IDIOLEX: Unified and Continuous Representations for Idiolectal and Stylistic Variation

Anonymous ACL submission

Abstract

Existing sentence representations primarily encode *what* a sentence says, rather than *how* it is expressed, even though the latter is important for many applications. In contrast, we develop sentence representations that capture style and dialect, decoupled from semantic content. We call this the task of *idiolectal representation learning*. We introduce IDIOLEX, a framework for training models that combines supervision from a sentence's provenance with linguistic features of a sentence's content, to learn a continuous representation of each sentence's style and dialect. We evaluate the approach on dialects of both Arabic and Spanish. The learned representations capture meaningful variation and transfer across domains for analysis and classification. We further explore the use of these representations as training objectives for stylistically aligning language models. Our results suggest that jointly modeling individual and community-level variation provides a useful perspective for studying idiolect and supports downstream applications requiring sensitivity to stylistic differences, such as developing diverse and accessible LLMs[1].

1 Introduction

Much of mainstream evaluation and optimization of large language models (LLMs) prioritizes semantic correctness (Clark et al., 2020; Lewis et al., 2020; Liu et al., 2023; Chiang et al., 2024; Singh et al., 2025; Kim et al., 2025). User-specific linguistic and stylistic adaptation receives comparatively less emphasis in standard benchmarks, despite substantial literature on controllable generation, style transfer, and personalization (Jin et al., 2022; Dong et al., 2023; Durandard et al., 2025a). Although semantic correctness is essential, maximizing understanding in languages with diverse speaker populations requires modeling variation across dialect, register, and style, which we collectively refer to as *idiolect* (Allan, 2013).

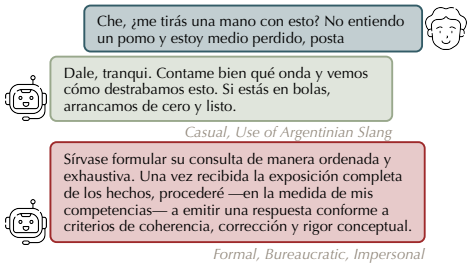

Figure 1: IDIOLEX used to compare the idiolectal alignment between user input and GPT 5.1 generations in casual Argentinian Spanish.

In 1948, Bernard Bloch coined the term *idiolect* to refer to "the totality of the possible utterances of one speaker at one time in using language to interact with one other speaker" (Bloch, 1948). In modern linguistics, the concept of idiolect encompasses all the person-, interlocutor-, and context-specific considerations of language use. Since LLMs are not stable interlocutors with consistent personalities and backgrounds (Durandard et al., 2025a), the lack of common linguistic ground can result in misunderstanding of both users' and systems' idiolect (Frisch and Giulianelli, 2024; Fleisig et al., 2024; Nias et al., 2025; Lin et al., 2025).

Prior NLP work related to idiolect has largely appeared under task-specific formulations such as dialect identification (DID; Tiedemann and Ljubešić, 2012; Lui and Cook, 2013) and authorship attribution (AA; Stamatatos et al., 1999; Zhu and Jurgens, 2021; Rivera-Soto et al., 2021; Wegmann et al., 2022). Although these tasks rely on similar linguistic signals to one another, they have historically been studied in isolation. Similarly, there have been attempts to learn style-aware representations for authorship verification using either contrastive learning over sentence pairs based on authorship (Wegmann et al., 2022) or explicit linguistic fea-

[1]Code: https://anonymous.4open.science/r/IdioleX-3D45

tures (Patel et al., 2023), but none that combine both or make use of levels of idiolectal proximity.

We propose IDIOLEX, a training framework for learning idiolectal sentence representations with minimal supervision. These representations allow us to capture *idiolectal similarity* between utterances and to support a range of applications, including comparisons between text segments and downstream uses such as post-training or evaluation of LLMs. IDIOLEX is language-agnostic and relies on weak supervision from hierarchical proximity and LLM-generated linguistic features rather than task-specific labels. The resulting representations are designed to de-prioritize semantic content while encoding stylistic and linguistic structure in a continuous space, enabling applications across analysis, classification, and alignment.

Specifically, we show that IDIOLEX representations are successful in multiple contexts:

1. **A Base for Existing Tasks** that outperforms prior neural approaches on DID and AA across Arabic and Spanish varieties (§4).
2. **Embeddings for Similarity Scoring** that encode linguistic structure and dialectal variation with minimal semantic interference (§5).
3. **A Style Alignment Training Objective** to improve idiolectal and dialectal alignment during post-training (§6).

2 Idiolectal Representation Learning

Idiolectal variation – differences in how individuals use language in terms of lexicon, morphology, syntax, and register – has long been studied in work on authorship attribution, dialect identification, and style transfer (Zhu and Jurgens, 2021; Wegmann et al., 2022; Patel et al., 2025). We build on these lines of research towards idiolectal representation learning (IRL) to encode stylistic and dialectal properties independent of the task.

Stylistic & Dialectal Similarity We assume that stylistic and dialectal similarity correlate with a *proximity hierarchy*: that language use tends to be most similar within the same individual and same document, followed by the same individual across documents, then among speakers within a shared dialect community, and least similar across different dialect communities. As such, idiolectal similarity can be modeled as a continuous space reflecting shared linguistic features across speakers and communities. We use this proximity hierarchy (same document > same author > same dialect > cross-dialect) to provide weak supervision during training while avoiding discrete categorization.

Relationship to Semantic Content We consider the following desideratum: if two documents are stylistically similar but discuss different topics, their representations should remain relatively close, whereas documents that are semantically similar but differ substantially in style or dialect should be more distant. While stylistic and semantic signals are intertwined, our objective emphasizes linguistic form over topical content.

3 IDIOLEX

We introduce IDIOLEX, a framework for training sentence representations that encode idiolectal variation by combining weak supervision from proximity information (defined above, §2) with LLM-extracted linguistic features. Our goal is to encourage representations that are consistent with the view of stylistic variation described in Section 2. Rather than relying on a single supervision signal, we integrate multiple sources that provide complementary perspectives on linguistic variation.

3.1 Data Schema

Data Sampling & Filtering We construct training data from publicly available Reddit comments associated with regional communities. For each language, we identify the main regional subreddits and collect comments from the inception of the subreddit until December 2024 using the Pushshift archive (Baumgartner et al., 2020) and filter for language and poor-quality. A full list of subreddits and filtering criteria are provided in Appendix A.

Proximity Scores Because stylistic similarity is not directly observable, we approximate it using relative comparisons derived from authorship and community structure. For each anchor sentence, we sample comparison sentences representing varying degrees of proximity: (3) same comment, (2) same author, (1) same dialect community, and (0) different communities. These groupings provide a weak but structured signal that encourages the representation space to reflect stylistic relationships without requiring explicit labels. Proximity scores are assigned for each pair of sentences between 0 (different community) and 3 (same comment) based on the most exclusive category the sentence belongs to, as seen in Figure 2. This ranking is inherently noisy; however, across a large number of samples,

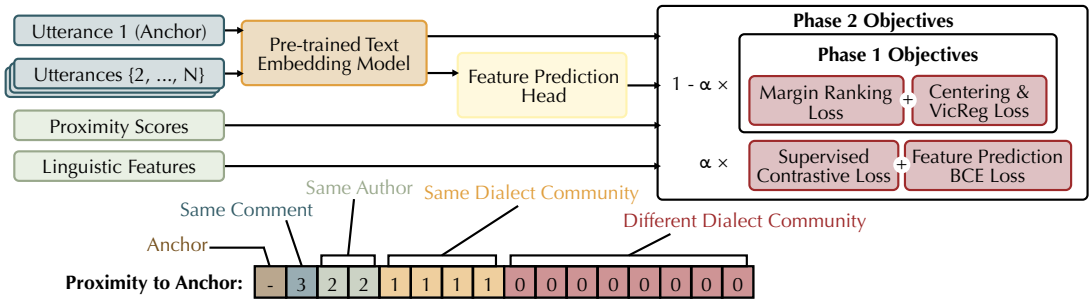

Figure 2: IDIOLEX training framework. During training, all batches are sampled such that every individual item can act as an anchor for contrastive learning, necessitating that there are 2^{3-n} samples for each proximity score $n \in [0, 3]$.

it still provides useful gradients. To mitigate this noise, we introduce an additional level of supervision using LLM-generated dialectal features.

Linguistic Features In addition to relational supervision, we incorporate automatically extracted linguistic features that reflect common dimensions of dialectal variation (Williams et al., 2008). These linguistic features are extracted from linguistic textbooks on Spanish (Lipski, 2012) and Arabic (Ryding and Wilmsen, 2021) dialectology. Feature extraction is performed using an LLM prompted to identify whether each attribute is present on a sentence-by-sentence level, since LLMs have been demonstrated to have meta-linguistic capabilities (Begus et al., 2025). In this work, we use GPT 5 mini[2] for feature extraction. We view these signals as approximate cues whose limitations (varying extraction quality) are mitigated through complementary supervision from relational structure. Details of prompts, feature definitions, and annotation procedures are provided in Appendix A.3.

