 0.2[, [0.2, 0.4[, [0.4, 0.6], [0.6, 0.8[, [0.8, 1[, 1*. In total, 140 samples are annotated for each dialect (70 of which are positive and the remaining 70 are negative), unless one of the bins had fewer

[4]Notably, the six recruited annotators previously participated in the annotation of NADI 2024's evaluation sets.

than 10 samples. The annotators follow the same guidelines of NADI 2024's first subtask (Abdul-Mageed et al., 2024), whereby they answer: ***Is it possible that the tweet is authored by someone who speaks one of your country's dialects?*** *Options: Yes, Not Sure/Maybe, or No.*

Results Inspecting the number of acceptable samples in each bin in Table 2, a large number of negative samples with *correctness* < 1 are rated as acceptable, indicating the effectiveness of using non-perfect correctness scores to flag wrongly assigned negative samples. Conversely, a majority of the positive samples are rated as acceptable for the different correctness bins, even for correctness scores that are less than 0.6, which could be attributed to the class imbalance toward negative samples.

Label		Correctness 0	]0,0.2[	[0.2,0.4[	[0.4,0.6[	[0.6,0.8[	[0.8,1[	1
EG	+ve	2	6	6	6	7	9	10
	-ve	10	10	10	9	8	4	2
IQ	+ve	4	2	1	6	4	5	7
	-ve	9	6/8	10	9	7	5	2
DZ	+ve	3	4	5	5	8	7	9
	-ve	9	5/5	8	9	9	8	4
TN	+ve	1	3	1	0	1	5	7
	-ve	2/2	4/4	8/9	7	5	4	1
JO	+ve	3	7	6	5	8	10	9
	-ve	2/2	5/5	10	10	9	9	2
SD	+ve	5/8	1/2	6	7	10	8	10
	-ve	5/5	3/3	8/8	7	8	10	1

Table 2: The number of acceptable samples for the different correctness bins. For each row, each bin contains 10 total samples, except for a few bins marked by the *acceptable/total* format. **Note:** Values marked in blue indicate that >50% of the bin's samples are acceptable, while values marked in red indicate that <50% of the bin's samples are acceptable. The high-correctness bins of the positive class are expected to have a large number of acceptable samples. Conversely, the high-correctness bins of the negative class are expected to have a small number of acceptable samples, which is not the case. Country codes follow the ISO 3166-1 alpha-2 standard.

Moving Forward Our analysis shows that a large proportion of negative samples with non-perfect correctness scores should be reassigned to the positive class. However, manual annotation is still required to assess these automatically flagged samples. Moreover, the class imbalance seems to have an impact on the model's dynamics. Future work could consider using a *human and model in the loop* setup (Vidgen et al., 2021). More specifically, the initial model's training dynamics are used to automatically identify ambiguous samples, which are then manually reassigned to the correct class. Afterward, another model is trained from scratch on the dataset after reassignment, with the new model's dynamics used to automatically identify new ambiguous samples.

4 Multi-Label ADI Dataset Creation

It seems inevitable that building multi-label ADI models requires the presence of multi-label ADI training datasets, especially to reduce the number of false negatives for each dialect (i.e., samples wrongly assumed unacceptable in that dialect). However, building large enough multi-label ADI datasets is expensive, since annotating just 1,120 samples by speakers of 9 countries could cost as much as \$1,700 (Abdul-Mageed et al., 2024). Hence, we propose two pseudo-labeling methods for building multi-label ADI datasets:

(1) Binary Dialect Classifiers. We again build 18 independent acceptability classifiers, one per country-level dialect dia, each predicting whether a sentence x is acceptable in dialect dia. These classifiers are trained only on the balanced NADI 2023 dataset to avoid skewed pseudo-labels. For a sentence x_i, we generate 18 acceptability pseudo-labels using the 18 classifiers as:
$\hat{\mathbf{y}}^{\text{BIN}}_{x_i} = (Accept_{dia_1}(x_i), \ldots, Accept_{dia_{18}}(x_i))$.

Here, we notably rely on the Arabic Level of Dialectness (ALDi; Keleg et al., 2023) score as a global signal characterizing the degree of dialectness of each sentence ($a_i = \text{ALDi}(x_i) \in [0, 1]$). Keleg et al. (2025) showed that ALDi moderately correlates with the number of dialects in which a sentence is acceptable. The higher the ALDi score of a sentence, the less the number of dialects in which the sentence is acceptable.

For the **positive** samples of dia's acceptability classifier $Accept_{dia}(x_i)$, we consider all sentences from the NADI datasets that are geolocated to dia, in addition to MSA sentences with $a_i < 0.11$, which are broadly acceptable across dialects, irrespective of their geolocation. Since treating all sentences geolocated to other regions as negative examples leads to systematic errors (as shown in Section 3), we prioritize the precision of selecting true **negative samples**. To this end, we only select sentences that are (1) geolocated to *non-neighbouring* dialect regions, (2) with high dialect-

Pseudo-Labeling Method	P_{macro}	R_{macro}	$F1_{macro}$	Acc.
(1) Binary Dialect Classifiers	**78.2**	51.4	60.4	76.2
(2) GPT-4o Pseudo-Labels	73.5	**66.3**	67.8	77.4
Hybrid ALDi-Based Labels (§4.2)	77.5	62.8	**68.5**	**79.3**

Table 3: Macro-averaged Precision, Recall, F1, and Accuracy of the two methods in addition to the aggregation method (§4.2) on the NADI2024 development set.

ness ($a_i > 0.77$), where linguistic overlap with dialect dia is less likely.[5]

(2) GPT-Based Multi-Label Annotation. For each sentence, we obtain multi-label annotations from GPT-4o by prompting the model to independently assess the acceptability of the sentence in each of the 18 dialects. This yields a binary relevance vector $\hat{\mathbf{y}}_i^{\text{GPT}} \in \{0,1\}^{18}$. The full prompting template is provided in Appendix E.

