Language Proficiency as a Recoverable Dimension in Multilingual LLM Embeddings

Rodrigo Wilkens
University of Exeter
r.wilkens@exeter.ac.uk

Abstract

Understanding whether proficiency is encoded as structured knowledge rather than inferred from surface correlates is critical for interpreting and applying LLMs in educational contexts. We investigate whether multilingual large language model (LLM) embeddings encode language proficiency as a structured recoverable dimension rather than merely supporting predictive classification. Using the UniversalCEFR benchmark, which spans 13 languages and the full proficiency range from A1 to C2, we evaluate the frozen LLM embedding space in two complementary ways. First, we test whether proficiency levels can be predicted directly from frozen embeddings across languages and model variants. The results show that embeddings without task-specific fine-tuning consistently support CEFR classification. Variation in results is strongly associated with the amount of annotated data and language family, suggesting that data availability and cross-linguistic structure matter more than architectural differences. Second, we examine how CEFR levels are organized inside embedding space. We find that texts from lower to higher proficiency levels align along a consistent ordered direction, with higher levels systematically positioned further along this gradient. Distances between levels increase proportionally to their ordinal gap (e.g., A1 vs. C2 is farther apart than B1 vs. B2), indicating a "consistent continuous gradient with overlapping adjacent levels rather than arbitrary clusters. Together, these findings show that CEFR is not only predictable from multilingual LLM embeddings but is also internally structured as an ordered representational dimension.

Keywords: Language Proficiency Assessment, Representation Probing, Multilingual LLM Embeddings, CEFR Modeling

1. Introduction

Language proficiency assessment plays a central role in education, large-scale testing, and increasingly in natural language processing (NLP) applications that support personalized learning and automated feedback (Shermis and Burstein, 2003; Yannakoudakis et al., 2011; Ke and Ng, 2019). Within this context, the Common European Framework of Reference for Languages (CEFR) has become a widely adopted standard for describing learner proficiency across languages and educational systems (of Europe, 2020; of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, 2001). Automated essay scoring (AES) and automatic readability assessment (ARA) constitute two major strands of computational research concerned with modeling language proficiency. AES focuses on the assessment of learner-produced texts, aiming to evaluate writing quality and proficiency level (Page, 1966; Attali and Burstein, 2006; Yannakoudakis et al., 2011; Ke and Ng, 2019). In contrast, ARA addresses the estimation of text difficulty and its suitability for readers at different proficiency stages (Collins-Thompson, 2014; Vajjala, 2022). Despite their distinct theoretical orientations (i.e., production in the case of AES and reception in the case of ARA), both lines of work have historically operationalized proficiency through observable properties of text. Early approaches relied on linguistic features such as lexical diversity, syntactic complexity, and cohesion measures (Zesch et al., 2015; Ke and Ng, 2019).

Deep learning marked a shift from explicit feature engineering to representation learning. Neural models based on CNNs and LSTMs demonstrated improved ability to capture semantic and discourse-level patterns (Taghipour and Ng, 2016; Alikaniotis et al., 2016; Dong and Zhang, 2016), but struggled with long-range dependencies and scalability. Transformer-based encoders, such as BERT (Devlin et al., 2019), have since become dominant in NLP, enabling fine-tuning for proficiency level prediction tasks (Mayfield and Black, 2020; Rodriguez et al., 2019). Hybrid models that combine contextual embeddings with handcrafted linguistic features have achieved further gains in both AES and ARA (Li et al., 2022; Faseeh et al., 2024; Liu and Lee, 2023; Wilkens et al., 2024).

Generative decoder-based large language models (LLMs) have reshaped the landscape of NLP. Instead of fine-tuning encoder architectures, researchers increasingly rely on prompt engineering, in-context learning, and instruction tuning of autoregressive models (Brown et al., 2020; Chung et al., 2024). This progression reflects a steady expansion of representational capacity and methodological flexibility in automated assessment. In particular, it is still unknown whether the CEFR levels are internally encoded in the models (i.e., corre-

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 227–242
11 May 2026.

spond to a coherent latent geometric direction in LLM embedding spaces) or whether the observed predictive success stems from superficial correlations (e.g., text length and lexical frequency). While probing studies have shown that linguistic properties can be recovered from contextual embeddings (Tenney et al., 2019; Hewitt and Manning, 2019; Pimentel et al., 2020), it remains unclear how higher-level constructs such as language proficiency are structured within multilingual embedding spaces. From an educational perspective, understanding whether proficiency is internally structured has direct implications. If proficiency corresponds to a recoverable and continuous representational dimension, it enables new forms of model usage beyond classification, including controllable text adaptation, calibrated feedback generation, and alignment between model representations and pedagogical scales. Conversely, if predictions rely primarily on superficial correlates, their interpretability and pedagogical validity remain limited.

This study addresses that gap by investigating whether frozen multilingual LLM embeddings encode recoverable CEFR structure. To achieve this goal, we formulate two research questions:

1. To what extent do frozen multilingual LLM embeddings support CEFR classification across languages and modeling configurations?

2. Does CEFR proficiency correspond to a recoverable latent geometric axis in embedding space that generalizes across languages and task modalities?

Using multilingual CEFR corpora, we conduct predictive evaluation alongside cross-validated geometric analysis of embedding space to investigate whether CEFR proficiency corresponds to a recoverable latent structure. Our findings indicate that while embeddings provide a multilingual proficiency signal, CEFR levels can be interpreted as continuous latent gradients embedded within higher-dimensional linguistic structure. Our findings contribute to the current discussions on whether LLMs capture linguistically valid proficiency constructs or merely exploit surface correlates.

The remainder of the paper reviews related work on automated assessment and representational analysis (Section 2), introduces the predictive and geometric methodology (Section 3), presents our findings addressing multilingual classification and latent proficiency structure (Section 4), and concludes with an analysis of representational implications and future directions (Sections 5-6).

2. Related Work

2.1. Language Proficiency Assessment

Automated Essay Scoring (AES), concerned with modeling learner language production, and automatic readability assessment (ARA), focused on estimating text difficulty for language reception, have historically relied on handcrafted linguistic indicators such as lexical diversity, syntactic complexity, discourse features, and surface statistics (Page, 1966; Attali and Burstein, 2006; Persing et al., 2010; Zesch et al., 2015; Collins-Thompson, 2014; François and Fairon, 2012). In both traditions, proficiency or difficulty has been operationalized through observable linguistic properties derived from text. Cross-lingual readability studies have further demonstrated that complexity modeling varies across languages and typological settings (Vajjala and Rama, 2018), highlighting the interaction between linguistic structure and proficiency prediction. Neural architectures based on CNNs and LSTMs shifted the focus toward representation learning (Alikaniotis et al., 2016; Taghipour and Ng, 2016; Dong and Zhang, 2016), reducing explicit feature engineering while retaining predictive evaluation as the primary objective. Transformer encoders (Devlin et al., 2019) subsequently became dominant, and fine-tuning approaches were widely adopted for AES and CEFR prediction (Rodriguez et al., 2019; Mayfield and Black, 2020). Hybrid architectures combining contextual embeddings with handcrafted linguistic and readability features consistently achieved strong performance, reinforcing the complementarity between structured linguistic metrics and distributed representations (Li et al., 2022; Faseeh et al., 2024; Liu and Lee, 2023; Wilkens et al., 2024). Across these methodological developments, research in both AES and ARA has remained primarily evaluation-driven, emphasizing predictive metrics over the analysis of representational structure.

2.2. Encoder and Decoder Models in Assessment

Transformer-based encoders dominated early LLM-based assessment through supervised fine-tuning (Devlin et al., 2019; Liu et al., 2019), while multilingual variants demonstrated strong cross-lingual transfer capabilities (Conneau et al., 2020). However, empirical evidence shows that multilingual performance varies with data size, domain, and language similarities (Lauscher et al., 2020; Bjerva et al., 2019).

Advances in autoregressive decoder models have shifted attention toward prompt-based and instruction-tuned evaluation (Brown et al., 2020; Chung et al., 2024). Decoder-based scoring can

achieve competitive results without task-specific fine-tuning, challenging the dominance of encoder-centric pipelines (Seßler et al., 2025; Yancey et al., 2023; Mizumoto and Eguchi, 2023; Liew and Tan, 2024). Nevertheless, comparative studies of encoder and decoder configurations largely assess the system's predictive accuracy.

Probing research demonstrates that linguistic properties can be linearly recovered from contextual embeddings (Tenney et al., 2019; Hewitt and Manning, 2019; Pimentel et al., 2020), and geometric analyses reveal that syntactic and semantic structure often corresponds to identifiable directions or subspaces in high-dimensional representation space (Mu and Andreas, 2020). While such studies show that structured linguistic information is embedded in transformer representations, it primarily targets discrete grammatical or lexical properties. Complex, socially defined constructs such as language proficiency, particularly under standardized frameworks like CEFR, have not been examined as latent geometric gradients in multilingual embedding spaces.