3.2 Architecture

The IDIOLEX framework builds on a pretrained sentence encoder that serves as the backbone for representation learning. This design allows the framework to be applied to different languages by substituting the underlying pretrained model. Importantly, the architecture is intentionally minimal: it does not assume task-specific structures or discrete label prediction, and can be used as a general-purpose representation learner for stylistic and dialectal variation.

Past work has demonstrated that transformer-based language models encode this information across different layers (Tenney et al., 2019). Sentence representations are obtained by applying learned layer-wise attention over all transformer layers (inspired by (Rei et al., 2020)) followed by token-level mean pooling. The resulting representations are mean-centered and L2-normalized during training.

Training proceeds in two stages due to limitations on the quantity of data that can be annotated with linguistic features. First, the encoder is optimized on the full train set using a relative similarity objective – a margin ranking loss – that encourages sentences with higher expected stylistic overlap to be closer in representation space than less related sentences. This pre-training primes the model to move from semantic to stylistic encoding and makes use of the significantly larger amount of unannotated data to maximize generalizability. In the second stage, feature-level supervision is incorporated alongside the margin ranking loss on a smaller subset of the training data through auxiliary objectives that encourage the representation to encode dimensions captured by the extracted linguistic features. Combining relational and feature signals allows the model to balance global structure with interpretable variation. We additionally apply normalization and regularization techniques to stabilize training and reduce representation collapse, promoting a well-conditioned embedding space suitable for downstream analysis.

Margin Ranking Loss The margin ranking loss calculates the pairwise difference in the semantic similarity s of each sentence with a given *anchor* sentence, a, and then compares it to the difference in proximity score r between sentences i and j:

$$\mathrm{MRL}(a, i, j) = \max(0, -(r(x_a, x_i) - r(x_a, x_j)) \times (s(x_a, x_i) - s(x_a, x_j)) + \lambda).$$

In other words, we enforce that higher-proximity pairs have a greater cosine similarity by at least

[2] gpt-5-mini-2025-08-07

λ. Because we structure our mini-batch such that every sentence can act as an anchor, see Figure 2, we sum over every possible anchor and pair of comparison sentences for the final loss. During training, the margin λ starts at 0, then increases linearly to 0.5.

Feature Prediction The feature prediction head takes the idiolectal sentence representations and applies a two-layer prediction model for the set of LLM-generated linguistic features, with a binary cross-entropy loss applied on the logits.

Supervised Contrastive Loss The supervised contrastive loss acts as an additional similarity loss, but in place of the ranking from earlier, it uses the Jaccard similarity $w(x_i, x_j)$ over the feature vectors x as the weights for a supervised contrastive loss for sentences in a batch B and $\tau = 0.07$:

$$\mathrm{SCL}(i,j) = w(x_i, x_j) \log \frac{e^{\mathrm{sim}(z_i,z_j)/\tau}}{\sum_{k \in B} e^{\mathrm{sim}(z_i,z_k)/\tau}}.$$

As with MRL, we sum over all combinations of $i, j \in B$. This allows us to better identify the true similarity of sentences from different dialects as a linguist would see them, but we are limited by the chosen features. Both feature losses are used after a larger pre-training stage because the features are only available for a subset of the data.

Centering & Variance Constraints We enforce mean centering around 0 using a learned μ and combine a variance and decorrelation loss to encourage a standard deviation of 1.0 across each dimension and decorrelation between dimensions to avoid anisotropy. The centering vector μ is learned jointly with the encoder parameters using batch-wise estimates.

Each language's IDIOLEX model is trained independently, starting from a language-specific pre-trained base model, AraBert v2[3] for Arabic and Bertin[4] for Spanish (Antoun et al., 2020; la Rosa et al., 2022). Data from 10 authors from each dialect is withheld for each of the development and test sets (150 in total per set). Hyperparameters, model configurations, and optimization settings are reported in Appendix B.1 to facilitate reproducibility.

[3] https://huggingface.co/aubmindlab/bert-base-arabertv2

[4] https://huggingface.co/bertin-project/bertin-roberta-base-spanish

4 Performance on Classification Benchmarks

We show that IDIOLEX provides strong performance on downstream style-related tasks with the addition of task-specific heads. We evaluate this ability for dialect identification (DID) and authorship attribution (AA). For each task, we evaluate three versions of the models trained via IDIOLEX.

1. **IDIOLEX**: The model representations as-is, with a trained classification head, demonstrating the performance of the representations in a cross-domain context.
2. **Fine-tuned IDIOLEX**: The model representations with a trained classification head after finetuning on the train split for each task.
3. **Fine-tuned IDIOLEX + Lexical**: The model representations with a trained classification head after finetuning, ensembled with a trained logistic regression head that is fed TF-IDF representations. The IDIOLEX framework prioritizes the encoding of the morphological and syntactic features associated with idiolect (see Appendix A.3), so the addition of explicit lexical features aids in tasks where lexical variation gives significant information.

We evaluate IDIOLEX against two baselines: (1) a fine-tuned BERT-based system (similar training to Fine-tuned IDIOLEX), and (2) a centroid clustering baseline. For any systems based on BERT, we use the same monolingual models used as the pre-trained encoder for training our IDIOLEX models. The centroid clustering baseline is constructed with frozen encoder representations from which dialect centroids are calculated on the train data. Where comparable, we also report winning shared-task scores on the datasets we evaluate on. More information on classification model training is found in Appendix B.2.1, and further details regarding the baselines can be found in Appendix B.2.2.

4.1 Dialect Identification

DID relies heavily on stylistic cues such as lexical choices, morphology, and syntactic patterns, which may overlap across dialects. For Spanish, we use the multi-label DSL-ML 2024 shared task dataset (Chifu et al., 2024; Zampieri et al., 2024). For Arabic, we evaluate on MADAR 26 (Bouamor et al., 2018, 2019), which contains sentences annotated with 25 city-level dialect labels along with Modern Standard Arabic.

Table 1 reports F1 scores and exact match accu-

Models	Spanish F1	Spanish EM	Arabic F1	Arabic EM
IDIOLEX	**0.85**	0.62	0.43	0.43
Fine-tuned IDIOLEX	0.84	**0.63**	0.61	0.61
Fine-tuned IDIOLEX + Lexical	**0.85**	**0.63**	0.66	0.66
Fine-tuned BERT	0.80	0.59	0.56	0.56
Centroid Clustering w/ BERT	0.77	0.57	0.20	0.22
Centroid Clustering w/ E5	0.74	0.54	0.15	0.17
Sälevä and Palen-Michel, 2024	0.82	0.50	–	–
Abu Kwaik and Saad, 2019	–	–	**0.67**	**0.67**
Samih et al., 2019	–	–	0.59	0.59

Table 1: DID F1-score and exact match accuracy (EM) results on Spanish (DSL-ML) and Arabic (MADAR 26). External baselines are: the top scoring submission to DSL-ML, the top scoring submission to MADAR 26, and the top scoring *neural* submission to MADAR 26.

Models	Accuracy
IDIOLEX	31%
Fine-tuned IDIOLEX	36%
Fine-tuned IDIOLEX + Lexical	**38%**
Fine-tuned BERT	28%
Centroid Clustering w/ BERT	16%
Centroid Clustering w/ E5	27%

Table 2: Sentence-wise open-set authorship attribution exact match accuracy on the PAN 19 benchmark.

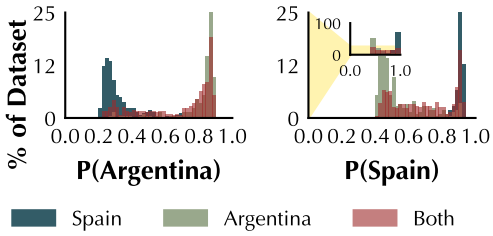

Figure 3: Non-fine-tuned IDIOLEX classification model's likelihood distribution over samples on multi-label Spanish DID.

racy (EM) on DSL-ML and MADAR 26. For DSL-ML, our models outperform both the winning VarDial 2024 submission by (Sälevä and Palen-Michel, 2024) and all baselines. Fine-tuning and lexical ensembling provide marginal benefits, demonstrating the strength of IDIOLEX representations despite the domain mismatch. Figure 3 shows the predicted likelihood distributions from the base IDIOLEX model for sentences labeled by annotators as Argentinian, Spanish, and both. A likelihood over the threshold of 0.5 results in the sentence being assigned that label. Single-label samples receive high likelihoods while multi-label samples exhibit broader distributions, reflecting mixed dialectal cues. This behavior is consistent with our goals of encoding dialect-ness in a continuous manner to allow for nuance in analysis.

On Arabic single-label DID, IDIOLEX achieves an accuracy of 0.66, outperforming Samih et al. (2019) by 7 points (0.59) and coming within 1 point of Abu Kwaik and Saad (2019) (0.67), the best reported result from the shared task. Samih et al. (2019)'s system, although neural, incorporates explicit heuristics such as named regional entities, whereas IDIOLEX relies purely on learned representations. In contrast, Abu Kwaik and Saad (2019) uses a non-neural, feature-engineered approach tailored specifically to this task, which likely contributes to its strong performance. Despite this, IDIOLEX, as a general-purpose representation framework evaluated across multiple tasks, achieves near state-of-the-art performance without task-specific engineering, highlighting the strength and generality of its learned representations.