4.1 Quality of Supervision Signals

To evaluate the two labeling methods, we use the development set of the NADI2024 shared task (Abdul-Mageed et al., 2024), which has 120 sentences annotated for 8 country-level dialects. This provides a useful benchmark for evaluating the performance of the two methods, as shown in Table 3.

Our conservative way of selecting the samples for the binary classifiers results in a high macro precision of 78.2 and a moderate macro recall of 51.4. In contrast, GPT-4o achieves substantially higher recall and a better overall F1 score. This hints that GPT-4o is more willing to assign multiple dialect labels, particularly in cases where dialectal overlap is present.

4.2 Pseudo-Labels Aggregation

To better understand the results in Table 3, we analyze performance across ALDi ranges in Table 4. For MSA samples ($[0, 0.11)$), both GPT-4o and the Binary Classifiers achieve perfect precision. For highly dialectal samples ($[0.77, 1]$), the Binary Classifiers attain higher precision, indicating more reliable negative supervision when strong dialectal cues are present.

In contrast, GPT-4o consistently achieves higher recall and F1 in the intermediate ALDi ranges ($[0.11, 0.77]$), where dialectal overlap is more common. This behavior is consistent with the conservative construction of the Binary Classifiers, whose

[5]See Appendices B,C and D for more implementation details.

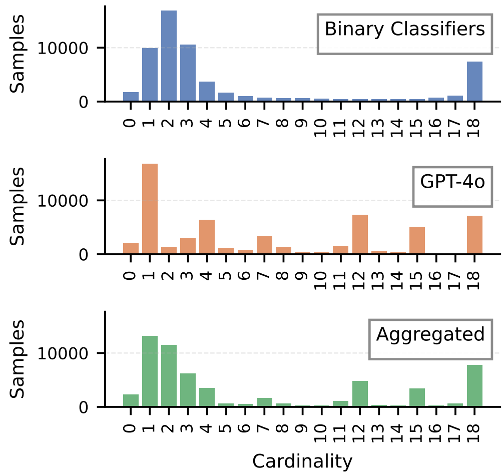

Figure 3: Number of samples for each label cardinality according to the three pseudo-labeling methods.

negative samples are restricted to highly dialectal cases ($a_i > 0.77$), improving precision at the extremes but limiting coverage in the mid-range.

ALDi Bin	Supervision Source	P_{macro}	R_{macro}	$F1_{macro}$
$[0, 0.11)$ (n=7)	(1) Binary Classifiers	100.00	**94.64**	**97.12**
	(2) GPT-4o	100.00	90.48	94.84
$[0.11, 0.44)$ (n=16)	(1) Binary Classifiers	76.99	46.81	55.99
	(2) GPT-4o	**84.72**	**61.36**	**69.72**
$[0.44, 0.77)$ (n=48)	(1) Binary Classifiers	63.28	36.72	44.38
	(2) GPT-4o	**63.86**	**56.94**	**57.15**
$[0.77, 1.0]$ (n=49)	(1) Binary Classifiers	**77.60**	51.38	59.03
	(2) GPT-4o	62.36	**68.14**	**62.62**

Table 4: Performance of the two pseudo-labeling methods on subsets of the NADI2024 development set.

Consequently, we aggregate the predictions of the two pseudo-labeling methods by using the Binary Classifiers at the ALDi extremes ($a_i < 0.11$ or $a_i > 0.77$) and GPT-4o in the intermediate range ($0.11 \leq a_i \leq 0.77$). This aggregation combines the strengths of both methods and achieves the best macro F1 and accuracy on the evaluation set (last row of Table 3), reflecting an improved precision–recall trade-off.

5 Multi-Label Dialect Classification Models

Using the pseudo-labeled multi-label ADI dataset described in Section 4, we fine-tune MARBERT (Abdul-Mageed et al., 2021a) as a multi-label classifier over the 18 Arabic dialects (country-level ADI). We refer to the resulting family of models fine-tuned under different settings, as LAHJAT-

BERT. We train using binary cross-entropy with logits loss, treating each label independently. During fine-tuning, we freeze the bottom 8 transformer layers of MARBERT and update only the top 4 layers and the classification head. We do not fully fine-tune all MARBERT layers since fine-tuning only a fraction of the final layers recovers most downstream effectiveness (Lee et al., 2019). Also, we discard zero-cardinality samples (i.e., instances for which the pseudo-labeling step assigns no dialect), since keeping them would treat the samples as negative for all 18 dialects and inject systematic noise. We report the complete training and inference hyperparameters in Appendix A.

An analysis of the dataset (Figure 3) reveals a strong skew in the cardinality distribution: most samples contain only one or two active dialects. This bias might encourage the model to predict low-cardinality outputs, substantially reducing recall for sentences acceptable in multiple dialects. To mitigate the dataset's inherent low-cardinality bias, we adopt two strategies for curriculum learning (CL). In both strategies, the complexity of the training examples is increased in a controlled manner during the model training process, to achieve better model generalization (Bengio et al., 2009). The exact ordering of both the cardinality buckets and the ALDi buckets is motivated in Appendix F.

5.1 Cardinality-Based Curriculum Learning

In this strategy, the samples' label cardinality is used as a proxy for their respective complexities. For a sentence x_i with a pseudo-labeled target vector $\mathbf{y}_i \in \{0,1\}^{18}$, the label cardinality $c(x_i) = \|\mathbf{y}_i\|_0$ represents the number of dialects in which the sentence is acceptable. This strategy is proposed to mitigate the skewness of the training dataset samples toward lower cardinalities.

