Probing Native-Language Signals in Learner English Representations

Anonymous ACL submission

Abstract

L1 transfer is a well-known source of variation in learner English. Native Language Identification has shown that L1-predictive information is present in learner text, but high classification accuracy alone does not reveal whether this signal reflects linguistic transfer or corpus-specific surface cues. We examine whether L1-associated variation is recoverable from frozen encoder representations, and whether the signal remains stable across corpora, topic shifts, and grammatical correction. We probe five encoders on essays from Write & Improve 2024 and EFCAMDAT, two learner corpora with different corpus designs. Across both corpora, all five encoders recover L1-associated signals well above baseline. Monolingual and multilingual encoders perform comparably, with no consistent advantage for multilingual pretraining. The signal remains measurable under topic-heldout evaluation and grammatical correction, indicating that it is not limited to prompt vocabulary or overt grammatical errors. In Write & Improve, stronger L1 separability aligns with specific error categories for particular L1 groups, connecting representation geometry to interpretable patterns in learner production.

1 Introduction

When second language learners produce English, their first language leaves detectable traces in the text. Native Language Identification (NLI) research has established that learner English contains recoverable signals of a writer's first language (Tetreault et al., 2013; Malmasi and Dras, 2015). What remains unclear is what kind of information makes L1 prediction possible. The signal may reflect linguistic transfer, corpus-specific cues, or some combination of the two (Uluslu et al., 2025).

The difficulty is that NLI accuracy is not self-explanatory. High performance may reflect transfer-related patterns in learner production, but it may also reflect topic vocabulary, prompt-specific writing patterns, cultural references, or other corpus-specific cues. A signal tied to prompt vocabulary or surface grammatical errors would say little about stable L1-associated variation. Less attention has been paid to whether this variation remains recoverable from frozen encoder representations across corpora, and whether the recoverable signal connects to interpretable patterns in learner production.

We focus on the representational status of this signal: where it appears in pretrained encoders and how stable it is across changes in corpus, topic, and surface form. Probing studies have shown that grammatical and semantic properties are accessible across layers of pretrained encoders (Tenney et al., 2019; Belinkov, 2022). We use this framework to ask whether L1-associated variation is also accessible in frozen representations of learner English.

We probe five pretrained encoders on essays from two learner corpora: Write & Improve 2024 (Nicholls et al., 2024), which includes self-reported L1 labels, and EFCAMDAT (Geertzen et al., 2013), where the L1 labels are corpus-defined proxies based on recorded nationality. The two-corpus design tests whether L1 decodability is stable across different corpus settings. We then evaluate two control conditions, topic-heldout evaluation and grammatical correction, and examine whether stronger L1 separability is associated with grammatical error profiles. Figure 1 summarizes the two-corpus probing pipeline and its three analysis components.

Across both corpora, all five encoders show consistent L1 decodability. Monolingual and multilingual encoders perform comparably, suggesting that the signal does not depend on cross-lingual pretraining. Robustness analyses further show that the signal remains measurable under topic-heldout evaluation and grammatical correction, showing that it is not reducible to prompt topics or overt grammatical errors. In Write & Improve, stronger

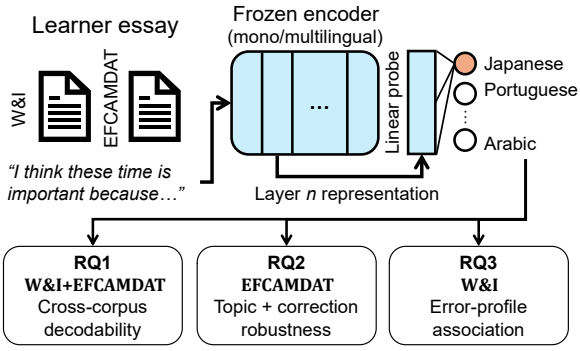

Figure 1: Experimental pipeline. Essays from two learner corpora are encoded by five frozen encoders; layer-wise mean-pooled representations are probed with a linear probe across three research questions.

L1 separability also aligns with error patterns that are consistent with known English–L1 structural contrasts.

This paper makes three contributions. First, we show that L1-associated signals are linearly decodable from frozen encoder representations across two learner corpora, establishing a representational basis for these signals beyond output-level NLI accuracy. Second, the signal withstands topic-heldout evaluation and grammatical correction, showing that it is not reducible to prompt vocabulary or overt grammatical errors. Third, in W&I, stronger L1 separability aligns with specific grammatical error profiles that correspond to known English–L1 structural contrasts, connecting representation geometry to interpretable patterns in learner production.