4.2 Authorship Attribution

AA is another task that relies on stylistic patterns rather than semantic content. To assess stylistic variation independent of dialect, we include the PAN 2019 Spanish authorship attribution dataset (Kestemont et al., 2019), an open-set benchmark. This benchmark specifically examines cross-domain capabilities: the test set data is all in a single domain that is not present during training. The original Spanish PAN 19 shared task was conducted on the story level, so we omit submissions as a baseline since we are predicting on a sentence-level. Sentence-level attribution is particularly challenging due to semantic constraints on feature realization and feature sparsity in short texts.

Since there are <UNK> author labels, we assign the <UNK> class label to any samples where the difference in likelihood between the top two labels is below a threshold, which was calculated per-problem via hyperparameter search over a development set (see Appendix B.2.1).

Table 2 reports that IDIOLEX models in all training formats outperform baselines, achieving 10% absolute improvement over the strongest compari-

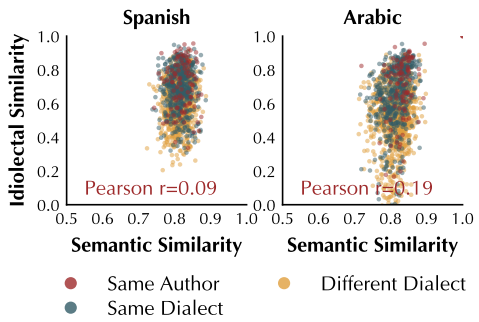

Figure 4: IDIOLEX-based stylistic similarity scores of a random sample of the withheld test set of the Reddit corpus plotted against Multilingual-E5 semantic similarity scores (Wang et al., 2024). We see a low, but positive correlation.

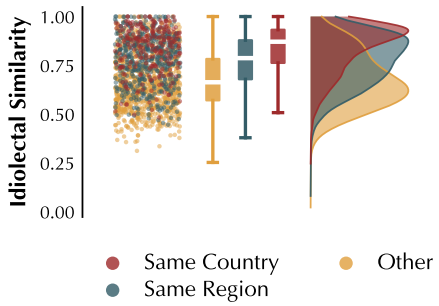

Figure 5: Stylistic similarity on the MADAR-26 dataset calculated via the Arabic IDIOLEX model on sentences with identical semantic content. We see clear differentiation by dialectal proximity.

son model. Since PAN 19 is a cross-domain benchmark, these results demonstrate that IDIOLEX captures stylistic linguistic features that generalize beyond semantic and topical domains.

5 Disentangling Idiolect from Semantics

To evaluate whether the learned embeddings reduce semantic interference, per Section 2, we examine their relationship to traditional semantic representations via comparison to idiolectal similarity scores computed by calculating the cosine similarity of representations.

5.1 Relationship to Semantic Similarity

To assess the degree to which IDIOLEX representations are influenced by semantic content, we compare the idiolectal similarity scores with semantic similarity scores computed using representations generated by Multilingual-E5 (Wang et al., 2024), one of the highest performing open-source semantic embedding models according to the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023).

In Figure 4, we plot stylistic similarity against semantic similarity for a random sample of sentence pairs from the Reddit test set in both Spanish and Arabic. Each point corresponds to a pair of sentences, colored by relational category (same author, same dialect, different dialect). For both languages, the correlation between semantic and idiolectal similarity is positive but weak (Pearson $\rho = 0.09$ for Spanish and $\rho = 0.19$ for Arabic). This indicates that while stylistic similarity is not fully independent of semantic similarity, the two are only loosely coupled. This is expected in any true idiolectal representation because some linguistic features are tied to semantic content (e.g., double negation is dependent the presence of a negative). Still, this is not usually the case, hence the low magnitude of correlation.

5.2 Controlling for Semantics

To more directly test whether IDIOLEX models capture dialectal variation independent of meaning, we utilize the MADAR 26 dataset (Bouamor et al., 2018), which contains parallel sentences across dialects. Because the sentences are semantically equivalent translations, any differences in similarity must arise from linguistic form rather than content.

In Figure 5, we plot idiolectal similarity scores on the MADAR 26 dataset, calculated on parallel samples, grouped by geographical proximity: same country (but different city), same dialectal region, or different dialectal region. For this analysis, MSA is considered to exist in its own dialectal region.

We observe a clear gradient where geographic proximity indicates higher average stylistic similarity. The differences between groups are statistically significant ($p < 0.005$). Because semantic content is held constant, this pattern provides strong evidence that the IDIOLEX framework results in models that encode dialectal structure in a content-independent manner, to the point of capturing fine-grained geographic variations.

We do not claim complete disentanglement of style and semantics; as discussed in Section 2, linguistic structure and semantic content depend on one another to various degrees. Rather, our objective is a shift in emphasis: IDIOLEX representations prioritize linguistic patterns over topicality. The empirical evidence suggests that this shift is achieved in practice.

6 Applications in LLM Post-Training

Because IDIOLEX provides a continuous measure of idiolectal similarity, it offers a possible objective for aligning LLM outputs when sensitivity to stylistic and dialectal variation is desired. As a proof of concept, we incorporate idiolectal representations into LLM post-training with two instruction-tuned LLMs: Llama 3.1 8B,[5] which is multilingual (Grattafiori et al., 2024), and Allam 7B,[6] which is pretrained on English and Arabic only (Bari et al.). We show that this leads to improved dialectal alignment without sacrificing fluency on the AMIYA shared task test set, using ADI2 dialectal adherence via the Al-Qasida evaluation framework (Robinson et al., 2025) and English → Dialectal Arabic (DA) translation quality (ChrF++) on MADAR.

6.1 Supervised Fine-Tuning with IDIOLEX

Standard supervised fine-tuning (SFT) optimizes cross-entropy loss, which does not explicitly encourage dialectal fidelity. As a result, models trained with standard SFT may produce fluent outputs that lack strong dialectal alignment. We find that with standard SFT, models that demonstrate the highest dialectal alignment ((Das Gollapalli et al., 2026) in Table 3) exhibit substantial degradation in fluency, as measured by ChrF++, highlighting a trade-off between alignment and generation quality.

To address this, we augment SFT with an IDIOLEX-based embedding alignment objective. For each training sample, we precompute the IDIOLEX embedding $\mathbf{e}_i$ of the ground-truth response. During training, we pool the LLM's hidden states $\bar{\mathbf{h}}_i$ over the generated tokens and project them into the IDIOLEX embedding space via a learned projection head g_θ. The resulting loss then combines standard cross-entropy with a cosine similarity term:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha \left(1 - \cos(g_\theta(\bar{\mathbf{h}}_i), \mathbf{e}_i)\right)$$

where $\alpha = 0.5$. Intuitively, this ensures that the same idiolectal information encoded by IDIOLEX representations is recoverable from the hidden state of the LLM. Following (Gaber et al., 2026), we construct instruction–response pairs by augmenting monolingual DA text from the AMIYA closed-data track using GPT 5o Mini[7] to form instruction-answer pairs (Robinson et al., 2026). Unlike prior work, we train on a relatively small dataset (approximately 50k pairs). Additional training details, including information on training data, are provided in Appendix B.2.3.

6.2 Alignment Benefits

Across all models in Table 3, using SFT with IDIOLEX consistently improves dialectal alignment while maintaining or improving translation quality. This effect is particularly strong for Moroccan, Egyptian, and Palestinian Arabic. For example, (Das Gollapalli et al., 2026) achieves high ADI2 scores but suffers from low ChrF++, while (Mutal et al., 2026) attains better translation performance at the cost of weaker alignment. In contrast, Allam 7B fine-tuned with IDIOLEX achieves both high dialectal alignment and competitive translation performance, despite being trained on fewer than 50k samples compared to the 2M used by (Mutal et al., 2026).

Performance varies with data availability. Moroccan and Egyptian Arabic have the largest training sets (10,612 and 8,317 monolingual samples, respectively), and correspondingly show the strongest gains. Palestinian Arabic, by contrast, has very limited data (743 samples), both in our setup and in the shared task. Prior work ((Gaber et al., 2026)) mitigates this by incorporating Jordanian data, which is dialectally similar. Although we also include Jordanian samples (7,217) without explicitly labeling them as Palestinian, our model still achieves strong performance on Palestinian. Our strong performance in Palestinian across metrics demonstrates that our model was able to align well dialectally to Palestinian despite the data scarcity – suggesting that IDIOLEX benefits more from coverage of related idiolects and dialects than models that are trained with discrete dialect labels.

Finally, unlike AMIYA submissions, which may fine-tune separate models for each dialect, our models are trained once without dialect-specific specialization. Our strong performance across dialects indicates that IDIOLEX with SFT not only improves dialectal alignment, but also enables models to adapt their output dialect appropriately based on the input.