We partition the training set into cardinality buckets B_c, where B_c contains samples with label cardinality c. We then define a curriculum ordering over these buckets using the mean loss of each cardinality. Let $\pi(e)$ denote the cardinality index of the bucket selected at curriculum stage e. At stage e, the model is trained on (1) all examples from $B_{\pi(e)}$, and (2) an equal number of randomly sampled examples from each previously introduced bucket $B_{\pi(1)}, \ldots, B_{\pi(e-1)}$.

The model first trains exclusively on the bucket introduced at the first curriculum stage. In the next stage, the bucket selected next is introduced together with an equal number of samples drawn from each previously introduced bucket. This progression continues until the full cardinality range in the dataset is incorporated. The overall training schedule is illustrated in Figure 4, which visualizes how buckets are gradually introduced while maintaining balanced exposure across earlier stages.

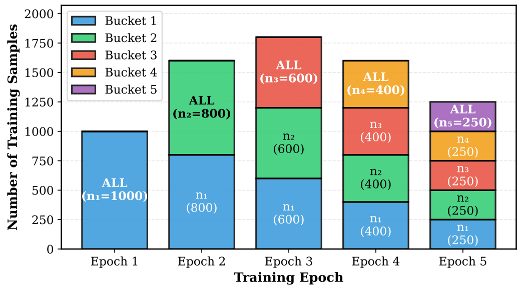

Figure 4: Illustration of the cardinality-based curriculum schedule, showing the progressive introduction of higher difficulty cardinality samples. The numerical values are illustrative and do not reflect the actual dataset.

5.2 ALDi-Based Curriculum Learning

ALDi provides another intuitive proxy for the samples' complexities. Samples with intermediate ALDi scores are expected to be harder than MSA samples with low scores and high-score samples that show clear cues of a specific dialect.

We divide the continuous ALDi range into four contiguous intervals: $I_1 = [0,\ 0.11), I_2 = [0.11,\ 0.44), I_3 = [0.44,\ 0.77), I_4 = [0.77,\ 1]$. For each interval I_k, we construct a bucket B_k that contains all training examples whose ALDi score falls in that range: $B_k = \{x_i \mid a(x_i) \in I_k\}$; where $a(x_i) \in [0,1]$ denotes sentence x_i's ALDi score.

We define a curriculum ordering over the ALDi buckets based on average training loss. Let $\pi(e)$ denote the index of the ALDi bucket selected at epoch e, with buckets ordered from lowest to highest loss. At epoch e, the model is trained on (1) all examples from the current bucket $B_{\pi(e)}$, and (2) a random subset of examples from each previously introduced bucket $B_{\pi(1)}, \ldots, B_{\pi(e-1)}$, following the same sampling strategy as the cardinality-based CL.

6 Results

We evaluate our multi-label dataset construction and training strategies on the NADI2024 development set, which provides multi-label annotations for **8 country-level dialects**. Since our training and supervision span **18 dialects**, we report metrics only on the overlapping label set.

6.1 Multi-Label Model Performance

We evaluate how a single multi-label classifier can learn from different supervision signals by training the same model architecture on each dataset without curriculum learning. This setup isolates the effect of the supervision labels themselves.

As shown in Table 5 (Block II), the model trained on GPT-4o labels achieves higher macro Recall but lower Precision, indicating a tendency to over-predict dialect labels. In contrast, training on the hybrid labels yields higher Precision and Accuracy while maintaining comparable macro F1. This behavior is consistent with the design of the hybrid labels, which combines the conservative behavior of the binary classifiers at ALDi extremes with the richer GPT-based labels in the ALDi mid-range.

Model	$\mathbf{P_{macro}}$	$\mathbf{R_{macro}}$	$\mathbf{F1_{macro}}$	**Acc.**
(I) Baseline				
NADI 2024 Baseline	71.2	30.9	39.7	69.3
(II) Multi-Label Model Performance				
Trained on GPT-4o labels	69.6	**65.7**	66.2	75.5
Trained on hybrid labels	**73.7**	63.7	**67.4**	**77.5**
(III) Curriculum Learning on Hybrid Dataset				
Hybrid + Cardinality-Based CL	69.0	**80.6**	**72.7**	77.5
Hybrid + ALDi-Based CL	**71.4**	71.3	70.3	**78.2**

Table 5: Macro-averaged Precision, Recall, F1, and Accuracy on the NADI2024 development set. Results are grouped into (i) multi-label model performance without CL, and (ii) the effect of CL.

6.2 Impact of Curriculum Learning

We next examine the effect of curriculum learning by comparing curriculum-based training with the baseline setting without a curriculum. As shown in Table 5 (Block III), both curriculum learning strategies improve macro F1 relative to training without curriculum learning, while consistently increasing recall and reducing precision.

In our setup, the curriculum order is derived from a loss-based criterion (Appendix §F), where examples introduced at the latest stages tend to be more ambiguous, corresponding to intermediate label cardinalities and intermediate ALDi scores.

One possible interpretation of the observed recall increase is that exposure at the latest curriculum stages to higher-loss, more ambiguous examples encourages the model to activate a larger set of labels per instance. These examples are characterized by intermediate ALDi scores and multiple valid dialect labels per instance. This broader label coverage may help recover more relevant dialect labels, leading to higher recall. At the same time, predicting more labels per instance can make the model less conservative, which is reflected in the accompanying reduction in precision.

This interaction also helps explain why curriculum learning is most effective when applied to the hybrid supervision. Since the hybrid annotations exhibit higher precision than GPT-4o labels alone, they help limit the precision loss associated with increased recall.