3. Methodology

Aiming to investigate whether multilingual LLM embeddings encode CEFR proficiency as recoverable information, we combine predictive evaluation (RQ1) with geometric analysis (RQ2). The former evaluates classification performance across modeling configurations, while the latter examines whether proficiency corresponds to a latent geometric axis in embedding space.

3.1. Predictive Framework

We evaluate whether frozen multilingual LLM embeddings encode sufficient signal to support CEFR classification across languages. This step establishes representational adequacy before examining geometric structure. We focus on EuroLLM (Martins et al., 2025b,a), a multilingual large language model trained with explicit European-language coverage, as it provides controlled proportions of pretraining across languages. It also supports both base (enriched) and instruction-tuned (instruct-enriched) variants.

Using EuroLLM allows us to directly examine whether differences in instruction tuning or pretraining exposure influence CEFR encoding across languages.[1] The comparison isolates the effect of instruction tuning. Otherwise, similar performance would suggest that the proficiency signal emerges during the base model's training. The information encoded in the model has been associated with scaling effects (Brown et al., 2020). Therefore, we evaluate two model sizes: 1.7B and 9B parameters. All models are run in frozen mode, no fine-tuning is performed, and the model input is exclusively the text document without additional instructions. This isolates pretrained representational capacity from task-specific adaptation. We extract embeddings using the HuggingFace library (Wolf et al., 2019) and selecting the representations from the final hidden layer.

To evaluate the separability of the proficiency signal, we train three downstream classifiers to predict the 6-CEFR levels (A1-C2): Logistic Regression (LR), Linear Support Vector Machine (SVM) with a linear kernel, and Multi-layer Perceptron (MLP). LR and SVM are used to test whether the CEFR signal is linearly separable in the embedding space.[2] The MLP introduces controlled non-linearity; thus, a performance improvement suggests that the CEFR signal is present but not linearly organized. These three models were trained using stratified N-fold cross-validation[3]. Models were evaluated using accuracy, macro F1, Quadratic Weighted Kappa (QWK), and Mean Absolute Error (MAE).[4]

Moving beyond machine learning model ranking, we investigate what drives variation in CEFR classification performance across languages. Therefore, we perform a secondary analysis at the level of language-corpus configurations. Concretely, after computing cross-validated performance scores (e.g., Macro F1) for each combination of language, corpus, model variant (Enriched vs Instruct-Enriched), model size (1.7B vs 9B), and downstream classifier, we treat each such configuration as a single data point in a regression analysis. In other words, for every language-corpus pair and modeling setup, we obtain one performance score. These scores constitute the dependent variable in our analysis. We estimate a series of linear regression models using Ordinary Least Squares (OLS). In this context, OLS estimates the average effect of each explanatory variable on classification performance while holding other variables constant. The resulting coefficients can be interpreted as the expected change in performance associated with a

[1] We focus on a single model family to maintain a controlled multilingual training setting, allowing us to isolate representational effects without confounding differences in architecture or training data.

[2] Linear probes are standard in representation analysis because they provide a minimal-capacity test of recoverability (Hewitt and Manning, 2019; Pimentel et al., 2020).

[3] The N is determined by dataset size: 5-fold CV when fewer than 500 instances, 10-fold CV when between 500 and 5,000 instances, and 5-fold CV in the other cases.

[4] QWK is included due to its ordinal sensitivity and established use in AES research (Taghipour and Ng, 2016). MAE provides complementary information regarding average ordinal deviation.

one-unit change in the predictor. To account for variability across languages, we also use mixed-effects models where language is treated as a random effect. Intuitively, this allows us to separate general trends (e.g., dataset size effects) from language-specific variation, ensuring that observed effects are not driven by particular languages but reflect broader patterns. At each step, we examine the increase in explained variance (R^2) to determine how much additional information each block of variables contributes. This incremental approach allows us to assess whether performance differences are primarily driven by architecture, pretraining exposure, supervision volume, or broader linguistic structure.

3.2. Representational Framework

Given our goal of testing whether proficiency can be recovered as a coherent and ordered geometric factor, we move beyond predictive accuracy and examine the internal structure of embedding space. In this sense, a purely unsupervised analysis would identify directions of maximal variance, which in multilingual embeddings are typically driven by language identity, topic distribution, or corpus effects. Proficiency, however, is not expected to dominate global variance. Therefore, the absence of visible clustering in an unsupervised projection would not imply absence of structured proficiency encoding. For this reason, we adopt a supervised probing strategy. Rather than asking which directions explain the most variance, we directly test whether a simple linear direction can be learned that aligns embeddings with CEFR levels. If proficiency is encoded in representation space, it should be recoverable through such a projection even if it does not correspond to a principal component.

For each LLM variant, we train a Ridge regression model to predict numeric CEFR levels from standardized document embeddings.[5] The learned regression coefficients define a direction in embedding space that best aligns with proficiency. Each document is then projected onto this direction, yielding a scalar value that represents its position along the candidate proficiency axis. To ensure generalizability, the direction is learned using cross-validation. A linear probe is intentionally chosen as a minimal-capacity test (Tenney et al., 2019; Hewitt and Manning, 2019). If CEFR corresponds to structured information encoded in embeddings, it should be recoverable without requiring complex nonlinear transformations.

If CEFR is encoded as an ordered geometric factor, projected values should increase with proficiency level. We quantify this alignment using Spearman rank correlation between projection values and CEFR labels.[6] Spearman's ρ is appropriate given the ordinal nature of CEFR. A high correlation indicates a consistent ordering of levels along the recovered axis.

Alignment alone does not guarantee meaningful geometric separation. A projection may correlate strongly with CEFR while still exhibiting substantial overlap between adjacent levels. We therefore examine dispersion patterns along the recovered axis. For each CEFR level, we compute: (1) the mean projection value (centroid), and (2) the variance of projection values within that level. We then quantify CEFR-level separation using the Fisher ratio, defined as the average squared distance between level centroids divided by the average within-level variance. This ratio evaluates whether between-level differences dominate internal variability. A high Fisher value indicates that CEFR levels form compact and well-separated bands along the axis, rather than diffuse, overlapping distributions.

To further evaluate ordinal geometry, we compute pairwise distances between level centroids and examine their relationship by measuring how far apart two CEFR levels are in ordinal terms. If proficiency is encoded as a continuous gradient, centroid distance should increase as the ordinal difference between levels increases.

To contextualize these findings, we also examine unsupervised principal component projections derived solely from embedding variance, without using CEFR labels or the learned proficiency axis. If proficiency were a dominant global variance component, CEFR levels would align along the leading principal components. By comparing unsupervised structure with supervised projections, we determine whether proficiency is a principal variance driver or a recoverable latent dimension embedded within higher-dimensional structure.

3.3. Corpus

We conduct our experiments on UniversalCEFR (Imperial et al., 2025), a large-scale multilingual benchmark for CEFR-based proficiency modeling. This dataset is composed of CEFR-annotated texts spanning 13 languages and covering the full proficiency spectrum from A1 to C2. Table 1 summarizes the number of documents per language in our experimental setup.

UniversalCEFR integrates both learner-produced texts and pedagogically curated reference materi-

[5] We use ridge regression to estimate a single linear projection from embedding space to ordinal CEFR levels. Unlike multi-class classification models, ridge regression yields a continuous direction that is interpretable as a proficiency axis. L2 regularization further ensures stability given the high dimensionality and collinearity of LLM embeddings.

[6] To assess robustness, we compute 95% confidence intervals using bootstrap resampling over documents.

Language	# Docs	CEFR Levels
Arabic (ar)	2,160	A1–C2
Czech (cz)	441	A1–C1
Welsh (cy)	1,372	A1–A2
German (de)	1,542	A1–C2
English (en)	15,513	A1–C2
Spanish (es)	31,355	A1–C2
French (fr)	2,013	A1–C2
Hindi (hi)	1,491	A1–C1
Italian (it)	813	A1–B2
Dutch (nl)	3,596	A1–C2
Portuguese (pt)	1,423	A1–C2
Russian (ru)	1,758	A1–C2

Table 1: Number of CEFR-annotated documents per language in our experimental setup.

als under a unified CEFR labeling scheme. The languages represented belong to diverse typological families, including Germanic, Romance, Slavic, Indo-Aryan, Celtic, and Semitic languages. This typological and task diversity provides a heterogeneous yet controlled environment for multilingual proficiency analysis.

UniversalCEFR is particularly suitable for our study because its large scale and full CEFR coverage enable robust estimation of proficiency-related structure across all six levels. Its multilingual composition allows us to test whether proficiency constitutes a coherent representational dimension across typologically diverse languages. Furthermore, the inclusion of both learner and reference texts enables the analysis of production- and reception-oriented proficiency within a shared embedding space. Together, these properties make UniversalCEFR an appropriate benchmark for evaluating both predictive recoverability (RQ1) and geometric coherence (RQ2).