[5] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[6] https://huggingface.co/humain-ai/ALLaM-7B-Instruct-preview

[7] gpt-5-mini-2025-08-07

Model	ADI2					ChrF++				
	mar	**egy**	**pse**	**syr**	**sau**	**mar**	**egy**	**pse**	**syr**	**sau**
Allam + IDIOLEX SFT	0.54	**0.48**	**0.14**	0.12	**0.11**	**33.2**	**36.4**	**36.1**	34.6	36.7
Allam + SFT	0.49	0.45	0.12	**0.13**	**0.11**	31.8	33.3	34.2	**35.1**	**37.3**
Allam	0.26	0.28	0.05	0.07	0.04	30.7	33.3	34.3	33.5	35.4
Llama 3.1+ IDIOLEX SFT	**0.60**	0.29	0.12	0.10	0.07	27.0	25.4	27.1	26.4	29.5
Llama 3.1 + SFT	0.57	0.27	0.11	0.08	0.05	22.0	22.4	25.4	23.9	28.2
Llama 3.1	0.09	0.12	0.02	0.02	0.01	21.1	20.2	23.9	21.0	29.0
(Gaber et al., 2026)	0.57	0.45	0.10	0.18	0.10	29.2	39.5	38.4	33.2	42.9
(Das Gollapalli et al., 2026)	0.70	0.63	0.12	0.10	0.07	4.6	5.5	5.8	6.0	7.4
(Mutal et al., 2026)	0.38	0.15	0.07	0.22	0.05	34.5	31.0	38.3	17.7	38.2

Table 3: Dialectalness (ADI2) and translation performance (ChrF++) scores on the AMIYA shared task evaluation set for Moroccan (**mar**), Egyptian (**egy**), Palestinian (**pse**), Syrian (**syr**), and Saudi (**sau**) Arabic. Our models, particularly Allam + IDIOLEX SFT, balance style and translation quality across dialects, as compared to reported scores for top submissions for the AMIYA shared task (Robinson et al., 2026) (bottom three rows). Best scores from our experiments are bolded, while best reported scores from the AMIYA shared task are underlined.

7 Related Work

Idiolectal and stylistic variation has been studied across authorship attribution, dialect identification, stylistic representation learning, and personalized language modeling. Our work builds on these lines by learning continuous, reusable sentence representations that capture stylistic and dialectal variation across tasks.

Authorship Attribution & Dialect Identification Authorship attribution (AA) models individual writing style through supervised or contrastive objectives for applications such as forensic linguistics (Stamatatos et al., 1999; Halvani et al., 2019; Bischoff et al., 2020; Fabien et al., 2020; Rivera-Soto et al., 2021; Huertas-Tato et al., 2025), while dialect identification (DID) focuses on linguistic variation across regions and communities for uses in text classification (Tiedemann and Ljubešić, 2012; Lui and Cook, 2013; Bouamor et al., 2018). Both tasks rely on an overlapping set of stylistic and linguistic signals and are typically formulated as discrete classification problems, but are rarely approached in conjunction with one another. In contrast, we leverage authorship and dialect structure as weak supervision to learn *continuous representations* rather than predictions.

Style Representations Standard sentence embeddings are optimized primarily for semantic similarity and have limited sensitivity to stylistic variation (Reimers and Gurevych, 2019; Krasnowska-Kieraś and Wróblewska, 2019; Wang et al., 2024). Recent work has proposed learning style-aware representations, including CISR (Wegmann et al., 2022), StyleDistance (Patel et al., 2025), and approaches that incorporate explicit linguistic features such as LISA vectors (Patel et al., 2023). Our approach is closely related, but differs by jointly leveraging proximity scores gleaned from authorship and community metadata and feature prediction to learn representations that reflect stylistic similarity.

Personalization and Stylistic Alignment Personalized language modeling and user representations capture individual variation in language use, often entangling stylistic and semantic preferences (Lin et al., 2017; Welch et al., 2020; Kocoń et al., 2021). More recent work explores controlling the stylistic properties of LLM outputs and evaluating dialectal fidelity (Durandard et al., 2025a,b; Robinson et al., 2025). We continue this line of work by using learned stylistic representations as training signals during post-training, enabling alignment with user-level linguistic variation.

8 Conclusion

We introduce IDIOLEX, a framework for idiolectal representation learning that models stylistic and dialectal variation in a continuous manner by combining supervision from proximity scores with LLM-extracted linguistic features. We show that these representations capture meaningful idiolectal structure, generalize across domains, and support multiple downstream uses like DID and AA. They exhibit low correlation with semantic similarity and provide an effective training objective for dialectal alignment in post-training. More broadly, IDIOLEX highlights the value of treating idiolect as a graded and compositional phenomenon through a shift from discrete labels to continuous representations. This approach provides a practical framework for studying and building for language variation.

Limitations

Our approach relies on multiple sources of noisy supervision, including automatically generated linguistic features and assumptions about stylistic similarity derived from relative comparisons. These signals may therefore reflect biases or inaccuracies present in the underlying models or data sources. As a result, the learned representations may emphasize certain surface-level or model-salient features of language variation while underrepresenting others that are linguistically meaningful but less easily captured through automated extraction.

This work also does not directly address perceptions on dialect production. In particular, dialectal and stylistic variation is not uniformly perceived or valued across speakers or communities, and the representation of such variation in learned embeddings does not account for speaker intent, context, or social meaning. While idiolectal representations may enable downstream systems to model or reproduce idiolectal patterns, it remains an open question whether users across cultures consistently desire language technologies to generate content in their own dialect. Evaluating these considerations requires user-centered studies that fall outside the scope of this work.

Our empirical evaluation is limited to a small number of languages, datasets, and task formulations, and focuses primarily on offline benchmarks. Consequently, the extent to which IDIOLEX representations generalize to other situations is not definitive. In particular, performance may differ in low-resource languages, in code-switched or mixed-dialect settings, or in domains with highly context-dependent stylistic norms.

Finally, while we demonstrate that idiolectal representations can support several downstream analyses and post-training applications, we do not evaluate their impact in end-to-end interactive systems. The integration of such representations into real-world language models raises additional challenges related to controllability, robustness, and unintended amplification of stylistic patterns. Future work is needed to assess how idiolectal representations interact with user preferences, safety constraints, and deployment considerations in practical settings.

Ethical Considerations

All experiments in this work are conducted using existing, publicly available datasets and do not involve direct interaction with human subjects. Datasets derived from online sources such as Reddit may contain informal or inconsistently labeled dialectal data, without expert linguistic validation. Consequently, both models evaluated using IDIOLEX representations trained on this data may exhibit uneven performance across language varieties.

While modeling dialectal and stylistic variation can enable more inclusive and expressive language technologies, it also raises questions about appropriateness and user preference. Speakers may not always want language models to generate text in their own idiolect, particularly in contexts where standard or neutral varieties are expected, or where incorrect or exaggerated dialectal usage could be perceived as inauthentic or undesirable. This work does not evaluate user attitudes toward dialectal generation, and we do not assume that improved modeling of dialectal variation necessarily implies that models should produce such variation by default.

We view these considerations as emphasizing the importance of careful evaluation, user control, and contextual sensitivity when incorporating dialectal representations into downstream systems. Further work is needed to better understand how such representations interact with user preferences and social norms in deployed language technologies.

References

Gheith A Abandah, Moath R Khaleel, Iyad F Jafar, Mohammad R Abdel-Majeed, Yousef H Hamdan, Ashraf E Suyyagh, Asma A Abdel-Karim, and Shorouq M AlAwawdeh. 2025. Jordanian Arabic to Modern Standard Arabic translation using a large model tuned on a purpose-built dataset and synthetic error injection. *Jordanian Journal of Computers & Information Technology*, 11(3).

Kathrein Abu Kwaik and Motaz Saad. 2019. ArbDialectID at MADAR shared task 1: Language modelling and ensemble learning for fine grained Arabic dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 254–258, Florence, Italy. Association for Computational Linguistics.

Tareq Al-Moslmi, Mohammed Albared, Adel Al-Shabi, Nazlia Omar, and Salwani Abdullah. 2018. Arabic senti-lexicon: Constructing publicly available language resources for arabic sentiment analysis. *Journal of Information Science*, 44(3):345–362.

Naif Alanazi, Mohammed Al-Batineh, and Hussein Abu-Rayyash. 2025. Saudial: The Saudi Arabic di-

alects game localization dataset. *Data in Brief*, page 111906.

Keith Allan. 2013. What is common ground? In *Perspectives on linguistic pragmatics*, pages 285–310. Springer.

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibaee, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, and 25 others. 2025. Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32871–32894, Vienna, Austria. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

M Saiful Bari, Yazeed Alnumay, Norah A Alzahrani, Nouf M Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z Alsubaie, Hassan A Alahmed, Ghadah Alabduljabbar, and 1 others. ALLaM: Large language models for Arabic and English. In *The Thirteenth International Conference on Learning Representations*.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.

Gasper Begus, Maksymilian Dabkowski, and Ryan Rhodes. 2025. Large linguistic models: Investigating LLMs' metalinguistic abilities. *IEEE Transactions on Artificial Intelligence*.

Sebastian Bischoff, Niklas Deckers, Marcel Schliebs, Ben Thies, Matthias Hagen, Efstathios Stamatatos, Benno Stein, and Martin Potthast. 2020. The importance of suppressing domain style in authorship analysis. *Preprint*, arXiv:2005.14714.

Bernard Bloch. 1948. A set of postulates for phonemic analysis. *Language*, 24(1):3–46.