6.3 Generalization on the MLADI Test Set

We evaluate whether the improvements on the development set generalize to the MLADI test set. Table 6 reports the performance of the three LahjatBERT variants compared with the NADI 2024 baseline and previously reported systems.

The NADI 2024 baseline is a single-label dialect identification model, but it is converted into a multi-label predictor at test time. For each sentence, it computes a softmax distribution over the 18 dialect classes, then selects the most likely dialects until their cumulative probability reaches a fixed value (we use the Top-90% setting, i.e., $P{=}0.9$), and returns those dialects as the predicted label set (Abdul-Mageed et al., 2024).

All LahjatBERT variants outperform the shared-task baseline and prior approaches in terms of macro F1, indicating that the gains obtained from the constructed supervision and training strategies extend beyond the development setting. Among the LahjatBERT variants, the ALDi-based curriculum achieves the highest macro F1, while the cardinality-based curriculum yields the highest recall, mirroring the precision-recall trade-offs observed on the development set.[6]

Model	$\mathbf{P_{macro}}$	$\mathbf{R_{macro}}$	$\mathbf{F1_{macro}}$	**Acc.**
LahjatBERT (no curriculum)	**69.0**	69.7	68.0	**79.1**
LahjatBERT + Cardinality CL	59.3	**81.0**	66.6	73.9
LahjatBERT + ALDi CL	65.0	76.4	**69.0**	78.2
Aya-32B (Dang et al., 2024)	49.5	64.5	54.5	65.6
Elyadata (Karoui et al., 2024)	50.2	56.9	52.4	67.0
NADI 2024 Baseline	64.8	39.9	47.0	72.3

Table 6: Macro-averaged Precision, Recall, F1, and Accuracy on the MLADI test set (Keleg et al., 2025), which contains 1,000 sentences annotated for 11 dialects.

[6] We analyze the models' predictions in the Appendix (§G).

7 Conclusion and Future Work

This work examines the limitations of reusing single-label Arabic dialect identification (ADI) datasets for multi-label ADI (MLADI). We specifically highlight that the difficulty of building binary dialect acceptability classifiers lies in the selection of negative samples. While training dynamics can help in automatically flagging wrongly-assigned negative samples, manual verification is still required to assess the acceptability of these flagged samples. Using single-label ADI data, we construct a pseudo-labeled multi-label dataset by aggregating predictions from GPT-4o and binary dialect acceptability classifiers, and introduce LAHJATBERT, a family of BERT-based multi-label models that outperform existing MLADI systems.

Future work could incorporate targeted human annotation for samples identified as ambiguous by the analysis, enabling iterative refinement of multi-label supervision. Finally, our CL approach substantially improves recall. Future work could examine whether similar gains extend to other multi-label tasks beyond dialect identification.

Limitations

Our evaluation is limited by the label coverage of available benchmarks. Although our training and supervision signals span 18 dialects, the `NADI2024` development set provides multi-label annotations for only 8 country-level dialects, so development metrics are computed only on the overlapping label subset. Likewise, the test set covers only 11 dialects, restricting conclusions about generalization to the full 18-dialect label space.

A second limitation originates from the fact that NADI relies on user geo-location as a proxy for dialect rather than direct linguistic annotation, which can be noisy when posting location and actual dialect do not align.

Ethics and Broader Impact

This work uses anonymized Arabic tweets from the publicly released NADI shared-task datasets, collected via the Twitter API under the platform's terms. Following NADI licensing restrictions, we release only tweet IDs and our automatically generated multi-label annotations (not the underlying tweet text), along with code and trained models to enable reproducibility. The research is methodological, and while dialect identification can support beneficial language technologies and sociolinguistic analysis, it also carries risks of misuse (e.g., profiling or surveillance), which we explicitly do not endorse and is outside the scope of this work.

References

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021. QADI: Arabic dialect identification in the wild. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 1–10, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The fourth nuanced Arabic dialect identification shared task. In *Proceedings of ArabicNLP 2023*, pages 600–613, Singapore (Hybrid). Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, et al. 2021a. Arbert & marbert: Deep bidirectional transformers for arabic. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 7088–7105.

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The fifth nuanced Arabic dialect identification shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 709–728, Bangkok, Thailand. Association for Computational Linguistics.

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

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

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

Maha J. Althobaiti. 2020. Automatic arabic dialect identification systems for written texts: A survey.

Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2022. Hierarchical aggregation of dialectal data for Arabic dialect identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4586–4596, Marseille, France. European Language Resources Association.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Gabriel Bernier-colborne, Cyril Goutte, and Serge Leger. 2023. Dialect and variant identification as a multi-label classification task: A proposal based on near-duplicate analysis. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy. Association for Computational Linguistics.

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, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier.

Nizar Y Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Vani Kanjirangat, Tanja Samardzic, Ljiljana Dolamic, and Fabio Rinaldi. 2024. NLP_DI at NADI 2024 shared task: Multi-label Arabic dialect classifications with an unsupervised cross-encoder. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 742–747, Bangkok, Thailand. Association for Computational Linguistics.

Amira Karoui, Farah Gharbi, Rami Kammoun, Imen Laouirine, and Fethi Bougares. 2024. ELYADATA at NADI 2024 shared task: Arabic dialect identification with similarity-induced mono-to-multi label transformation. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 758–763, Bangkok, Thailand. Association for Computational Linguistics.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. ALDi: Quantifying the Arabic level of dialectness of text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. Association for Computational Linguistics.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2025. Revisiting common assumptions about Arabic dialects in NLP. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3327, Vienna, Austria. Association for Computational Linguistics.

Amr Keleg and Walid Magdy. 2023. Arabic dialect identification under scrutiny: Limitations of single-label classification. In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.