4. Results

4.1. Predictive Performance

Table 2 reports cross-validated performance for each model configuration across languages. Frozen embeddings from all EuroLLM variants support CEFR classification at levels exceeding the prompt-based approach reported in Imperial et al. (2025).[7]

Performance differences between Enriched and Instruct-Enriched are small and inconsistent across metrics. Similarly, the difference between 1.7B and 9B models is modest. These results indicate that CEFR-relevant signal is already present in pretrained embeddings and does not depend strongly on model variant.

Across classifiers, Logistic Regression and linear SVM perform comparably to the MLP. The limited improvement from the nonlinear MLP suggests that CEFR signal is largely linearly recoverable from embeddings.

The relationship between pretraining proportion and CEFR performance reveals no meaningful association. The Spearman correlation between pretraining proportion and F1 is negligible ($\rho = -0.012$). Although pretraining proportion appears statistically significant in intermediate regression models, its effect disappears once language family is included. Pretraining exposure alone, therefore, does not robustly predict CEFR performance. Figure 1 provides a visual illustration of this pattern.

In contrast, dataset size exhibits a consistent association with performance. Adding log-transformed dataset size increases explained variance from $R^2 = 0.017$ to $R^2 = 0.139$ ($\Delta R^2 = 0.122$). In OLS models, the coefficient for dataset size remains positive and statistically significant. In mixed-effects models controlling for language-level, the effect is no longer statistically significant.

Language family accounts for a substantial portion of performance variability. In OLS regression, introducing language family increases explained variance to $R^2 = 0.49$ (Adj. $R^2 = 0.455$), representing the largest improvement across model specifications. Romance and Celtic languages exhibit significantly higher performance relative to the reference category. To account for non-independence across languages, we further estimate a mixed-effects model with language as a random factor. In this specification, family-level patterns remain observable, although effect sizes are attenuated. This indicates that part of the variance attributed to language family reflects language-specific structure rather than purely genealogical grouping. Taken together, the regression results indicate that dataset size and language family explain substantially more variance than model configuration or pretraining proportion.[8]

4.2. Representational Structure

Given the capacity of simple machine learning models to predict the CEFR level using embeddings from a frozen model, we now examine whether CEFR is encoded as a coherent latent direction in embedding space.

Table 3 shows that all three models yield strong ordinal alignment between projection values and CEFR levels, with cross-validated Spearman correlations above 0.89. The narrow bootstrap confidence intervals indicate that this alignment is sta-

[7] We consider the best F1 using prompt-based method reported on Imperial et al. (2025) for the EuroLLM9B.

[8] Performance distributions across languages and families are illustrated in Appendix Figure 7.

Lang	Acc	F1	QWK	MAE	σ_{F1}	Prompt F1
es	0.986	0.986	0.987	0.025	0.302	0.28
cy	0.959	0.959	0.914	0.041	0.035	0.26
it	0.843	0.627	0.740	0.157	0.034	0.42
pt	0.613	0.567	0.696	0.601	0.177	0.21
de	0.681	0.563	0.815	0.354	0.125	0.38
cs	0.733	0.535	0.762	0.270	0.018	0.33
ar	0.495	0.461	0.661	0.616	0.068	0.35
en	0.516	0.421	0.696	0.554	0.106	0.23
fr	0.438	0.402	0.643	0.711	0.044	0.28
nl	0.443	0.395	0.553	0.665	0.008	0.32
ru	0.425	0.395	0.742	0.760	0.037	0.21
hi	0.367	0.366	0.659	0.961	0.021	0.21

Table 2: Weighted performance by language. Metrics are averaged across all model configurations, with dataset size as the weight. σ_{F1} reports the standard deviation of F1 across configurations. Table 4 (Appendix) shows the results for all model configurations.

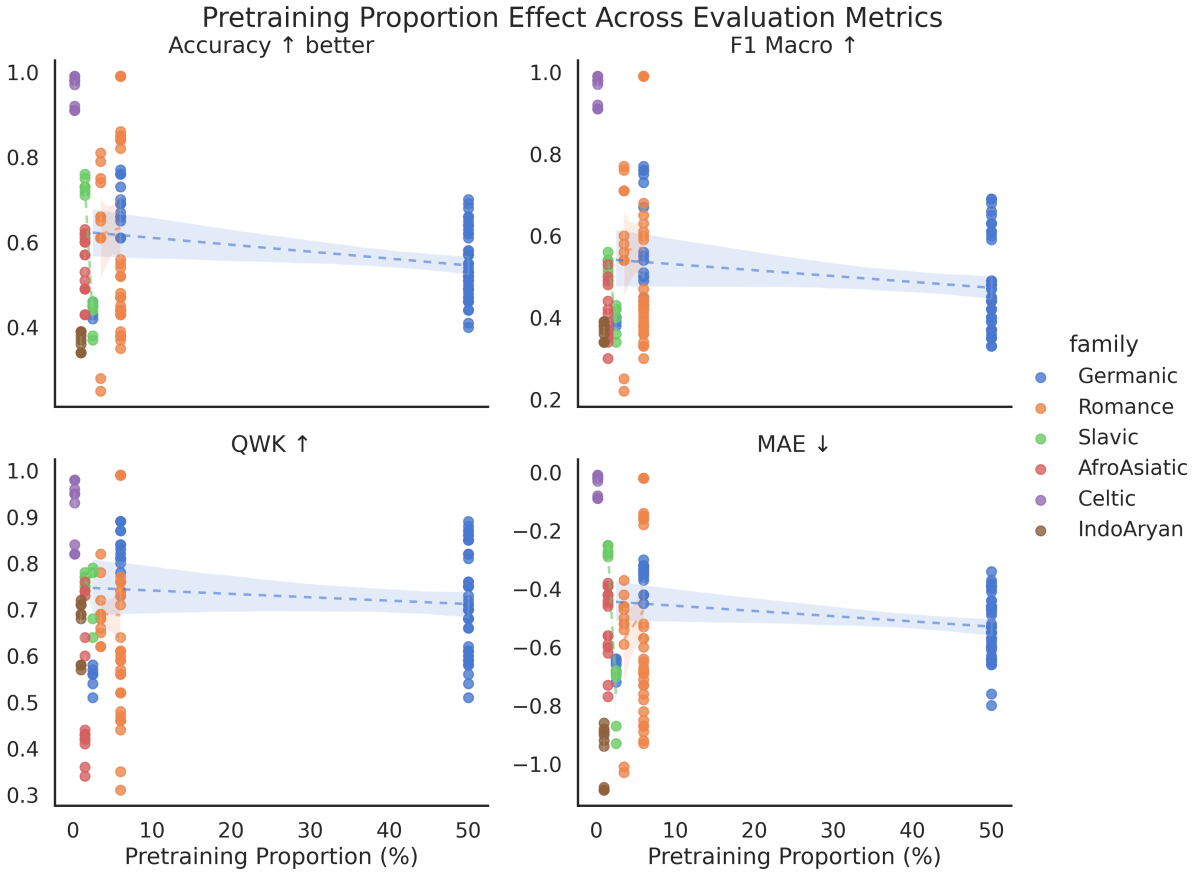

Figure 1: Relationship between pretraining proportion and CEFR classification performance across metrics.

ble and not driven by sampling variability. The 9B model achieves the highest correlation ($\rho_{CV} = 0.902$) and the largest Fisher separation ratio (12.30), suggesting slightly sharper geometric separation of proficiency bands. However, differences across model variants are modest. These results indicate that CEFR is not merely recoverable through classification (RQ1) but is encoded as a coherent geometric direction within embedding space. The high Fisher ratios further suggest structured banding rather than random overlap.

To visually ground these quantitative results, cross-validated projections onto the learned proficiency axis show monotonic increases in projection values from A1 to C2 across models. Although adjacent levels partially overlap, global ordinal struc-

Model	ρ_{CV}	95% CI	Fisher Ratio
1.7B	0.895	[0.893, 0.897]	11.24
1.7B-Instruct	0.896	[0.894, 0.898]	11.26
9B	0.902	[0.900, 0.904]	12.30

Table 3: Cross-validated ordinal alignment and geometric separation metrics.

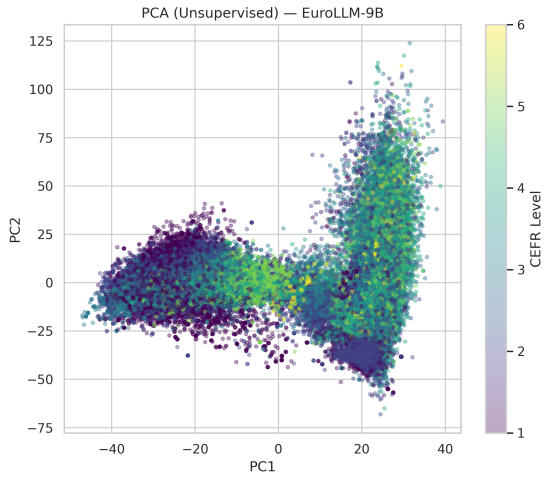

Figure 2: Unsupervised PCA projection (EuroLLM-9B example).

ture is preserved. Examining centroid geometry further reveals that between-level distances increase approximately in proportion to ordinal differences, supporting the interpretation of CEFR as a continuous representational gradient rather than a set of discrete clusters.[9]

Figure 2 shows that CEFR levels do not align with dominant variance directions. The absence of an ordered gradient indicates that proficiency is not a primary organizing factor in embedding space. In contrast, Figure 3 shows the same embeddings reorganized using the supervised proficiency direction. A clear horizontal gradient from A1 to C2 becomes visible, while orthogonal dispersion remains substantial. The contrast between Figures 2 and 3 indicates that proficiency does not dominate global variance but is recoverable as a task-relevant latent dimension.