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).

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.

Abdelaziz Bounhar. 2025. Atlaset: A Moroccan Darija text corpus. `https://huggingface.co/datasets/atlasia/Atlaset/`.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8359–8388. PMLR.

Adrian-Gabriel Chifu, Goran Glavaš, Radu Tudor Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 1–15, Mexico City, Mexico. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Sujatha Das Gollapalli, Mouad Hakam, Mingzhe Du, and See-Kiong Ng. 2026. NUS-IDS at AMIYA/VarDial 2026: Improving Arabic dialectness in LLMs with reinforcement learning. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.

Yi Dong, Zhilin Wang, Makesh Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023. SteerLM: Attribute conditioned SFT as an (user-steerable) alternative to RLHF. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11275–11288, Singapore. Association for Computational Linguistics.

Noé Durandard, Saurabh Dhawan, and Thierry Poibeau. 2025a. Language style matching in large language models. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 620–636, Avignon, France. Association for Computational Linguistics.

Noé Durandard, Saurabh Dhawan, and Thierry Poibeau. 2025b. LLMs stick to the point, humans to style: Semantic and stylistic alignment in human and LLM communication. In *Proceedings of the 26th Annual*

Meeting of the Special Interest Group on Discourse and Dialogue, pages 206–213, Avignon, France. Association for Computational Linguistics.

Mahmoud El-Haj. 2020. Habibi - a multi dialect multi national Arabic song lyrics corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1318–1326, Marseille, France. European Language Resources Association.

Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA : BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, Florida, USA. Association for Computational Linguistics.

Ivar Frisch and Mario Giulianelli. 2024. LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 102–111, St. Julians, Malta. Association for Computational Linguistics.

Rana Gaber, Yara Allam, Serag Amin, Ranwa Aly, and Bashar Alhafni. 2026. MBZUAI at AMIYA shared task 2026: Adapting open-source LLMs for dialectal Arabic. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models.

Oren Halvani, Christian Winter, and Lukas Graner. 2019. Assessing the applicability of authorship verification methods. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, ARES '19, New York, NY, USA. Association for Computing Machinery.

Javier Huertas-Tato, Alejandro Martin, and David Camacho. 2025. PART: Pre-trained authorship representation transformer. *Preprint*, arXiv:2209.15373.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast, and Benno Stein. 2019. PAN19 authorship analysis: Cross-domain authorship attribution.

Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, Sue Hyun Park, Hyeonbin Hwang, Jinkyung Jo, Hyowon Cho, Haebin Shin, Seongyun Lee, Hanseok Oh, Noah Lee, Namgyu Ho, and 13 others. 2025. The BiGGen bench: A principled benchmark for fine-grained evaluation of language models with language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5877–5919, Albuquerque, New Mexico. Association for Computational Linguistics.

Jan Kocoń, Marcin Gruza, Julita Bielaniewicz, Damian Grimling, Kamil Kanclerz, Piotr Miłkowski, and Przemysław Kazienko. 2021. Learning personal human biases and representations for subjective tasks in natural language processing. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1168–1173. IEEE.

Katarzyna Krasnowska-Kieraś and Alina Wróblewska. 2019. Empirical linguistic study of sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5729–5739, Florence, Italy. Association for Computational Linguistics.

Javier De la Rosa, Eduardo G. Ponferrada, Manu Romero, Paulo Villegas, Pablo González de Prado Salas, and María Grandury. 2022. BERTIN: Efficient pre-training of a Spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wynter, Xun Wang, Si-Qing Chen, Michael J. Wooldridge, Janet B. Pierrehumbert, and Furu Wei. 2025. Assessing dialect fairness and robustness of large language models in reasoning tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6317–6342, Vienna, Austria. Association for Computational Linguistics.

Zih-Wei Lin, Tzu-Wei Sung, Hung-Yi Lee, and Lin-Shan Lee. 2017. Personalized word representations carrying personalized semantics learned from social network posts. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 533–540. IEEE.

John M. Lipski. 2012. *Geographical and Social Varieties of Spanish: An Overview*, chapter 1. John Wiley & Sons, Ltd.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is your code generated by ChatGPT really correct? Rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*, volume 36, pages 21558–21572. Curran Associates, Inc.

Marco Lui and Paul Cook. 2013. Classifying English documents by national dialect. In *Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013)*, pages 5–15, Brisbane, Australia.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Jonathan Mutal, Perla Al Almaoui, Simon Hengchen, and Pierrette Bouillon. 2026. AladdinFTI @ AMIYA - three wishes for Arabic NLP: Fidelity, diglossia, and multidialectal generation. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.

Jaye Nias, Dwayne Brandon, Thane Douglass, Ashley Haynes, Niara Patterson, and Krystal Cooper. 2025. From culture to code: An intersectional analysis of AAVE slang in large language models. In *Proceedings of the 5th Biennial African Human Computer Interaction Conference*, pages 141–150.

Ajay Patel, Delip Rao, Ansh Kothary, Kathleen McKeown, and Chris Callison-Burch. 2023. Learning interpretable style embeddings via prompting LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15270–15290, Singapore. Association for Computational Linguistics.

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2025. StyleDistance: Stronger content-independent style embeddings with synthetic parallel examples. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8662–8685, Albuquerque, New Mexico. 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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nathaniel R. Robinson, Shahd Abdelmoneim, Anjali Kantharuban, Otba Alsboul, Salima Lamsiyah, Kelly Marchisio, and Kenton Murray. 2026. AMIYA shared task: Arabic modeling in your accent at VarDial 2026. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–17, Rabat, Morocco. Association for Computational Linguistics.

Nathaniel Romney Robinson, Shahd Abdelmoneim, Kelly Marchisio, and Sebastian Ruder. 2025. ALQASIDA: Analyzing LLM quality and accuracy systematically in dialectal Arabic. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22048–22065, Vienna, Austria. Association for Computational Linguistics.

Karin Ryding and David Wilmsen. 2021. *The Cambridge Handbook of Arabic Linguistics*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Jonne Sälevä and Chester Palen-Michel. 2024. Brandeis at VarDial 2024 DSL-ML shared task: Multilingual models, simple baselines and data augmentation. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 241–251, Mexico City, Mexico. Association for Computational Linguistics.

Younes Samih, Hamdy Mubarak, Ahmed Abdelali, Mohammed Attia, Mohamed Eldesouki, and Kareem Darwish. 2019. QC-GO submission for MADAR shared task: Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 290–294, Florence, Italy. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin

Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 1999. Automatic authorship attribution. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 158–164, Bergen, Norway. Association for Computational Linguistics.

Taghreed Tarmom, William Teahan, Eric Atwell, and Mohammad Ammar Alsalka. 2020. Compression versus traditional machine learning classifiers to detect code-switching in varieties and dialects: Arabic as a case study. *Natural Language Engineering*, 26(6):663–676.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634, Mumbai, India. The COLING 2012 Organizing Committee.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.

Charles Welch, Jonathan K. Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Compositional demographic word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4076–4089, Online. Association for Computational Linguistics.

Angie Williams, Peter Garrett, and Nikolas Coupland. 2008. Dialect recognition. In *Handbook of Perceptual Dialectology*, pages 345–358. John Benjamins Publishing Company.

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.

Jian Zhu and David Jurgens. 2021. Idiosyncratic but not arbitrary: Learning idiolects in online registers reveals distinctive yet consistent individual styles. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 279–297, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Language	Subreddit	Dialect Label	# Authors	# Sentences Pre-Train	Train	Dev	Test
Arabic	r/algeria	Algerian	3728	24003	5977	149	182
	r/Egypt	Egyptian	21117	536429	16188	209	129
	r/Iraq	Iraqi	2542	21595	7810	90	224
	r/jordan	Jordanian	14745	427230	15159	272	204
	r/Kuwait	Kuwaiti	1739	13099	6027	88	228
	r/lebanon	Lebanese	1705	12020	7083	413	227
	r/Libya	Libyan	788	6344	6202	298	238
	r/Morocco	Moroccan	3462	27940	6131	121	243
	r/Oman	Omani	516	1515	1515	106	102
	r/Palestine	Palestinian	530	2016	2016	151	252
	r/qatar	Qatari	402	794	794	93	119
	r/saudiarabia	Saudi	25982	639810	10743	201	463
	r/Sudan	Sudanese	795	9297	9297	327	125
	r/Syria	Syrian	1765	48265	13355	598	394
	r/UAE	Emirati	318	458	458	93	98
	Total		**80134**	**1770815**	**108755**	**3209**	**3228**
Spanish	r/argentina	Argentinian	122963	21532596	41160	399	4607
	r/BOLIVIA	Bolivian	6431	196137	12431	401	742
	r/chile	Chilean	65314	7563763	35592	850	1819
	r/Colombia	Colombian	56421	2166716	19553	1246	159
	r/cuba	Cuban	2846	65180	9191	529	1458
	r/Dominican	Dominican	3892	65488	6966	455	105
	r/ecuador	Ecuadorian	5628	163765	9003	145	244
	r/ElSalvador	El Salvadorian	4119	154321	13914	91	189
	r/guatemala	Guatemalan	10756	404417	15783	1299	621
	r/Honduras	Honduran	6882	281450	19740	364	1989
	r/mexico	Mexican	151978	8729427	50212	792	1500
	r/Panama	Panamanian	13608	1016097	17441	459	950
	r/Paraguay	Paraguayan	9845	479757	15455	675	225
	r/PERU	Peruvian	23411	792837	12280	2054	431
	r/spain	Peninsular	18903	395424	7792	262	384
	r/uruguay	Uruguayan	29093	3274405	33568	784	1694
	r/vzla	Venezuelan	25078	2202730	16640	481	2120
	Total		**557168**	**49484510**	**336721**	**11286**	**19237**

Table 4: Data split sizes for Reddit data by dialect.