Jaejun Lee, Raphael Tang, and Jimmy Lin. 2019. What would elsa do? freezing layers during transformer fine-tuning.

Javier A. Lopetegui, Arij Riabi, and Djamé Seddah. 2025. Common ground, diverse roots: The difficulty of classifying common examples in Spanish varieties. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 168–181, Abu Dhabi, UAE. Association for Computational Linguistics.

Helene Olsen, Samia Touileb, and Erik Velldal. 2023. Arabic dialect identification: An in-depth error analysis on the MADAR parallel corpus. In *Proceedings of ArabicNLP 2023*, pages 370–384, Singapore (Hybrid). Association for Computational Linguistics.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Omar F. Zaidan and Chris Callison-Burch. 2011. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, Portland, Oregon, USA. Association for Computational Linguistics.

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

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Mahesh Bangera. 2024. Language variety identification with true labels. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10100–10109, Torino, Italia. ELRA and ICCL.

A Experimental Details

We provide the full training and inference hyperparameters for the multi-label setup described in section 5. We fine-tune a BERT-based model for multi-label country-level Arabic dialect identification with 18 binary labels, where the classification head outputs one logit per country. Training minimizes binary cross-entropy with logits, which implements a combination of sigmoid and binary cross-entropy and treats each label independently. We split the data into 90% training and 10% validation with a fixed random seed of 42. During fine-tuning, we freeze the bottom 8 transformer layers of MARBERT and update only the top 4 layers and the classification head. We set both the hidden-state and attention dropout probabilities to 0.3. We train for 3 epochs with a batch size of 24 for both training and evaluation, evaluating once per epoch. The best checkpoint is selected based on validation micro F1. At inference time, we apply a sigmoid to the logits and use a threshold of 0.3 to obtain binary label assignments. We chose this value by maximizing validation micro F1 on the held-out split, the default threshold of 0.5 yields lower validation macro F1, so we use 0.3 for all reported results.

B ALDi Threshold Selection

We adopt two ALDi thresholds, $a_i = 0.11$ and $a_i = 0.77$, to distinguish sentences with minimal dialectal evidence from those with strong dialectal evidence. These thresholds are grounded in the annotation scheme and definition of the Arabic Level of Dialectness (ALDi) score introduced by Keleg et al. (2023).

ALDi Annotation Scheme. Following the ALDi annotation protocol described by Keleg et al. (2023), each sentence is annotated by three native Arabic speakers. Annotators assign one of four ordinal labels reflecting the degree of dialectness: *MSA* (0), *Little* ($\frac{1}{3}$), *Mixed* ($\frac{2}{3}$), and *Most* (1). The ALDi score for a sentence is defined as the mean of the three annotations. Consequently, ALDi values lie in $[0, 1]$ and take discrete steps of $\frac{1}{9}$.

Derivation of the Thresholds. Under this definition, the smallest non-zero ALDi value is

$$\frac{0 + 0 + \frac{1}{3}}{3} = \frac{1}{9} \approx 0.11,$$

which corresponds to the weakest possible evidence of dialectal content, where two annotators label the sentence as MSA and one assigns a *Little* dialect label. Conversely, an ALDi score of

$$\frac{1 + 1 + \frac{1}{3}}{3} = \frac{7}{9} \approx 0.77$$

corresponds to strong dialectal evidence, where two annotators assign the highest dialectness label (*Most*) and the third assigns at least *Little*. These two values therefore mark natural boundaries between minimal dialectal signal ($a_i < 0.11$), strong dialectal signal ($a_i > 0.77$), and an intermediate range characterized by mixed or graded annotator judgments.

ALDi Score Distribution. ALDi scores are concentrated near the MSA ($a_i < 0.11$) and highly dialectal ($a_i > 0.77$) ranges, with fewer samples in the mid-range ($0.11 \leq a_i \leq 0.77$), as shown in Figure B1.

C Non-Neighbouring Dialect Regions

When constructing negative samples for the binary dialect classifiers (section 4), we restrict negatives to sentences geolocated in non-neighbouring countries. Neighbouring countries are defined as those sharing a land border with the target country, and negative samples are drawn exclusively from non-bordering countries.

By restricting negative samples to non-bordering countries, we avoid cases where sentences may be linguistically compatible with the target dialect, resulting in more reliable negative supervision.

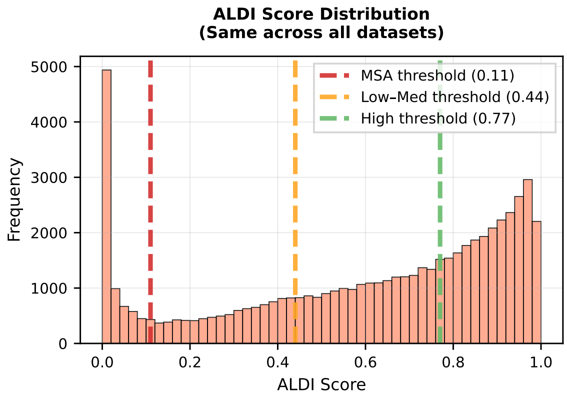

Figure B1: Distribution of ALDi scores in the dataset. Vertical dashed lines indicate the thresholds distinguishing MSA text ($a_i < 0.11$), low–medium dialectness (0.11–0.44), medium dialectness (0.44–0.77), and highly dialectal text ($a_i > 0.77$).

D Dataset Diagnostics

This appendix presents diagnostics of the constructed multi-label dataset, focusing on how label cardinality varies with the Arabic Level of Dialectness (ALDi) score for the supervision sources considered in this work.

D.1 Label Cardinality Across ALDi Ranges

Figure D2 shows the distribution of label cardinality across ALDi categories for each labeling method, as well as for the aggregated dataset.