2 Related Work

2.1 Native Language Identification and Cross-Corpus Generalization

NLI classifies a writer's first language from their second-language production. Since the introduction of TOEFL11 (Blanchard et al., 2013; Tetreault et al., 2013), the task has been studied with feature-engineered classifiers (Malmasi and Dras, 2015; Malmasi et al., 2017) and more recently zero-shot approaches using large language models (Zhang and Salle, 2023). Strong performance on standard benchmarks shows that learner writing contains recoverable L1-associated signals. However, high within-corpus accuracy does not guarantee that these signals reflect stable linguistic transfer: NLI systems can exploit topic-correlated cues, including cultural references, proper nouns, and prompt-specific vocabulary that happen to align with L1 labels (Uluslu et al., 2025).

Cross-corpus evaluation is therefore an important test case. Malmasi and Dras (2015) showed that cross-corpus performance is substantially lower than within-corpus accuracy, suggesting that some learned features are corpus-specific rather than reflecting stable transfer properties. This raises a representation-level question: whether L1-associated signals remain recoverable across different corpus designs.

2.2 Probing Encoder Representations

Probing classifiers test what linguistic information is linearly recoverable from pretrained encoder representations. Prior work has identified layer-wise patterns for syntactic and semantic properties in BERT-family models (Tenney et al., 2019; Hewitt and Manning, 2019; Rogers et al., 2020), and probe accuracy is interpreted as evidence that a property is accessible in the representation under the chosen classifier family (Belinkov, 2022; Agarwal et al., 2025).

Prior work has shown that multilingual encoders such as mBERT and XLM-R retain language identity and typological information across layers (Libovický et al., 2020; Zheng and Liu, 2022), but whether this carries over to learner-produced English, where all inputs share the same surface language, is less clear.

2.3 L1 Transfer and Learner Error Profiles

L1 transfer refers to systematic influence from a learner's first language on second-language production, appearing in morphosyntactic patterns, lexical choice, and discourse organization (Odlin, 1989; Ellis et al., 1994; Jarvis and Pavlenko, 2008). Large learner corpora such as EFCAMDAT (Geertzen et al., 2013) and more recently LENS (Acharya et al., 2025) have made it possible to study learner production across L1 backgrounds, proficiency levels, and populations. Recent work has shown that LLMs can reproduce such patterns in generated text (Gao et al., 2025), though it is less clear whether comparable patterns are recoverable from the representations of actual learner writing.

Grammatical error profiles help make the probing result more interpretable. Error categories such as article use, verbal morphology, and prepositions are often related to structural differences between English and learners' first languages (Chrabaszcz and Jiang, 2014). Large-scale corpus work has also

shown that L1 background affects English article use across proficiency levels (Öksüz et al., 2025). These findings suggest that L1-associated decodability may be linked to systematic error patterns, not only to topic or learner identity.

We therefore ask two questions: whether the signal remains recoverable after grammatical correction, and whether stronger L1 separability is associated with error profiles consistent with English–L1 structural contrasts.

3 Methods

3.1 Learner Corpora and Label Sources

We use two learner English corpora that differ in how L1 labels are assigned. The two-corpus design tests whether L1-associated signals in encoder representations remain recoverable across different label sources and corpus settings.

Write & Improve 2024. Write & Improve 2024 (Nicholls et al., 2024) is a collection of English essays submitted to an online writing feedback platform. Each essay includes a self-reported L1 label and a CEFR proficiency rating. We use first-version essays, learners' initial unrevised submissions, because they reflect unedited production prior to any platform feedback or revision. Restricting to the nine labels in the shared inventory yields 3,077 essays from 410 learners. Because many learners contribute multiple essays, all probing experiments use learner-level group splits to prevent essays from the same learner appearing in both training and test folds.

EFCAMDAT. EFCAMDAT (Geertzen et al., 2013) is a large-scale learner corpus from an online English learning platform. EFCAMDAT does not provide self-reported individual L1 labels. Instead, it provides a corpus-defined L1 variable that groups learners by the dominant language associated with their recorded nationality: Brazilian learners are grouped as Portuguese, Saudi Arabian learners as Arabic, and so on. This is a population-level grouping variable, not a claim about any individual learner's native language. EFCAMDAT uses 15 internal proficiency levels that map onto CEFR bands: levels 1–3 correspond to A1, 4–6 to A2, 7–9 to B1, 10–12 to B2, and 13–15 to C1. For sampling, we combine the B2 and C1 ranges into a single upper-proficiency band because the highest levels contain fewer essays. We then sample 600 essays for each label in each of four proficiency bands, yielding 21,600 essays in total.