A Training Data

A.1 Sampling

We collect publicly available Reddit comments using the Pushshift archive for each subreddit listed in Table 4, spanning from each subreddit's inception through December 2024 (Baumgartner et al., 2020). Comments are segmented into sentences by punctuation using regex, and language filtered using fastText language identification. Only sentences classified as the target language are retained.

Users that appear as `[deleted]`, `[removed]`, and `AutoModerator` are thrown out since they do not belong to one consistent user. Any sentences with fewer than 5 words (under the definition of *word* as a space-delineated token) are removed and any comments with fewer than 2 sentences or users with fewer than 2 comments are removed due to constraints around batch construction.

A.2 Splitting

From the original set of authors for each dialect, 10 are randomly selected to be in each of the dev and test splits. This ensures that the dev and test splits are made up of completely unseen authors to avoid contamination with the train set.

Of the remaining authors, all are present in the pre-train set. The train set contains a subset of the pre-train set and is the set for which feature extraction is performed. This set is made up of 200 authors per dialect, or all of the authors in the pre-train set, whichever is lower.

Spanish Binary Features	
Subject Pronouns & Verbal Morphology	
contains explicit subject yo	contains 2pl present suffix ais
contains explicit subject tu	contains 2pl present suffix eis
contains explicit subject usted	contains 2pl present suffix is
contains explicit subject vos	contains voseo present suffix as stressed
contains explicit subject vosotros or vosotras	contains voseo present suffix es stressed
contains explicit subject ustedes	contains voseo imperative form
contains explicit subject nosotros or nosotras	
contains explicit subject ellos or ellas	
Diminutives	
contains diminutive suffix ito or ita	contains diminutive suffix illo or illa
contains diminutive suffix ico or ica	contains diminutive suffix ino or ina
contains diminutive suffix in regional	contains diminutive suffix ete
contains diminutive suffix ingo or inga	contains diminutive suffix iquio or iquia
contains diminutive suffix icho or icha	
Clitic & Orthographic Variation	
contains differential object marking a before animate direct object	contains preverbal object clitic
contains accusative clitic doubling lo a NP	contains enclitic on infinitive
contains accusative clitic doubling la a NP	contains enclitic on gerund
contains accusative clitic doubling los a NP	contains multiple clitics in sequence
contains accusative clitic doubling las a NP	contains se plus object clitic
contains dative clitic doubling le a NP	contains strong pronoun object doubling
contains dative clitic doubling les a NP	contains indirect object strong pronoun doubling
contains inverted question mark	contains inverted exclamation mark
contains all caps word	contains repeated punctuation

Table 5: Spanish linguistic features extracted via LLM for IDIOLEX training. All features are sentence-level binary indicators.

A.3 Feature Extraction

A.3.1 Feature Inventory

We use sets of linguistic features extracted from language-specific dialect resources (Ryding and Wilmsen, 2021; Lipski, 2012). These features focus primarily on morphology/morphosyntax, affixes and functional lexemes, and orthography. A full list of Spanish features can be found in Table 5 and a full list of Arabic features can be found in Table 6.

A.3.2 Prompting

To annotate sentences with the linguistic features they contain, we prompt an LLM (in our case, GPT 5 mini[8]). The LLM is asked to return a JSON dictionary of binary features with the following prompt:

> You are a deterministic feature extractor.
>
> Task: For each feature key, output 1 if the sentence contains an explicit surface cue described by that feature; otherwise, output 0.
>
> Use a recall-oriented policy: if the cue is reasonably present, set 1. Return VALID JSON ONLY with exactly one top-level key: 'features'.
>
> 'features' maps each feature string EXACTLY (copy verbatim) to 0 or 1.
>
> Feature keys: features Sentence: sentence

In cases where the LLM returns an invalid response (i.e. due to safety filtering), we default to marking every feature as 0.

A.4 Batching

Each batch begins with an **anchor** sentence being sampled, which acts as the seed for the rest of the batch. This anchor sentence is randomly sampled from the dataset.

The other sentences in the batch fall into the following categories:

1. *Same Comment*: A different sentence from the same content. $r = 3$

[8] gpt-5-mini-2025-08-07

Arabic Binary Features	
Morphosyntax & Clause Structure	
contains case endings u a i	contains future prefix sa
contains tanwin un an in	contains future particle sawfa
contains dual suffix an or ayn	contains dialectal future marker rah ha ghadi bash
contains 2sg feminine suffix ki	contains progressive marker 3am qa3id jaalis ka bi
contains 2sg feminine dialect suffix sh or ish	contains imperfect prefix bi or ba
contains pronoun hum	contains imperfect prefix ka
contains pronoun hunna	contains imperfect prefix 3a
contains dialect possessive suffix o or u	contains analytic passive in it ta
is verb initial clause	contains internal passive kutiba
is subject initial clause	contains passive prefix ta it et
contains overt present copula	contains relative alladhi forms
contains zero copula equational	contains relative illi yalli elli
Negation, Interrogatives & Discourse Particles	
contains negator laysa	contains yes no particle hal
contains negator lan	contains dialect yes no particle huwwa wash ish
contains negator lam	contains interrogative ayy
contains negator lamma	contains interrogative eh
contains ma verbal negation	contains interrogative shu
contains ma verb sh negation	contains interrogative wesh
contains negator mish or mesh	contains interrogative shinu
contains negator mu or mo	contains discourse wallahi
contains egyptian ayywa	contains discourse ya3ni
contains msa adverb haqqan	
Dialectal Lexical Markers & Orthography	
contains egyptian izaay	contains egyptian demonstrative da di dol
contains egyptian keda	contains levantine demonstrative hada haydi
contains egyptian ba2a	contains maghrebi demonstrative had
contains levantine baddi	contains gulf wain
contains levantine halla	contains gulf negator mo
contains levantine shu	contains gulf yalla
contains levantine demonstrative hada	contains maghrebi barcha
contains maghrebi hshuma	contains maghrebi prefix ka
contains yemeni maish	contains yemeni shill
contains nonstandard qaf gaf	contains nonstandard hamza omission
contains nonstandard alef maqsura variation	contains arabic question mark
contains question mark	contains exclamation mark !
contains repeated punctuation	contains ellipsis
contains tatweel	contains letter lengthening
contains laughter token	

Table 6: Arabic linguistic features extracted via LLM for IDIOLEX training. All features are sentence-level binary indicators.

2. *Same Author*: A sentence from a different comment by the author. $r = 2$
3. *Same Subreddit*: A sentence from the same subreddit/region but written by a different author. $r = 1$
4. *Different Subreddit*: A sentence from a different subreddit/region versus the anchor. $r = 0$

This is done by sampling randomly over the set of sentences from the comment, the set of comments by the same author, the set of authors from the same subreddit, and the set of subreddits.

For training efficiency, we structure the batch such that every sentence has at least one sentence that falls into one of these categories. In practice, to ensure this, we use mini-batches of size 16 where for each anchor x_a, there are 2^{3-n} samples with proximity score $n \in [0, 3]$. In other words, for every sentence, there will be 1 other sentence from the same comment, 2 from the same author but a different comment, 4 from a different author but the same subreddit, and 8 from another subreddit.

B Experimental Details

B.1 Training Under the IDIOLEX Framework

B.1.1 Base Models

We used the the most prolific pre-trained monolingual language model for each language. We focused on models with fewer than 200 M parameters to ensure that the resulting model would be lightweight enough to use as a training objective or reward model. Specifically, for Spanish we use BERTIN[9], a

[9] https://huggingface.co/bertin-project/bertin-roberta-base-spanish

RoBERTa-base model trained on the Spanish component of mC4 (la Rosa et al., 2022). For Arabic, we use AraBERT v2[10], a BERT-base model trained on Arabic news corpora (Antoun et al., 2020). Both models follow the standard BERT-base architecture with 12 transformer layers and a hidden dimensionality of 768.

Tokenization is performed using the HuggingFace tokenizer corresponding to each pretrained model. Inputs are truncated to a maximum length of 512 tokens.

B.1.2 Auxiliary Prediction Heads

To incorporate linguistic supervision, we add two auxiliary heads on top of the encoder.

Feature Prediction A two-layer feedforward network predicts the binary feature vector associated with each sentence representation. The architecture consists of:

$$\text{Linear}(d, 2d) \rightarrow \text{ReLU} \rightarrow \text{Linear}(2d, F)$$

where d is the encoder hidden size and F is the number of linguistic features (different per language). The model is trained using binary cross-entropy with logits.