Binary Dialect Classifiers. The binary classifiers exhibit a strong dependence on the ALDi score. For sentences with low ALDi values (MSA text), they frequently activate many dialect labels, resulting in high cardinality. For sentences with high ALDi values (strongly dialectal text), they tend to activate very few labels. In the intermediate ALDi ranges ($0.11 \leq a_i \leq 0.77$), where explicit negative supervision is absent, classifier outputs concentrate at either high or low cardinalities, and intermediate label counts are rarely observed.

GPT-4o. GPT-4o displays a smoother relationship between ALDi and label cardinality. While cardinality generally decreases as ALDi increases, GPT-4o assigns moderate numbers of labels more frequently in the Low–Med and Medium ALDi ranges. This pattern is consistent with graded representations of dialectal overlap and contrasts with the polarized outputs of the binary classifiers in the same ALDi regions.

Aggregation Effects. The aggregated dataset (right panel of Figure D2) combines these two behaviors. At low and high ALDi values, where binary classifier predictions are stable, the aggregated cardinality closely matches their outputs. In the intermediate ALDi ranges, where binary classifier predictions are concentrated at extreme cardinalities, the aggregation relies on GPT-4o, resulting in intermediate label counts and a smoother dependence of cardinality on ALDi.

D.2 Overall Label Cardinality Distribution

Figure 3 shows the overall distribution of label cardinality for the supervision sources used in this work, independent of ALDi conditioning.

E Prompt Template for GPT-Based Annotation

Instruction: You are a native Arabic speaker and highly qualified linguist with expert-level understanding of regional Arabic dialects.
Given the sentence provided, evaluate its dialectal characteristics independently for each of the following dialects: Iraq, Egypt, Morocco, Libya, UAE, Saudi Arabia, Bahrain, Syria, Lebanon, Oman, Palestine, Algeria, Jordan, Tunisia, Kuwait, Yemen, Sudan, and Qatar.
Return findings in JSON format:

```
{
  "Iraq": 0/1,
  "Egypt": 0/1,
  ...
  "Qatar": 0/1
}
Input sentence: {tweet}
```

Figure E3: GPT-based annotation prompt template.

F Curriculum Learning Ordering

To obtain a principled ordering for both the cardinality-based curriculum learning and the ALDi-based curriculum learning, we first train a baseline MARBERT model on the aggregated multi-label dataset without any curriculum learning. After fine-tuning, we freeze this baseline model and run inference on the full training set, recording the per-example loss.

We then aggregate these losses in two ways: (i) by label cardinality, computing the mean loss for each cardinality value, and (ii) by ALDi bin, also computing the mean loss but for each ALDi interval. The resulting mean-loss profiles are visualized

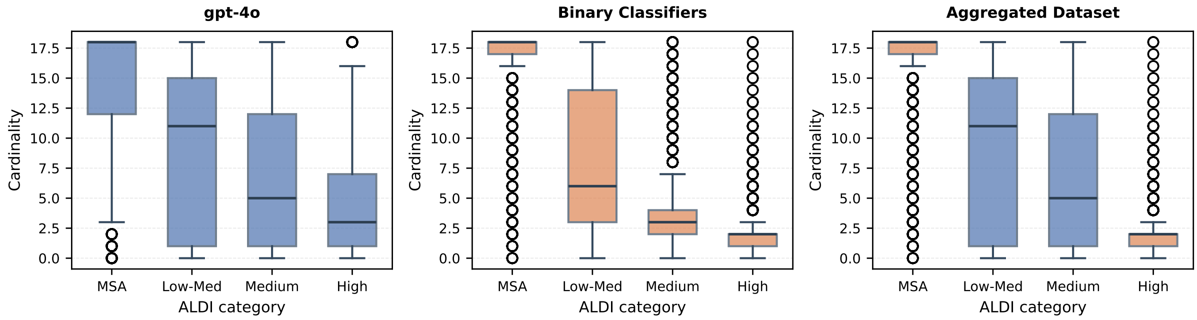

Figure D2: Cardinality across ALDi categories for the three labeling methods. Due to instabilities, we rely on GPT-4o for ambiguous cases and on the binary classifiers when the dialectness is clearly low (MSA) or high. This leads to more consistent and linguistically plausible multi-label patterns.

in Figure F4 (mean loss per cardinality) and Figure F5 (mean loss per ALDi bin).

We treat the mean loss of a bucket as a proxy for its difficulty, and order the stages from lower-loss (easier) buckets to higher-loss (harder) buckets. This difficulty-based ordering directly determines the progression of stages in both the cardinality-based and ALDi-based curriculum learning schedules described in section 5.

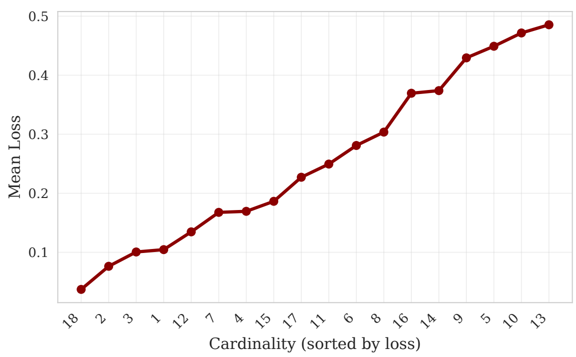

Figure F4: Sorted mean loss per cardinality, to measure the difficulty for the baseline model in predicting different cardinalities.