Finally, Figure 3 shows that both task types align along the same horizontal proficiency axis. At the same time, in both space representations, two vertically separated modality clusters (i.e., learner vs. reference texts) emerge along the orthogonal dimension.[10]. This indicates that the recovered direction captures shared proficiency-related structure,

[9]Cross-validated axis projections and centroid distance diagnostics are provided in Appendix Figures 4 and 5.

[10]Learner and reference clusters are illustrated in Figure 6.

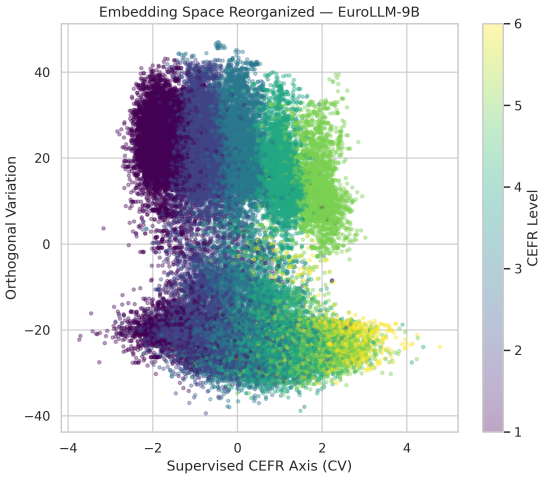

Figure 3: Embedding space reorganized using the supervised CEFR axis.

while modality-specific variation is expressed primarily in dimensions orthogonal to the proficiency axis.

5. Discussion

Our results show that frozen multilingual LLM embeddings consistently support CEFR classification across languages. This finding extends prior work in Automated Essay Scoring (AES) and readability assessment, where encoder-based models typically require task-specific fine-tuning or feature augmentation to achieve competitive performance (Taghipour and Ng, 2016; Alikaniotis et al., 2016).

We demonstrate that pretrained embeddings alone (without fine-tuning or linguistic features) already encode a signal for multilingual CEFR prediction. This challenges the prevailing assumption that assessment tasks necessarily require task-adapted architectures or sophisticated prompts. Instead, our findings suggest that CEFR-relevant linguistic structure emerges during large-scale pretraining.

Importantly, performance differences between Enriched and Instruct-Enriched variants are minimal. This indicates that CEFR signal does not primarily arise from instruction-following alignment, but is already present in base pretrained representations.

A central finding concerns the limited explanatory power of pretraining proportion. While generative LLM literature often emphasizes scaling and data exposure as primary drivers of performance (Brown et al., 2020), our regression analyses show that pretraining proportion does not robustly predict CEFR classification once language family is controlled.

In contrast, supervision volume (dataset size) and linguistic grouping explain substantially more variance. This aligns with multilingual NLP re-

search showing that cross-lingual transfer depends strongly on typological proximity and annotation density (Lauscher et al., 2020; Bjerva et al., 2019).

For CEFR assessment, this implies that increasing annotated training data may yield greater gains than modifying embedding architectures or increasing raw pretraining exposure.

The results provide evidence that CEFR is encoded as a recoverable dimension. High cross-validated ordinal correlations and substantial Fisher separation ratios indicate that proficiency is not merely classifiable.

Proficiency does not align with dominant principal components. This shows that CEFR is not a primary global variance factor (such as language or topic), but a latent dimension embedded within a higher-dimensional structure. This distinction reframes CEFR modeling: rather than asking whether models can predict proficiency, we show that proficiency corresponds to an internally organized representational gradient.

The approximately linear relationship between centroid distance and ordinal gap suggests that CEFR is encoded as a continuous proficiency gradient rather than as sharply separated categories. Adjacent levels overlap substantially, while global ordering remains stable. This geometric perspective aligns with theoretical interpretations of CEFR as a continuum of communicative competence rather than as strictly discrete bands.

The recovered proficiency axis is shared across learner (production) and reference (reception) texts. While learner texts exhibit greater orthogonal dispersion, particularly at lower levels, the horizontal ordering remains consistent.

These findings point to a recovered dimension that captures shared linguistic complexity rather than task-specific artefacts. Production variability manifests primarily in orthogonal directions, while proficiency alignment remains stable. These findings are consistent with an interpretation in terms of linguistic competence rather than purely corpus-specific conventions.

6. Conclusion

This study investigated whether multilingual LLM embeddings encode CEFR proficiency as recoverable and structured information. Investigating the extent to which frozen multilingual LLM embeddings support CEFR classification across languages and modeling configurations (*RQ1*), we showed that frozen EuroLLM embeddings support multilingual CEFR classification without task-specific fine-tuning, and that supervision volume and linguistic grouping explain substantially more variance than model configuration or pretraining proportion.

Assessing whether CEFR proficiency corresponds to a recoverable latent geometric axis in embedding space that generalizes across languages and task modalities (*RQ2*), we demonstrated that CEFR corresponds to a coherent linear direction in embedding space: proficiency is not merely predictable, but geometrically organized as a continuous gradient across languages and task modalities.

These findings shift the perspective from purely predictive evaluation toward representational validity in proficiency assessment. Rather than asking only whether LLMs can classify CEFR levels, we show that they internally encode structured information aligned with standardized proficiency scales. Future work should examine how this geometric structure transfers across languages, interacts with fine-tuning, and relates to interpretable linguistic features.

Limitations

First, embeddings are evaluated in a frozen setting. Fine-tuning may alter the relative impact of pretraining exposure and architectural differences.

Second, language family is used as a coarse proxy for linguistic structure. It does not capture detailed typological variables such as morphological complexity or syntactic configuration.

Third, pretraining proportion is treated as a scalar measure and does not account for domain similarity or data quality.

Fourth, CEFR labels themselves may reflect corpus-specific annotation practices, which can influence performance independently of linguistic competence.

Finally, the recovered proficiency axis may capture linguistic properties that are strongly correlated with CEFR rather than proficiency as an abstract construct. Features such as lexical sophistication, syntactic depth, and discourse complexity are associated with CEFR levels and may serve as proxy variables. While this does not undermine the finding that CEFR is recoverable as a structured representational dimension, it implies that the learned direction may partially reflect correlated linguistic properties associated with CEFR.

Plain Summary

This paper looks at whether large language models (LLMs) understand language proficiency in a meaningful way, rather than just guessing it from surface patterns. Language proficiency is often described using levels like A1 (beginner) to C2 (advanced). In this paper, we ask: do these levels actually exist inside the model's internal representations? In other words, is language proficiency something the model "organises" naturally, or is it just something

we can predict with a classifier? To investigate this, the study uses data covering many languages and proficiency levels. It analyses the internal representations (embeddings) of a multilingual model without fine-tuning it. The findings show two main things:

- The model's representations already contain enough information to predict proficiency levels quite well.
- Proficiency appears to form a continuous scale inside the model: texts from beginner to advanced levels are arranged along a consistent direction, rather than forming unrelated groups.

This suggests that the model does not just memorise patterns, but organises language proficiency in a structured way. The paper also finds that performance differences are mainly explained by factors like dataset size and language differences, rather than by the model itself. Overall, this work shows that language proficiency is not only predictable from LLMs, but is also reflected as an underlying structure in their representations. This could be useful for applications such as adapting text to different learners or generating feedback based on proficiency levels.

Acknowledgements

This study was supported by Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Brazil.

Bibliographical References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 715–725.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring–an empirical study. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1072–1077.

Muhammad Faseeh, Abdul Jaleel, Naeem Iqbal, Anwar Ghani, Akmalbek Abdusalomov, Asif Mehmood, and Young-Im Cho. 2024. Hybrid approach to automated essay scoring: Integrating deep learning embeddings with handcrafted linguistic features for improved accuracy. *Mathematics*, 12(21):3416.

Thomas François and Cédrick Fairon. 2012. An "ai readability" formula for french as a foreign language. In *Proceedings of the 2012 joint conference on empirical methods in Natural Language Processing and computational natural language learning*, pages 466–477.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4129–4138.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *IJCAI*, volume 19, pages 6300–6308.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

Xia Li, Huali Yang, Shengze Hu, Jing Geng, Keke Lin, and Yuhai Li. 2022. Enhanced hybrid neural network for automated essay scoring. *Expert Systems*, 39(10):e13068.

Pei Yee Liew and Ian KT Tan. 2024. On automated essay grading using large language models. In *Proceedings of the 2024 8th international conference on computer science and artificial intelligence*, pages 204–211.