Contrastive Similarity Represenations For supervised contrastive learning, representations are passed through a projection head:

$$\text{Linear}(d, d) \rightarrow \text{ReLU} \rightarrow \text{Linear}(d, 256).$$

The resulting vectors are ℓ_2 normalized before computing contrastive similarity based on the Jaccard similarity of the feature vectors.

B.1.3 Training Objectives

Training combines three losses.

Margin Ranking Loss (MRL) The margin ranking loss encourages sentences with higher proximity score to be closer in embedding space. The margin λ is linearly increased from 0 to 0.5 during training.

Supervised Contrastive Loss (SCL) We additionally apply a supervised contrastive loss where the similarity between samples is weighted using the Jaccard similarity between their feature vectors.

Feature prediction loss. The feature prediction head is trained using binary cross-entropy on the predicted feature logits.

Final Objective The final training objective is

$$L = (1 - \alpha)L_{rank} + \alpha L_{feat}, \tag{1}$$

where

$$L_{feat} = 0.25L_{BCE} + L_{supcon}.$$

B.1.4 Variance and Decorrelation Regularization

To prevent representation collapse and reduce embedding anisotropy, we apply a VICReg-style regularization term consisting of:

- a variance constraint encouraging the standard deviation of each embedding dimension to exceed 1,
- a decorrelation penalty on off-diagonal covariance entries.

[10] https://huggingface.co/aubmindlab/bert-base-arabertv2

B.1.5 Sentence Representation Construction

Sentence embeddings are obtained from the transformer using a layer-wise attention mechanism similar to that used by COMET (Rei et al., 2020). Let h_l denote the hidden states from layer l. A learned scalar weight w_l is applied to each layer and normalized using a softmax:

$$\tilde{h} = \sum_{l=0}^{L} \alpha_l h_l, \quad \alpha_l = \frac{\exp(w_l)}{\sum_{k=0}^{L} \exp(w_k)}. \tag{2}$$

The resulting token representations are then aggregated using average pooling over all non-padding tokens to obtain a sentence-level representation.

During training, embeddings are mean-centered using a running estimate of the global embedding mean across all distributed workers. After centering, vectors are ℓ_2 normalized.

B.1.6 Training Procedure

Training proceeds in two stages.

Stage 1: Ranking Pretraining The encoder is trained using only the margin ranking loss on the full dataset. This stage uses all available sentences.

Stage 2: Feature-Aware Training Training is then continued on the subset of sentences with feature annotations. During this stage the feature prediction and contrastive losses are added.

B.1.7 Optimization

Models are trained using the Adam optimizer with a learning rate of 1×10^{-5} and linear warmup for 25k steps.

Training uses a batch size of 32 sentences. Supervised contrastive learning uses a temperature $\tau = 0.07$, and Jaccard weights are truncated to the top 5 neighbors.

B.1.8 Training Infrastructure

Training is performed using PyTorch DistributedDataParallel across 4 GPUs. All models are implemented in PyTorch and trained using the HuggingFace Transformers library.

B.1.9 Training Hyperparameters

Table 7 summarizes the primary hyperparameters used for training using the IDIOLEX framework. Unless otherwise specified, the same configuration is used across all languages, with only the underlying pretrained encoder differing.

B.2 Evaluation Protocols

B.2.1 IDIOLEX-based Classification Procedure

To evaluate whether models trained via the IDIOLEX framework can support downstream classification tasks, we fine-tune classification heads on top of the pretrained encoder models. This section describes the training procedure, optimization settings, and implementation details used to produce the classification results reported in Section 4.

Models For each classification task, we initialize a sequence classification model using the same pretrained encoder architecture used in the representation learning stage; that is, AraBERT v2 (Antoun et al., 2020) for Arabic and BERTIN (la Rosa et al., 2022) for Spanish. The encoder weights are loaded directly from the IDIOLEX checkpoint containing pretrained encoder parameters. The IDIOLEX checkpoints contain only encoder weights; the classification layer is newly initialized to match the number of target classes.

The classification head consists of a linear projection from the encoder's pooled representation to the label space. For multi-label tasks (DSL-ML), the model is configured for multi-label classification, whereas single-label (MADAR-26, PAN-19) tasks use standard softmax classification.

Parameter	Value
Hidden size	768
Transformer layers	12
Maximum sequence length	512
Batch size	32
Ranking group size	16
Pretraining epochs	3
Full training epochs	10
Validation frequency	250 steps
Margin ranking margin	$0 \rightarrow 0.5$
Feature loss weight (α)	0.5
Feature BCE weight	0.25
Supervised contrastive temperature τ	0.07
Jaccard top-k positives	5
Feature head hidden size	$2d$
Projection head output size	256
Feature vector dimension	41
Learning rate	1×10^{-5}
Learning rate warmup	25k steps
Early stopping patience	25
GPUs	4
Hours	$\leq$ 48 hrs

Table 7: Training hyperparameters used for IDIOLEX.

Training Training is performed using the AdamW optimizer with linear learning rate decay and warmup. The learning rate schedule follows a linear warmup followed by linear decay over the course of training steps. Unless otherwise stated, the training hyperparameters found in Table 8 are used.

Hyperparameter	Value
Learning Rate	1×10^{-5}
Batch Size	32
Weight Decay	0.01
Warmup Ratio	0.1
Early Stopping Patience	3 epochs
Max Sequence Length	512 tokens

Table 8: Hyperparameters used for classification fine-tuning.

Gradient clipping with a maximum norm of 1.0 is applied to stabilize training. Mixed precision training (FP16) is optionally used when GPU support is available (always in our case). Because MADAR 26 consists of 26 classes, we apply an additional label smoothing of 0.1 to improve performance for that classifier. Training proceeds for a maximum of 25 epochs, with early stopping based on validation macro F1 score.

In tasks that allow unknown classes (e.g., open-set authorship attribution), an additional <UNK> label is assigned to examples whose predicted confidence falls below a tuned threshold. This threshold is determined via a held out set of 50% of the validation data.

We demonstrate the importance of combining both loss components in Table!9. Here, we follow the same training paradigm and hyperparameters, modifying only $\alpha = 0.0$ for feature-based losses only or $\alpha = 1.0$ for MRL only. We see that there is significant performance degration using either of the losses in isolation,

Lexical Ensembling For some experiments, a lexical model is combined with the neural classifier, reflecting performance improvements that past work has found for specific datasets. The lexical component consists of a TF-IDF representation over character n-grams followed by a logistic regression classifier trained using the SAGA solver. The lexical classifier's model settings are reflected in Table 10.

Models	Spanish F1	Spanish EM	Arabic F1	Arabic EM
IDIOLEX MRL Only	0.81	0.49	0.30	0.30
IDIOLEX Features Only	0.79	0.33	0.20	0.21
IDIOLEX	**0.85**	0.62	0.43	0.43
Fine-tuned IDIOLEX	0.84	**0.63**	0.61	0.61
Fine-tuned IDIOLEX + Lexical	**0.85**	**0.63**	0.66	0.66
Fine-tuned Base	0.80	0.59	0.56	0.56
Centroid Clustering Base	0.77	0.57	0.20	0.22
Centroid Clustering E5	0.74	0.54	0.15	0.17
Sälevä and Palen-Michel, 2024	0.82	0.50	–	–
Abu Kwaik and Saad, 2019	–	–	**0.67**	**0.67**
Samih et al., 2019	–	–	0.59	0.59

Table 9: DID F1-score and exact match accuracy (EM) results on Spanish (DSL-ML) and Arabic (MADAR 26). External baselines are: the top scoring submission to DSL-ML, the top scoring submission to MADAR 26, and the top scoring *neural* submission to MADAR 26.

Language	Model	KMeans F1-macro	KMeans Accuracy	Centroid-Based F1-macro	Centroid-Based Accuracy	Retrieval Accuracy*
Arabic (MADAR 26)	BERT	0.11	0.14	0.20	0.22	0.10
	E5	0.04	0.07	0.15	0.17	0.06
Spanish (DSL-ML)	BERT	0.50	0.46	0.77	0.57	0.07
	E5	0.44	0.45	0.74	0.54	0.06
Spanish (PAN 19)	BERT	0.07	0.32	0.14	0.16	0.07
	E5	0.10	0.33	0.21	0.27	0.08

Table 11: Clustering and retrieval-based evaluation of dialectal structure in sentence embedding spaces.

Hyperparameter	Value
Learning Rate	1×10^{-5}
Batch Size	32
Weight Decay	0.01
Warmup Ratio	0.1
Early Stopping Patience	3 epochs
Max Sequence Length	512 tokens

Table 10: Hyperparameters used for the lexical component of the classifier.

The lexical classifier outputs probability distributions over classes, which are linearly combined with the neural model predictions:

$$P = \omega P_{\text{lex}} + (1 - \omega) P_{\text{neural}}$$

The ensemble weight ω is tuned on the validation set to maximize macro F1.

Evaluation Metrics Evaluation follows the protocols described in Section 4. Predictions are evaluated using macro-averaged F1 score and exact match accuracy (important for multi-label tasks, where it evaluates how well the model predicts *all* the correct labels for each item).