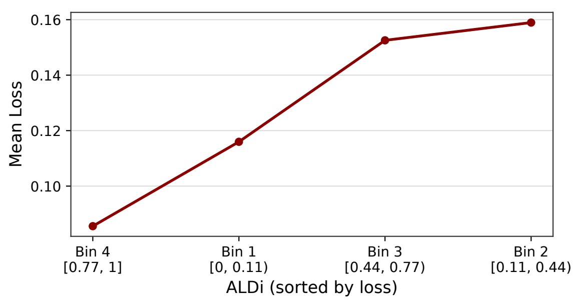

Figure F5: Sorted mean loss per ALDi bin, to measure the difficulty for the baseline model in predicting different cardinalities.

G Model's Predictions on MADAR's Samples

To analyze the behavior of the three newly introduced models, we study their predictions on 200 sentences from the MADAR-26 corpus (Salameh et al., 2018), for each of the 6 anchor dialects identified by the corpus creators. Additionally, we contrast their predictions to two baseline models: a single-label DI model **DID-Country**, and the multi-label DI **NADI2024 shared task's baseline**.

Comparing our three LAHJATBERT models to the NADI2024 baseline, it is clear that the baseline is more susceptible to predicting the sentence to be acceptable in countries from different regions than our newly introduced model. For instance, out of the 200 sentences in Rabat's dialect (spoken in Morocco), the baseline model unintuitively predicts that 190 of them are acceptable in Egypt, in comparison to less than 20 of them predicted to be acceptable by the three LAHJATBERT models. Hence, assuming that the labels within the top-p (p=90%) of a single-label DI model as acceptable dialects for the input sentence is not an optimal strategy for multi-label dialect identification.

For the three LAHJATBERT models, the Cardinality-based one seems to frequently predict labels from other regions, which explains its lower precision yet higher recall than the two other models. The ALDi-based model seem to be achieving the best compromise between the precision and the recall, as indicated by the fact it achieves the highest overall macro-F1 score.