Fengkai Liu and John SY Lee. 2023. Hybrid models for sentence readability assessment. In *Proceedings of the 18th workshop on innovative use of nlp for building educational applications (bea 2023)*, pages 448–454.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *arXiv preprint arXiv:1904.09482*.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M Alves, José Pombal, Nicolas Boizard, et al. 2025a. Eurollm-9b: Technical report. *arXiv preprint arXiv:2506.04079*.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2025b. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune bert for automated essay scoring? In *Proceedings of the fifteenth workshop on innovative use of NLP for building educational applications*, pages 151–162.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Jesse Mu and Jacob Andreas. 2020. Compositional explanations of neurons. *Advances in Neural Information Processing Systems*, 33:17153–17163.

Council of Europe. 2020. *Common European framework of reference for languages: Companion volume*. Council of Europe.

Council of Europe. Council for Cultural Cooperation. Education Committee. Modern Languages Division. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Ellis B Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 229–239.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622.

Pedro Uria Rodriguez, Amir Jafari, and Christopher M Ormerod. 2019. Language models and automated essay scoring. *arXiv preprint arXiv:1909.09482*.

Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. 2025. Can ai grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th international learning analytics and knowledge conference*, pages 462–472.

Mark D Shermis and Jill C Burstein. 2003. *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891.

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.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 5366–5377.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal cefr classification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 147–153.

Rodrigo Wilkens, Patrick Watrin, Rémi Cardon, Alice Pintard, Isabelle Gribomont, and Thomas François. 2024. Exploring hybrid approaches to readability: experiments on the complementarity between linguistic features and transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2316–2331.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, and A. M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short l2 essays on the cefr scale with gpt-4. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 576–584.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 180–189.

Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. 2015. Task-independent features for automated essay grading. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications*, pages 224–232.

7. Language Resource References

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R Jablonkai, et al. 2025. Universalcefr: Enabling open multilingual research on language proficiency assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9714–9766.

A. Additional Geometric Diagnostics

A.1. Cross-Validated Axis Projection

Figure 4 displays cross-validated projections of embeddings onto the learned proficiency axis for all model variants. Each panel shows the distribution of projection values per CEFR level.

This visualization provides a direct view of ordinal alignment in one dimension. Median projection values increase monotonically from A1 to C2, indicating that higher proficiency levels are systematically shifted along the recovered axis. Although adjacent levels partially overlap, the global ordering remains stable across folds and across model sizes.

Importantly, this figure reflects out-of-fold projections and therefore illustrates generalizable structure rather than in-sample fitting.

A.2. Centroid Distance as a Function of Ordinal Gap

Figure 5 examines geometric separation from a complementary perspective. Instead of showing full distributions, it summarizes each CEFR level by its centroid along the supervised axis and measures the distance between level centroids as a function of absolute ordinal difference.

Distances increase approximately linearly with CEFR gap for all models. This indicates that geometric spacing between levels scales proportionally with ordinal distance, reinforcing the interpretation of proficiency as a continuous gradient rather than a collection of arbitrarily separated clusters.

While Figure 4 emphasizes within-level dispersion and overlap, Figure 5 focuses on between-level scaling. Together, they provide complementary evidence of structured ordinal geometry.

B. Learner and reference dispersion

Figure 6 presents a two-dimensional PCA projection of EuroLLM-9B embeddings without using CEFR labels or task supervision. Points are colored according to task category (learner vs. reference).

The first principal component (PC1) reveals a strong separation between learner and reference texts. Reference texts cluster predominantly on the right-hand side of the projection, while learner texts occupy the left region, with limited overlap in the central area. This indicates that task category accounts for a substantial portion of the dominant variance in embedding space.

In contrast, the second principal component (PC2) primarily captures dispersion within each category rather than cross-category separation. The

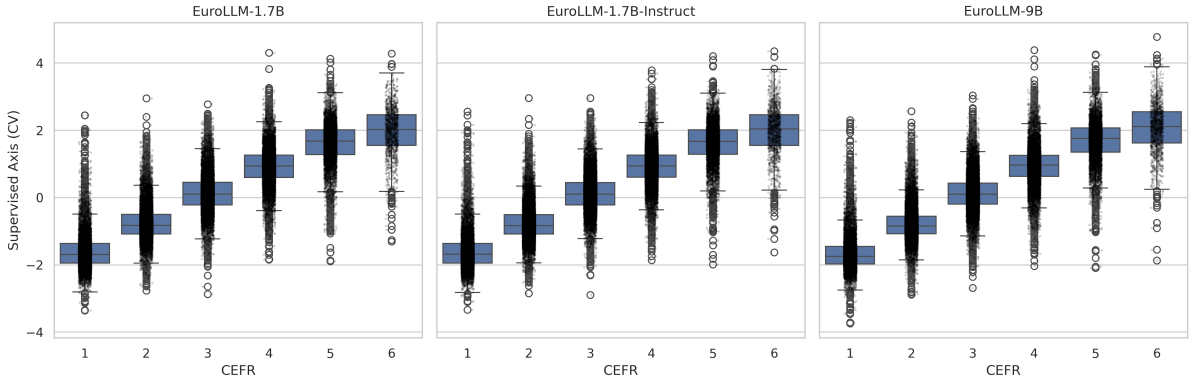

Figure 4: Cross-validated supervised projection onto the CEFR proficiency axis.

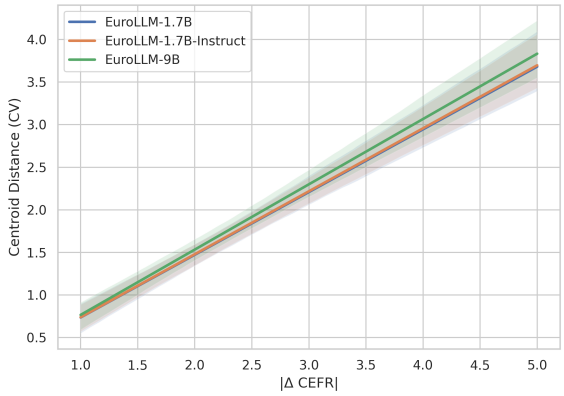

Figure 5: Centroid distance as a function of absolute CEFR gap under cross-validation.

vertical spread suggests substantial internal heterogeneity within both learner and reference groups.

B.1. Performance by Language and Family

Figure 7 shows performance distributions across languages, colored by language family. Romance and Celtic languages tend to exhibit higher average performance, whereas Indo-Aryan languages show comparatively lower scores. However, dispersion within families indicates that language-specific properties also contribute to performance variability, consistent with the mixed-effects analysis.

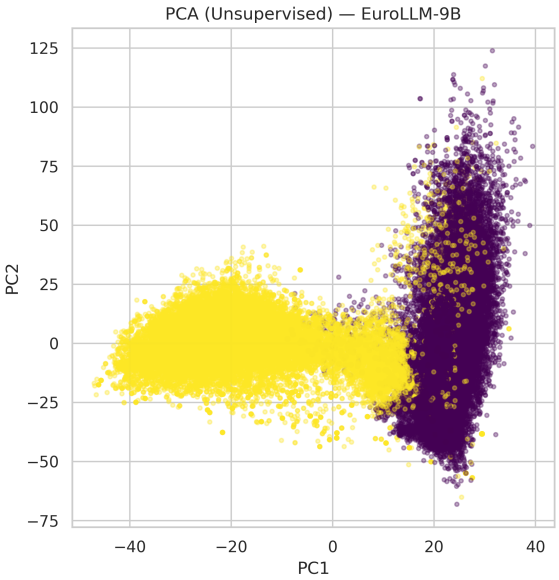

Figure 6: PCA projecting with colors indicating learner and references corpus modalities.