B.2.2 Classification Baselines

Closed-set Classification We fine-tune the same monolingual models used in our IDIOLEX training, BERTIN (la Rosa et al., 2022) for Spanish tasks and AraBERTv2 (Antoun et al., 2020) for Arabic tasks, on the three classification datasets used in our experiments: MADAR 26 (Arabic), DSL-ML (Spanish), and PAN 19 (Spanish). To maintain consistency with IDIOLEX's representation-based evaluation, we use the same logistic regression probe as in the IDIOLEX classification setup. Specifically, we extract the [CLS] representations from the fine-tuned baseline models and use them as input features to the logistic regression probe; these constitute the reported baseline results.

Clustering To evaluate how well embedding models capture dialect-specific structure without supervised classification, we employ a set of clustering-based baselines over sentence-level embeddings. We conduct all experiments using BERTIN (la Rosa et al., 2022) for Spanish and AraBERTv2 (Antoun et al., 2020) for Arabic (the same monolingual BERT-based models used in our closed-set baseline), as well as Multilingual E5 (Wang et al., 2024), a state-of-the-art multilingual sentence-embedding model commonly used in recent embedding evaluation work. Using frozen encoder representations, we fit KMeans on the training-set embeddings and then apply the fitted model to the test-set embeddings, assessing the resulting structure using *F1* and *Accuracy*.

Our Clustering-Based evaluation framework consists of three complementary metric categories, each capturing different aspects of dialect separation in the embedding space:

Centroid-Based Classification. We compute class centroids by averaging embeddings from the training set for each dialect. We then classify test samples by computing the Euclidean distance from each test embedding to all training-set centroids and assigning each test sample to the nearest centroid. This provides a simple nearest-centroid baseline that evaluates how well embeddings support geometric separation of dialects. F1-macro and accuracy are computed by comparing the predicted dialect labels (from the nearest centroid) to the true dialect labels for each test sample. We report both F1-macro and accuracy, where higher values indicate better dialect separation in the embedding space.

KMeans Clustering Metrics. We fit KMeans clustering on the training-set embeddings, using the number of unique dialect labels as the number of clusters. We then apply the fitted model to the test-set embeddings and evaluate the clustering quality using F1-macro and accuracy, which measure how well the unsupervised clusters align with the true dialect labels. F1-macro and accuracy are computed by mapping each cluster to its dominant dialect label (the most frequent true label in that cluster) and then comparing the predicted cluster assignments to the true dialect labels. Higher values indicate that the embedding space naturally separates dialects without explicit supervision.

Retrieval Baseline. For each anchor sentence, we sample 3 positive examples (same dialect) and 3 negative examples (different dialects), then rank all 6 candidates by cosine similarity to the anchor. We report the fraction of trials where all positives rank above all negatives (*Accuracy**), which measures the model's ability to distinguish same-dialect from different-dialect samples in a retrieval setting.

Tables 11 presents comprehensive results across all metrics and datasets. The tables show that while both models capture some dialect structure, there is significant variation across languages and metric types, with E5 generally performing better on centroid-based metrics while BERTIN shows competitive performance on retrieval tasks.

B.2.3 Post-Training

We incorporate IDIOLEX into the supervised fine-tuning (SFT) stage of post-training by integrating it into the training objective. By operating on the model's internal representations rather than on the generated text, we avoid issues stemming from length mismatches between ground truth (or model length limits) and generations.

Data The original AMIYA training data is primarily monolingual text data, not instruction-tuning pairs. To ensure our fine-tuning didn't degrade instruction-following abilities, we use a method similar to that by Gaber et al. (2026) to convert these corpora into instruction-response pairs, using a combination of template-based methods and LLM-based augmentation.

We draw from 9 corpora spanning 5 target dialects (Egyptian, Moroccan, Palestinian, Saudi, Syrian) as well as 5 additional dialects (Jordanian, Iraqi, Yemeni, Tunisian, and Kuwaiti):

For parallel corpora containing aligned translations, we create instruction pairs without any LLM calls. Each source sentence is prefixed with a randomly sampled translation instruction template:

1. English → DA: "Translate this to [dialect]: [source]", "Say this in [dialect]: [source]", "How would you say this in [dialect]? [source]"

2. MSA → DA: "Convert this to [dialect]: [source]", "Rewrite this in [dialect]: [source]"

Source	Type	Dialects	Size	Method
MADAR-26 (train) (Bouamor et al., 2018)	Bitext	Multi-Dialectal	2k sentences per	Template
FLORES-200 (Goyal et al., 2022)	Bitext	Multi-Dialectal	1k sentences per	Template
SauDial (Alanazi et al., 2025)	Bitext	Saudi	1k sentences	Template
PALM (train) (Alwajih et al., 2025)	Monolingual	Multi-Dialectal	4k sentences	LLM
HABIBI (El-Haj, 2020)	Monolingual	Multi-Dialectal	525k verses	LLM
EDC (Tarmom et al., 2020)	Monolingual	Egyptian	13k sentences	LLM
Atlaset (Bounhar, 2025)	Monolingual	Moroccan	1.1M tokens	LLM
MASC (Al-Moslmi et al., 2018)	Monolingual	Multi-Dialectal	6k sentences	LLM
JODA (Abandah et al., 2025)	Monolingual	Jordanian	50k sentences	LLM

Table 12: AMIYA data used for post-training and how it was augmented to make instruction-following data.

The dialectal translation serves as the ground-truth response. This yields high-quality, naturally dialectal training pairs.

For monolingual corpora (containing only dialectal text without paired translations), we use a two-step LLM augmentation process to generate full instruction-response pairs. First, for each dialectal corpus text, we prompt GPT-5-mini to generate a dialectal question that the text could naturally answer.

> You are a native speaker of [dialect]. Below is a text in [dialect]. Write a short question (1 sentence) in [dialect] that someone might ask a chatbot, where this text would be a natural answer. Rules: Write ONLY the question. The question must be in [dialect], NOT Modern Standard Arabic. Use dialectal question words like: [dialect-specific examples]. The question should be about [randomly sampled topic].

Topics are sampled from 8 categories aligned with the AL-QASIDA evaluation: food/recipes, proverbs, cultural traditions, geography, history, dialect expressions, nature, and arts. Dialect-specific question words are provided to encourage authentic dialectal phrasing.

A naive approach would use the original corpus text directly as the ground-truth response. However, we found this produced poor SFT results: corpus fragments are typically 3-5 words (e.g., song lyrics, social media posts), and models trained on these learned to produce extremely short, often incoherent responses. We instead prompt the LLM to generate a full conversational answer incorporating the corpus text:

> You are a helpful chatbot that ONLY speaks [dialect]. A user asked you: [generated question] Write a helpful, detailed answer (3-6 sentences) in [dialect]. You MUST incorporate this phrase naturally in your answer: [corpus text]

This produces training pairs where the question is dialectal, the answer is a full multi-sentence response in dialect, and the original human-written dialectal text is preserved within the answer, ensuring the ground truth contains authentic dialectal vocabulary and phrasing rather than just LLM-generated approximations.

Due to computational constraints, we randomly sample at most 5k sentences per dataset for our monolingual training data, resulting in about 39k instruction-answer pairs total plus 10k translation pairs.

Architecture Our approach adds two components to a standard LoRA-based SFT. First, we use an encoder trained under the IDIOLEX framework on the language of interest. The parameters of this model stay frozen during SFT. To reduce per-step overhead, we pre-compute the IDIOLEX embeddings of the ground truth data and store it in a cache. Then, we introduce a small two-layer projection head g_θ that maps from the LLM's hidden dimension to the IDIOLEX embedding dimension. The output of this projection head is L2-normalized in the same manner as the IDIOLEX representations are to ensure they fall in the same normalized embedding space.

For each training sample i with input tokens $\mathbf{x}_i$ and response starting at position s_i, the combined loss is:

$$\mathcal{L}_i = \mathcal{L}_{\text{CE}}(\mathbf{x}_i, s_i) + \alpha \cdot \mathcal{L}_{\text{A}}(\mathbf{x}_i, s_i, \mathbf{e}_i)$$

The cross-entropy loss ($\mathcal{L}_{\text{CE}}$) is computed only over response tokens. The alignment loss ($\mathcal{L}_{\text{A}}$) pools the LLM's last-layer hidden states over response tokens and projects to IdioleX space:

$$\mathcal{L}_{\text{emb}} = 1 - \text{cosine_similarity}\left(g_\theta(\mathbf{h}_i), \mathbf{e}_i\right)$$

where $\mathbf{h}_t$ is the mean-pooled hidden state at position t from the last transformer layer.

Experimental Details We use $\alpha = 0.5$, LoRA rank 32 on all linear layers, a learning rate of 2×10^{-4} with cosine annealing, and train for 1 epoch. The projection head adds approximately 6M parameters (for a 4096-dimensional LLM projecting to 768-dimensional IDIOLEX embedding space).

IDIOLEX embeddings are precomputed once on a single GPU before distributed training begins, cached to disk, and loaded by all workers. The LLM with LoRA adapters and the projection head are wrapped into a single ‘nn.Module‘ for compatibility with DeepSpeed ZeRO-2, which requires a single model instance. For efficiency, training was done with gradient checkpointing and mixed precision (bfloat16).