Dialect	DID-Country (Salameh et al., 2018)	NADI2024 Baseline (Abdul-Mageed et al., 2024)	LahjatBERT	LahjatBERT + ALDi CL	LahjatBERT + Cardinality CL
BEI (200)	**LB 129**, <u>SY 32</u>, <u>JO 15</u>, <u>PS 12</u>, EG 4, SA 3, IQ 2, OM 2, MA 1	**LB 200**, <u>SY 200</u>, <u>JO 199</u>, <u>PS 199</u>, EG 180, IQ 174, TN 160, SD 127, BH 86, MA 78, KW 68, DZ 57, SA 56, QA 53, LY 41, OM 36, YE 25, AE 19	<u>SY 165</u>, **LB 149**, <u>PS 131</u>, <u>JO 121</u>, IQ 36, SA 35, AE 34, KW 32, BH 31, OM 31, QA 29, YE 27, EG 23, LY 18, SD 17, TN 10, DZ 9, MA 9	<u>SY 174</u>, **LB 146**, <u>PS 135</u>, <u>JO 129</u>, OM 35, SA 33, AE 33, IQ 31, KW 31, BH 30, QA 30, YE 26, EG 25, SD 19, LY 18, TN 9, DZ 8, MA 7	<u>SY 188</u>, **LB 184**, <u>PS 181</u>, <u>JO 173</u>, SA 45, AE 44, BH 42, OM 42, KW 41, QA 41, YE 41, IQ 35, EG 31, SD 23, LY 22, MA 9, TN 9, DZ 7
CAI (200)	**EG 147**, SA 12, SY 12, <u>SD 9</u>, JO 7, DZ 5, LY 4, TN 4, MA 3, YE 3, IQ 2, PS 2, QA 2, KW 1, LB 1, BH 1, OM 1, AE 1	**EG 200**, YE 193, SA 192, <u>SD 192</u>, PS 188, JO 187, LY 182, LB 178, SY 167, MA 159, OM 142, TN 130, IQ 73, KW 54, QA 51, AE 40, DZ 20	**EG 190**, PS 131, <u>SD 115</u>, LB 54, SA 54, JO 52, SY 52, LY 51, KW 46, IQ 44, BH 44, QA 44, AE 43, OM 42, YE 42, TN 14, MA 12, DZ 10	**EG 188**, PS 152, <u>SD 132</u>, LY 49, SA 49, JO 46, LB 45, IQ 44, SY 44, KW 43, BH 41, OM 41, QA 41, AE 41, YE 41, TN 11, MA 9, DZ 8	PS 193, **EG 189**, <u>SD 165</u>, SA 135, LB 103, JO 99, SY 93, QA 83, OM 81, AE 80, BH 79, YE 79, KW 74, LY 73, IQ 60, TN 14, MA 12, DZ 10
DOH (200)	**QA 144**, <u>SA 17</u>, IQ 7, JO 5, SY 5, <u>OM 4</u>, MA 3, SD 3, TN 3, PS 3, LY 2, EG 2, YE 2	<u>SA 200</u>, **QA 199**, <u>OM 198</u>, LY 194, JO 191, SY 181, IQ 163, SD 156, YE 137, <u>AE 122</u>, MA 112, <u>KW 86</u>, LB 80, TN 77, PS 76, <u>BH 64</u>, EG 59, DZ 14	<u>SA 173</u>, <u>KW 154</u>, <u>BH 142</u>, <u>AE 136</u>, **QA 133**, <u>OM 128</u>, IQ 115, YE 111, PS 93, JO 91, SY 91, LB 89, SD 55, LY 54, EG 52, MA 24, TN 24, DZ 21	<u>SA 179</u>, <u>KW 165</u>, <u>BH 151</u>, **QA 147**, <u>AE 142</u>, <u>OM 134</u>, YE 117, IQ 111, PS 105, LB 102, JO 100, SY 100, EG 53, SD 49, LY 46, MA 17, TN 17, DZ 14	<u>SA 197</u>, <u>BH 194</u>, **QA 193**, <u>KW 190</u>, <u>AE 190</u>, <u>OM 175</u>, YE 149, IQ 132, PS 115, JO 111, LB 96, SY 96, SD 53, EG 50, LY 46, MA 14, TN 13, DZ 11
RAB (200)	**MA 176**, <u>DZ 6</u>, <u>TN 4</u>, JO 3, <u>LY 3</u>, SA 3, PS 2, OM 1, SY 1, QA 1	**MA 198**, <u>DZ 196</u>, <u>LY 194</u>, EG 190, QA 188, LB 180, AE 165, SA 159, SY 154, SD 151, <u>TN 151</u>, PS 147, KW 143, BH 125, IQ 70, JO 41, OM 32, YE 11	<u>DZ 172</u>, **MA 171**, <u>TN 84</u>, <u>LY 27</u>, EG 16, PS 15, IQ 13, SY 13, JO 11, LB 11, SA 11, OM 10, AE 10, KW 9, BH 8, QA 8, YE 8, SD 7	<u>DZ 181</u>, **MA 175**, <u>TN 68</u>, <u>LY 23</u>, PS 14, EG 10, IQ 8, SY 8, JO 7, SA 7, LB 6, AE 6, QA 6, KW 5, BH 5, OM 5, YE 4, SD 1	**MA 178**, <u>DZ 174</u>, <u>TN 130</u>, <u>LY 52</u>, PS 32, SY 30, JO 29, LB 27, AE 25, KW 24, OM 24, SA 24, QA 24, YE 24, IQ 21, BH 20, EG 19, SD 16
TUN (200)	**TN 167**, <u>DZ 6</u>, SA 5, JO 4, IQ 3, OM 3, PS 3, <u>LY 2</u>, <u>MA 2</u>, SY 2, QA 2, YE 1	<u>LY 198</u>, **TN 197**, LB 192, IQ 191, <u>DZ 190</u>, EG 189, <u>MA 183</u>, OM 181, SD 180, QA 160, YE 150, KW 104, PS 99, AE 60, SY 42, SA 40, JO 28, BH 14	**TN 159**, <u>LY 149</u>, <u>DZ 143</u>, <u>MA 94</u>, IQ 24, SY 24, PS 24, LB 23, JO 21, EG 19, KW 17, SA 17, AE 16, YE 15, BH 14, OM 14, SD 14, QA 14	**TN 163**, <u>DZ 155</u>, <u>LY 153</u>, <u>MA 57</u>, PS 28, IQ 25, JO 23, LB 23, SY 23, EG 21, KW 20, SA 20, AE 20, OM 19, YE 18, BH 17, QA 16, SD 15	<u>LY 170</u>, **TN 162**, <u>DZ 159</u>, <u>MA 103</u>, PS 48, SY 44, JO 42, LB 41, IQ 35, SA 33, KW 32, OM 31, AE 31, QA 30, YE 30, BH 29, EG 28, SD 21
MSA (200)	**OM 162**, **SA 162**, **SD 152**, **LY 151**, **SY 151**, **DZ 150**, **IQ 149**, **JO 149**, **EG 149**, **PS 148**, **KW 147**, **LB 147**, **BH 147**, **MA 147**, **QA 147**, **TN 147**, **AE 147**, **YE 147**	**OM 200**, **SA 200**, **SD 200**, **IQ 195**, **YE 189**, **JO 165**, **AE 151**, **QA 145**, **LY 144**, **EG 143**, **DZ 111**, **KW 100**, **BH 91**, **TN 75**, **MA 57**, **SY 56**, **PS 30**, **LB 22**	**IQ 196**, **JO 195**, **LB 195**, **PS 195**, **SA 194**, **SY 194**, **AE 194**, **KW 193**, **LY 193**, **EG 193**, **OM 192**, **YE 192**, **BH 191**, **SD 191**, **QA 190**, **TN 188**, **MA 187**, **DZ 183**	**IQ 198**, **EG 198**, **LY 197**, **PS 197**, **SA 197**, **JO 196**, **KW 196**, **LB 196**, **SD 196**, **SY 196**, **AE 196**, **YE 196**, **BH 195**, **OM 195**, **QA 195**, **MA 189**, **TN 189**, **DZ 188**	**IQ 200**, **JO 200**, **KW 200**, **BH 200**, **OM 200**, **PS 200**, **QA 200**, **SA 200**, **AE 200**, **YE 200**, **LB 199**, **SY 199**, **LY 197**, **SD 197**, **EG 196**, **MA 190**, **TN 190**, **DZ 188**

Table G1: The number of times a country label is predicted by the model for the 200 sentences of the six anchor dialects of the MADAR CORPUS-26 corpus (Salameh et al., 2018): Beirut (BEI), Cairo (CAI), Doha (DOH), Rabat (RAB), Tunis (TUN), and MSA. Each dialect's representative country is in **bold**, and the country's same-region countries are <u>underlined</u>, following the regional grouping of Baimukan et al. (2022). **Note #1:** the DID-country model is a single-label DI model trained on the MADAR corpus. Its province-level predictions are mapped into country-level ones, with MSA predictions mapped to all the 18 country-level dialects we consider. **Note #2:** Country abbreviations - **AE**: UAE, **BH**: Bahrain, **DZ**: Algeria, **EG**: Egypt, **IQ**: Iraq, **JO**: Jordan, **KW**: Kuwait, **LB**: Lebanon, **LY**: Libya, **MA**: Morocco, **OM**: Oman, **PS**: Palestine, **QA**: Qatar, **SA**: Saudi Arabia, **SD**: Sudan, **SY**: Syria, **TN**: Tunisia, **YE**: Yemen.