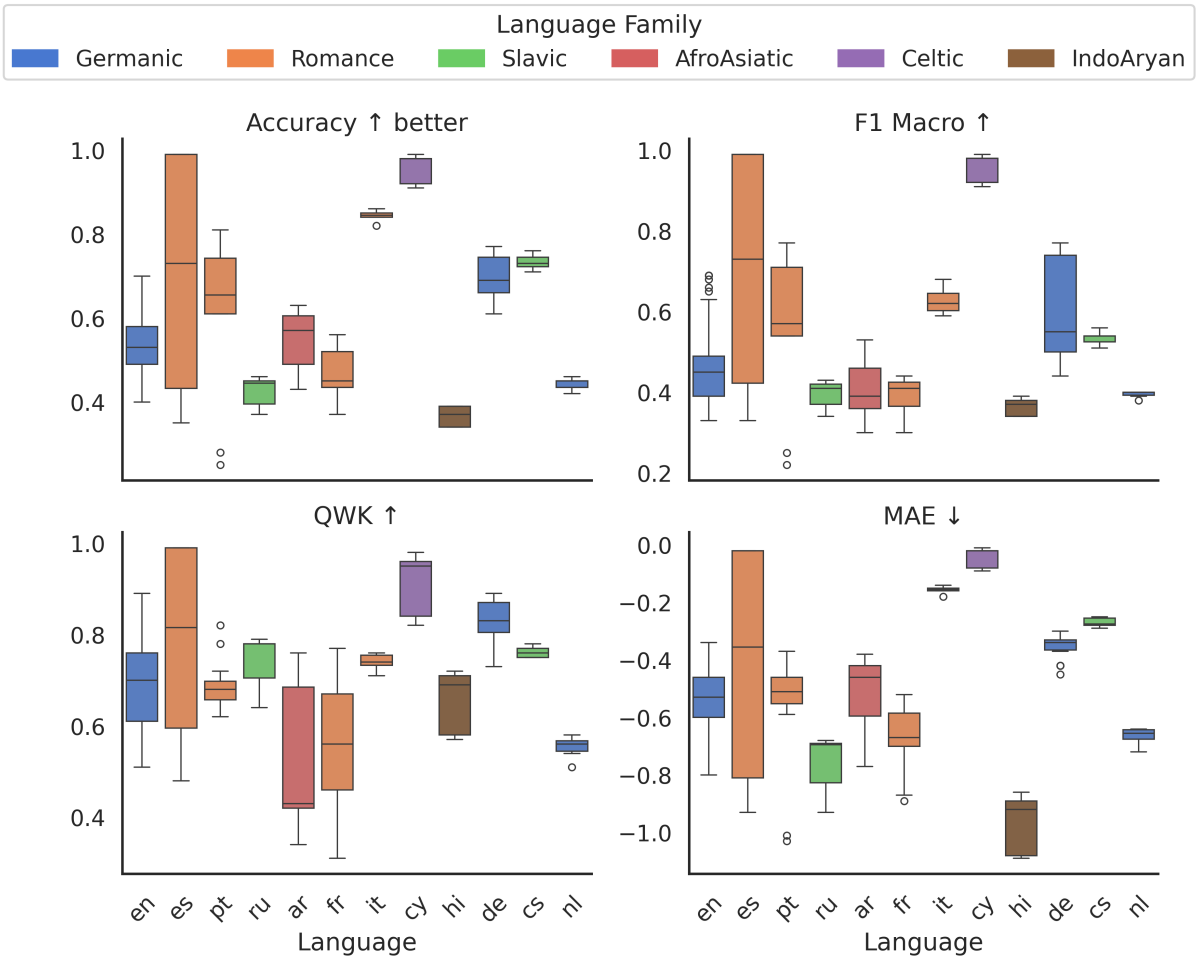

Figure 7: Performance distribution by language (colored by language family) across evaluation metrics.

C. All ML Results

Table 4 presents the complete set of configuration-level results underlying the RQ1 analyses. Each entry corresponds to a unique combination of LLM variant, classifier, language, and corpus. Metrics are reported as mean ± standard deviation across evaluation folds. This detailed table is provided to ensure transparency and reproducibility of all statistical analyses discussed in the main text.

LLMtype	LLMsize	MLmodel	lang	corpus	Acc	F1	QWK	MAE
e	1.700000	LinearSVM	ar	readme	0.43 ± 0.03	0.42 ± 0.03	0.6 ± 0.04	0.77 ± 0.05
e	1.700000	LogReg	ar	readme	0.49 ± 0.02	0.48 ± 0.03	0.73 ± 0.02	0.62 ± 0.02
e	1.700000	MLP	ar	readme	0.51 ± 0.04	0.5 ± 0.04	0.74 ± 0.03	0.59 ± 0.06
e	1.700000	LinearSVM	ar	zaebuc	0.57 ± 0.09	0.41 ± 0.16	0.41 ± 0.16	0.45 ± 0.1
e	1.700000	LogReg	ar	zaebuc	0.62 ± 0.07	0.37 ± 0.09	0.43 ± 0.07	0.39 ± 0.06
e	1.700000	MLP	ar	zaebuc	0.61 ± 0.06	0.35 ± 0.08	0.42 ± 0.11	0.42 ± 0.08
e	1.700000	LinearSVM	cy	learn	0.99 ± 0.01	0.99 ± 0.01	0.98 ± 0.01	0.01 ± 0.01
e	1.700000	LogReg	cy	learn	0.98 ± 0.01	0.98 ± 0.01	0.95 ± 0.02	0.02 ± 0.01
e	1.700000	MLP	cy	learn	0.91 ± 0.02	0.91 ± 0.02	0.82 ± 0.05	0.09 ± 0.02
e	1.700000	LinearSVM	de	elg	0.77 ± 0.06	0.77 ± 0.08	0.89 ± 0.03	0.33 ± 0.08
e	1.700000	LogReg	de	elg	0.76 ± 0.04	0.76 ± 0.05	0.89 ± 0.02	0.33 ± 0.06
e	1.700000	MLP	de	elg	0.73 ± 0.05	0.73 ± 0.05	0.87 ± 0.04	0.37 ± 0.08
e	1.700000	LinearSVM	de	merlin	0.61 ± 0.05	0.45 ± 0.07	0.73 ± 0.05	0.42 ± 0.06
e	1.700000	LogReg	de	merlin	0.67 ± 0.03	0.5 ± 0.06	0.81 ± 0.03	0.34 ± 0.03
e	1.700000	MLP	de	merlin	0.7 ± 0.03	0.56 ± 0.08	0.84 ± 0.02	0.3 ± 0.03
e	1.700000	LinearSVM	en	cambridge	0.62 ± 0.04	0.6 ± 0.04	0.82 ± 0.04	0.49 ± 0.05
e	1.700000	LogReg	en	cambridge	0.68 ± 0.05	0.68 ± 0.05	0.87 ± 0.03	0.39 ± 0.07
e	1.700000	MLP	en	cambridge	0.64 ± 0.05	0.63 ± 0.05	0.89 ± 0.03	0.4 ± 0.07
e	1.700000	LinearSVM	en	cefr	0.44 ± 0.05	0.33 ± 0.06	0.61 ± 0.08	0.66 ± 0.07
e	1.700000	LogReg	en	cefr	0.51 ± 0.04	0.37 ± 0.05	0.7 ± 0.04	0.55 ± 0.05
e	1.700000	MLP	en	cefr	0.47 ± 0.04	0.35 ± 0.06	0.68 ± 0.05	0.6 ± 0.05
e	1.700000	LinearSVM	en	icle500	0.51 ± 0.03	0.44 ± 0.05	0.54 ± 0.08	0.6 ± 0.05
e	1.700000	LogReg	en	icle500	0.55 ± 0.02	0.47 ± 0.06	0.59 ± 0.05	0.53 ± 0.03
e	1.700000	MLP	en	icle500	0.55 ± 0.06	0.44 ± 0.09	0.62 ± 0.07	0.53 ± 0.09
e	1.700000	LinearSVM	en	readme	0.4 ± 0.03	0.35 ± 0.04	0.56 ± 0.05	0.8 ± 0.06
e	1.700000	LogReg	en	readme	0.46 ± 0.03	0.4 ± 0.04	0.7 ± 0.03	0.64 ± 0.04
e	1.700000	MLP	en	readme	0.49 ± 0.02	0.42 ± 0.03	0.73 ± 0.02	0.59 ± 0.03
e	1.700000	LinearSVM	es	caes	0.99 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.02 ± 0.0
e	1.700000	LogReg	es	caes	0.99 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.02 ± 0.0
e	1.700000	MLP	es	caes	0.99 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.02 ± 0.0
e	1.700000	LinearSVM	es	kwiqiz	0.39 ± 0.04	0.36 ± 0.04	0.57 ± 0.07	0.85 ± 0.07
e	1.700000	LogReg	es	kwiqiz	0.44 ± 0.05	0.45 ± 0.05	0.61 ± 0.05	0.73 ± 0.07
e	1.700000	MLP	es	kwiqiz	0.38 ± 0.08	0.38 ± 0.08	0.48 ± 0.08	0.93 ± 0.11
e	1.700000	LinearSVM	fr	kwiqiz	0.55 ± 0.05	0.42 ± 0.05	0.47 ± 0.1	0.57 ± 0.05
e	1.700000	LogReg	fr	kwiqiz	0.56 ± 0.04	0.44 ± 0.06	0.56 ± 0.11	0.52 ± 0.06
e	1.700000	MLP	fr	kwiqiz	0.44 ± 0.03	0.34 ± 0.08	0.31 ± 0.05	0.71 ± 0.02
e	1.700000	LinearSVM	fr	readme	0.37 ± 0.04	0.36 ± 0.04	0.59 ± 0.05	0.89 ± 0.07
e	1.700000	LogReg	fr	readme	0.44 ± 0.04	0.43 ± 0.04	0.74 ± 0.04	0.68 ± 0.05
e	1.700000	MLP	fr	readme	0.43 ± 0.04	0.42 ± 0.05	0.76 ± 0.02	0.67 ± 0.04
e	1.700000	LinearSVM	hi	readme	0.34 ± 0.04	0.34 ± 0.04	0.58 ± 0.03	1.08 ± 0.07
e	1.700000	LogReg	hi	readme	0.36 ± 0.04	0.36 ± 0.04	0.68 ± 0.03	0.94 ± 0.07
e	1.700000	MLP	hi	readme	0.38 ± 0.03	0.38 ± 0.04	0.71 ± 0.03	0.89 ± 0.06
e	1.700000	LinearSVM	it	merlin	0.82 ± 0.04	0.59 ± 0.09	0.71 ± 0.07	0.18 ± 0.04

Continued on next page

LLMtype	LLMsize	MLmodel	lang	corpus	Acc	F1	QWK	MAE
e	1.700000	LogReg	it	merlin	0.85 ± 0.04	0.65 ± 0.1	0.76 ± 0.06	0.15 ± 0.04
e	1.700000	MLP	it	merlin	0.85 ± 0.03	0.61 ± 0.11	0.74 ± 0.06	0.15 ± 0.03
e	1.700000	LinearSVM	nl	elg	0.45 ± 0.02	0.4 ± 0.03	0.56 ± 0.02	0.66 ± 0.02
e	1.700000	LogReg	nl	elg	0.46 ± 0.02	0.4 ± 0.04	0.58 ± 0.03	0.64 ± 0.03
e	1.700000	MLP	nl	elg	0.45 ± 0.03	0.39 ± 0.03	0.56 ± 0.04	0.65 ± 0.04
e	1.700000	LinearSVM	ru	readme	0.37 ± 0.03	0.34 ± 0.03	0.64 ± 0.04	0.93 ± 0.07
e	1.700000	LogReg	ru	readme	0.45 ± 0.04	0.42 ± 0.04	0.78 ± 0.03	0.69 ± 0.06
e	1.700000	MLP	ru	readme	0.45 ± 0.03	0.42 ± 0.04	0.78 ± 0.02	0.69 ± 0.04
e	9.000000	LinearSVM	ar	readme	0.43 ± 0.02	0.44 ± 0.01	0.64 ± 0.03	0.73 ± 0.03
e	9.000000	LogReg	ar	readme	0.49 ± 0.02	0.49 ± 0.04	0.74 ± 0.02	0.6 ± 0.03
e	9.000000	MLP	ar	readme	0.53 ± 0.04	0.53 ± 0.05	0.76 ± 0.03	0.56 ± 0.06
e	9.000000	LinearSVM	ar	zaebuc	0.49 ± 0.08	0.36 ± 0.09	0.42 ± 0.12	0.56 ± 0.1
e	9.000000	LogReg	ar	zaebuc	0.63 ± 0.05	0.34 ± 0.07	0.44 ± 0.1	0.38 ± 0.05
e	9.000000	MLP	ar	zaebuc	0.6 ± 0.06	0.36 ± 0.07	0.36 ± 0.1	0.44 ± 0.07
e	9.000000	LinearSVM	cs	merlin	0.71 ± 0.05	0.52 ± 0.04	0.75 ± 0.04	0.29 ± 0.05
e	9.000000	LogReg	cs	merlin	0.73 ± 0.04	0.54 ± 0.04	0.75 ± 0.06	0.27 ± 0.05
e	9.000000	MLP	cs	merlin	0.75 ± 0.02	0.51 ± 0.07	0.77 ± 0.04	0.25 ± 0.03
e	9.000000	LinearSVM	cy	learn	0.98 ± 0.01	0.98 ± 0.01	0.96 ± 0.02	0.02 ± 0.01
e	9.000000	LogReg	cy	learn	0.97 ± 0.01	0.97 ± 0.01	0.93 ± 0.02	0.03 ± 0.01
e	9.000000	MLP	cy	learn	0.92 ± 0.01	0.92 ± 0.01	0.84 ± 0.03	0.08 ± 0.01
e	9.000000	LinearSVM	de	merlin	0.65 ± 0.04	0.51 ± 0.06	0.78 ± 0.03	0.36 ± 0.04
e	9.000000	LogReg	de	merlin	0.69 ± 0.03	0.54 ± 0.07	0.82 ± 0.02	0.32 ± 0.04
e	9.000000	MLP	de	merlin	0.69 ± 0.01	0.55 ± 0.06	0.83 ± 0.02	0.32 ± 0.02
e	9.000000	LinearSVM	en	cefr	0.49 ± 0.01	0.39 ± 0.01	0.59 ± 0.01	0.64 ± 0.01
e	9.000000	LinearSVM	en	cefr	0.46 ± 0.03	0.37 ± 0.05	0.66 ± 0.05	0.65 ± 0.03
e	9.000000	LogReg	en	cefr	0.57 ± 0.01	0.48 ± 0.03	0.75 ± 0.01	0.46 ± 0.01
e	9.000000	LogReg	en	cefr	0.52 ± 0.04	0.4 ± 0.04	0.71 ± 0.03	0.53 ± 0.03
e	9.000000	MLP	en	cefr	0.57 ± 0.02	0.49 ± 0.03	0.75 ± 0.02	0.46 ± 0.03
e	9.000000	MLP	en	cefr	0.49 ± 0.06	0.35 ± 0.03	0.7 ± 0.04	0.55 ± 0.06
e	9.000000	LinearSVM	en	elg	0.66 ± 0.06	0.66 ± 0.1	0.85 ± 0.04	0.39 ± 0.08
e	9.000000	LogReg	en	elg	0.7 ± 0.05	0.69 ± 0.08	0.88 ± 0.03	0.34 ± 0.06
e	9.000000	MLP	en	elg	0.65 ± 0.05	0.63 ± 0.08	0.86 ± 0.04	0.39 ± 0.07
e	9.000000	LinearSVM	en	icle500	0.53 ± 0.06	0.48 ± 0.08	0.6 ± 0.07	0.55 ± 0.08
e	9.000000	LogReg	en	icle500	0.54 ± 0.06	0.48 ± 0.08	0.61 ± 0.05	0.53 ± 0.06
e	9.000000	MLP	en	icle500	0.57 ± 0.05	0.47 ± 0.08	0.61 ± 0.03	0.52 ± 0.05
e	9.000000	LinearSVM	en	readme	0.41 ± 0.04	0.36 ± 0.05	0.6 ± 0.03	0.76 ± 0.05
e	9.000000	LogReg	en	readme	0.48 ± 0.02	0.42 ± 0.04	0.72 ± 0.03	0.61 ± 0.04
e	9.000000	MLP	en	readme	0.5 ± 0.02	0.44 ± 0.03	0.75 ± 0.02	0.57 ± 0.03
e	9.000000	LinearSVM	es	caes	0.99 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.02 ± 0.0
e	9.000000	LogReg	es	caes	0.99 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.02 ± 0.0
e	9.000000	MLP	es	caes	0.99 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.02 ± 0.0
e	9.000000	LinearSVM	es	kwiqiz	0.43 ± 0.07	0.42 ± 0.07	0.59 ± 0.08	0.82 ± 0.1
e	9.000000	LogReg	es	kwiqiz	0.44 ± 0.05	0.44 ± 0.06	0.61 ± 0.07	0.76 ± 0.1
e	9.000000	MLP	es	kwiqiz	0.44 ± 0.05	0.43 ± 0.05	0.61 ± 0.04	0.78 ± 0.07
e	9.000000	LinearSVM	fr	kwiqiz	0.48 ± 0.03	0.39 ± 0.06	0.44 ± 0.11	0.69 ± 0.07
e	9.000000	LogReg	fr	kwiqiz	0.52 ± 0.04	0.41 ± 0.06	0.52 ± 0.12	0.57 ± 0.08
e	9.000000	MLP	fr	kwiqiz	0.47 ± 0.05	0.33 ± 0.06	0.46 ± 0.09	0.64 ± 0.08
e	9.000000	LinearSVM	hi	readme	0.34 ± 0.03	0.34 ± 0.03	0.57 ± 0.06	1.09 ± 0.08
e	9.000000	LogReg	hi	readme	0.39 ± 0.03	0.39 ± 0.03	0.71 ± 0.04	0.88 ± 0.06
e	9.000000	MLP	hi	readme	0.39 ± 0.05	0.39 ± 0.05	0.72 ± 0.03	0.86 ± 0.06
e	9.000000	LinearSVM	it	merlin	0.84 ± 0.02	0.6 ± 0.11	0.74 ± 0.04	0.16 ± 0.02
e	9.000000	LogReg	it	merlin	0.86 ± 0.03	0.68 ± 0.14	0.76 ± 0.05	0.14 ± 0.03
e	9.000000	MLP	it	merlin	0.84 ± 0.03	0.63 ± 0.12	0.73 ± 0.04	0.16 ± 0.03
e	9.000000	LinearSVM	nl	elg	0.42 ± 0.02	0.38 ± 0.03	0.51 ± 0.02	0.72 ± 0.02
e	9.000000	LogReg	nl	elg	0.45 ± 0.02	0.4 ± 0.03	0.57 ± 0.03	0.64 ± 0.03
e	9.000000	MLP	nl	elg	0.43 ± 0.03	0.4 ± 0.04	0.54 ± 0.03	0.68 ± 0.04
e	9.000000	LinearSVM	pt	cople2	0.75 ± 0.05	0.71 ± 0.06	0.72 ± 0.06	0.52 ± 0.09
e	9.000000	LogReg	pt	cople2	0.81 ± 0.05	0.77 ± 0.08	0.82 ± 0.04	0.37 ± 0.08
e	9.000000	MLP	pt	cople2	0.25 ± 0.05	0.22 ± 0.05	0.62 ± 0.03	1.03 ± 0.09
e	9.000000	LinearSVM	pt	peapl2	0.66 ± 0.06	0.6 ± 0.07	0.69 ± 0.07	0.46 ± 0.09
e	9.000000	LogReg	pt	peapl2	0.66 ± 0.04	0.58 ± 0.06	0.69 ± 0.09	0.46 ± 0.07
e	9.000000	MLP	pt	peapl2	0.65 ± 0.05	0.56 ± 0.07	0.68 ± 0.09	0.47 ± 0.07
e	9.000000	LinearSVM	ru	readme	0.38 ± 0.04	0.36 ± 0.04	0.68 ± 0.03	0.87 ± 0.06
e	9.000000	LogReg	ru	readme	0.46 ± 0.04	0.43 ± 0.05	0.79 ± 0.02	0.68 ± 0.05
e	9.000000	MLP	ru	readme	0.44 ± 0.03	0.4 ± 0.03	0.78 ± 0.02	0.7 ± 0.05
ei	1.700000	LinearSVM	ar	zaebuc	0.6 ± 0.09	0.39 ± 0.1	0.43 ± 0.16	0.42 ± 0.11
ei	1.700000	LogReg	ar	zaebuc	0.62 ± 0.05	0.37 ± 0.08	0.42 ± 0.08	0.39 ± 0.05
ei	1.700000	MLP	ar	zaebuc	0.57 ± 0.05	0.3 ± 0.02	0.34 ± 0.12	0.46 ± 0.09
ei	1.700000	LinearSVM	cs	merlin	0.73 ± 0.03	0.54 ± 0.03	0.75 ± 0.03	0.28 ± 0.03
ei	1.700000	LogReg	cs	merlin	0.76 ± 0.02	0.56 ± 0.02	0.78 ± 0.02	0.25 ± 0.03
ei	1.700000	MLP	cs	merlin	0.72 ± 0.03	0.54 ± 0.08	0.77 ± 0.03	0.28 ± 0.03
ei	1.700000	LinearSVM	cy	learn	0.99 ± 0.01	0.99 ± 0.01	0.98 ± 0.01	0.01 ± 0.01
ei	1.700000	LogReg	cy	learn	0.98 ± 0.01	0.98 ± 0.01	0.95 ± 0.03	0.02 ± 0.01
ei	1.700000	MLP	cy	learn	0.91 ± 0.02	0.91 ± 0.02	0.82 ± 0.03	0.09 ± 0.02
ei	1.700000	LinearSVM	de	elg	0.76 ± 0.07	0.76 ± 0.08	0.87 ± 0.04	0.35 ± 0.09
ei	1.700000	LogReg	de	elg	0.76 ± 0.04	0.75 ± 0.05	0.89 ± 0.03	0.33 ± 0.06
ei	1.700000	MLP	de	elg	0.69 ± 0.05	0.67 ± 0.07	0.84 ± 0.04	0.45 ± 0.09
ei	1.700000	LinearSVM	de	merlin	0.61 ± 0.03	0.44 ± 0.06	0.74 ± 0.05	0.42 ± 0.05
ei	1.700000	LogReg	de	merlin	0.66 ± 0.04	0.49 ± 0.06	0.8 ± 0.03	0.35 ± 0.04
ei	1.700000	MLP	de	merlin	0.66 ± 0.04	0.5 ± 0.04	0.81 ± 0.03	0.34 ± 0.05
ei	1.700000	LinearSVM	en	cambridge	0.63 ± 0.02	0.61 ± 0.02	0.82 ± 0.04	0.48 ± 0.04
ei	1.700000	LogReg	en	cambridge	0.69 ± 0.05	0.69 ± 0.06	0.87 ± 0.03	0.37 ± 0.07
ei	1.700000	MLP	en	cambridge	0.62 ± 0.03	0.61 ± 0.04	0.85 ± 0.02	0.46 ± 0.03

Continued on next page

LLMtype	LLMsize	MLmodel	lang	corpus	Acc	F1	QWK	MAE
ei	1.700000	LinearSVM	en	cefr	0.53 ± 0.01	0.44 ± 0.02	0.68 ± 0.01	0.55 ± 0.02
ei	1.700000	LogReg	en	cefr	0.57 ± 0.01	0.48 ± 0.02	0.75 ± 0.01	0.47 ± 0.01
ei	1.700000	MLP	en	cefr	0.58 ± 0.01	0.49 ± 0.02	0.76 ± 0.01	0.44 ± 0.01
ei	1.700000	LinearSVM	en	elg	0.61 ± 0.05	0.63 ± 0.06	0.81 ± 0.04	0.47 ± 0.08
ei	1.700000	LogReg	en	elg	0.66 ± 0.06	0.65 ± 0.07	0.86 ± 0.02	0.38 ± 0.06
ei	1.700000	MLP	en	elg	0.64 ± 0.04	0.59 ± 0.06	0.85 ± 0.02	0.41 ± 0.06
ei	1.700000	LinearSVM	en	icle500	0.49 ± 0.03	0.42 ± 0.05	0.51 ± 0.08	0.63 ± 0.05
ei	1.700000	LogReg	en	icle500	0.52 ± 0.02	0.45 ± 0.05	0.58 ± 0.04	0.58 ± 0.02
ei	1.700000	MLP	en	icle500	0.53 ± 0.07	0.45 ± 0.08	0.58 ± 0.07	0.57 ± 0.08
ei	1.700000	LinearSVM	es	caes	0.99 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.02 ± 0.0
ei	1.700000	LogReg	es	caes	0.99 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.02 ± 0.0
ei	1.700000	MLP	es	caes	0.99 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.02 ± 0.0
ei	1.700000	LinearSVM	es	kwiqiz	0.35 ± 0.07	0.33 ± 0.07	0.52 ± 0.13	0.92 ± 0.12
ei	1.700000	LogReg	es	kwiqiz	0.47 ± 0.06	0.47 ± 0.06	0.64 ± 0.06	0.69 ± 0.09
ei	1.700000	MLP	es	kwiqiz	0.38 ± 0.09	0.38 ± 0.1	0.52 ± 0.05	0.87 ± 0.12
ei	1.700000	LinearSVM	fr	kwiqiz	0.52 ± 0.05	0.4 ± 0.05	0.46 ± 0.11	0.6 ± 0.07
ei	1.700000	LogReg	fr	kwiqiz	0.54 ± 0.05	0.42 ± 0.04	0.56 ± 0.11	0.53 ± 0.06
ei	1.700000	MLP	fr	kwiqiz	0.43 ± 0.04	0.3 ± 0.07	0.35 ± 0.12	0.73 ± 0.07
ei	1.700000	LinearSVM	fr	readme	0.38 ± 0.04	0.37 ± 0.04	0.6 ± 0.06	0.87 ± 0.06
ei	1.700000	LogReg	fr	readme	0.44 ± 0.04	0.43 ± 0.05	0.74 ± 0.03	0.68 ± 0.05
ei	1.700000	MLP	fr	readme	0.45 ± 0.02	0.44 ± 0.02	0.77 ± 0.03	0.65 ± 0.04
ei	1.700000	LinearSVM	hi	readme	0.34 ± 0.03	0.34 ± 0.03	0.58 ± 0.03	1.09 ± 0.04
ei	1.700000	LogReg	hi	readme	0.37 ± 0.05	0.37 ± 0.05	0.69 ± 0.02	0.92 ± 0.06
ei	1.700000	MLP	hi	readme	0.39 ± 0.02	0.38 ± 0.02	0.69 ± 0.02	0.9 ± 0.05
ei	1.700000	LinearSVM	pt	cople2	0.74 ± 0.05	0.71 ± 0.07	0.68 ± 0.08	0.59 ± 0.1
ei	1.700000	LogReg	pt	cople2	0.79 ± 0.04	0.76 ± 0.07	0.78 ± 0.05	0.42 ± 0.08
ei	1.700000	MLP	pt	cople2	0.28 ± 0.03	0.25 ± 0.03	0.62 ± 0.04	1.01 ± 0.06
ei	1.700000	LinearSVM	pt	peapl2	0.61 ± 0.04	0.54 ± 0.04	0.68 ± 0.03	0.5 ± 0.04
ei	1.700000	LogReg	pt	peapl2	0.61 ± 0.03	0.54 ± 0.06	0.66 ± 0.05	0.52 ± 0.04
ei	1.700000	MLP	pt	peapl2	0.61 ± 0.03	0.54 ± 0.05	0.65 ± 0.08	0.54 ± 0.06

Table 4: Results of all model