	W&I	EFCAMDAT
Label source	self-reported	nationality-based proxy
Labels	9	9
Essays	3,077	21,600
Sampling	N/A	stratified by label and proficiency

Table 1: Dataset statistics for the two learner corpora.

Aligned-9 L1 label inventory. We retain the nine L1 labels present in both corpora with sufficient essays to support learner-level cross-validation: Portuguese, Chinese, Spanish, Arabic, German, Italian, French, Japanese, and Turkish. Table 1 summarizes the two corpus configurations.

Error annotation. For the error-profile analysis in Section 4.3, we use W&I M2 annotations produced with the ERRANT framework (Bryant et al., 2017). EFCAMDAT error annotations are available only for a subset of essays and use a different annotation scheme; we therefore restrict the error-profile analysis to W&I. We aggregate ERRANT error types into four grammatically interpretable categories: verb morphology (VERB:TENSE, VERB:FORM, VERB:SVA), article and determiner use (DET), prepositions (PREP), and word order (WO). These categories correspond to grammatical dimensions extensively linked to L1 structural influence in SLA research (Jarvis and Pavlenko, 2008; Chrabaszcz and Jiang, 2014; Öksüz et al., 2025). Spelling, punctuation, and other surface-level categories are excluded. Error rates are normalized as counts per 100 words.

3.2 Encoder Representations and Probing

We probe five pretrained transformer encoders that span monolingual and multilingual pretraining: English BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), multilingual BERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and mDeBERTa-v3 (He et al., 2023). This comparison tests whether L1-associated signals in learner English are more recoverable in multilingual encoders, or are also accessible in English-only encoders.

For each essay, we extract hidden states from every transformer layer and compute a single essay representation by mean-pooling token-level hidden states, truncating input sequences to 256 tokens. Mean pooling provides an architecture-agnostic summary of the full input without relying on a dedicated classification token, making it well-suited to comparing encoders with different pretraining

objectives (Reimers and Gurevych, 2019). CLS-pooling results are reported in Appendix B.1.

At each layer, we train a multinomial logistic regression probe to predict the L1 label from the pooled representation. We use a linear probe to keep the classifier capacity limited and focus the analysis on information recoverable from the frozen representation (Belinkov, 2022).

Evaluation uses stratified 5-fold cross-validation with folds defined by learner ID, so no learner contributes essays to both train and test. W&I labels are unevenly distributed across L1 groups, ranging from 152 essays (German) to 834 (Spanish); class weights are set inversely proportional to label frequency (full distribution in Appendix A.1).

We report macro-F1 as the primary metric and identify a peak layer per encoder–corpus combination as the layer with the highest macro-F1. This peak-layer representation is used in the error-profile regression of Section 4.3. The peak layer here is selected based on mean macro-F1 across all five cross-validation folds.

We evaluate against four baselines. Two compare encoder probes to surface-feature classifiers: a TF-IDF baseline uses unigram lexical features, and a metadata-only baseline uses essay length and proficiency level. These show how much L1 information can be recovered from surface text and coarse metadata alone.

The random-label baseline shuffles L1 labels while keeping representations fixed, checking whether above-chance performance requires a consistent mapping between labels and representations. The shuffled-representation baseline permutes the dimensions of each vector independently per essay (Belinkov, 2022), checking whether above-chance performance requires the dimensional arrangement of the representations, not just their value distribution. Both controls are expected to perform near chance.

3.3 Robustness Settings

We evaluate robustness in two EFCAMDAT settings. The topic-heldout setting tests whether L1 decodability persists when test prompts are unseen during training. We reserve 25 topics for testing and enforce no learner overlap between train and test; the remaining 95 topics form the training pool. The same linear probe and TF–IDF baseline are trained on the training split and evaluated on the held-out topics.

The grammatical-correction setting tests how much L1 decodability remains after overt grammatical errors are corrected. We use the original–corrected matched subset from the Cleaned Error-coded Subcorpus (Öksüz et al., 2025). For each essay pair, we extract encoder representations for both the original and the corrected version and compare probe performance under the same label inventory and evaluation setup.

3.4 Reproducibility Details

All experiments use frozen encoder representations with no fine-tuning. Hidden states are extracted from every transformer layer of each encoder. Encoder checkpoints are `bert-base-cased` (English BERT), `roberta-base` (RoBERTa), `bert-base-multilingual-cased` (mBERT), `xlm-roberta-base` (XLM-R), and `microsoft/mdeberta-v3-base` (mDeBERTa-v3), all loaded from the Hugging Face `transformers` library.

We implement the probe using `scikit-learn` `LogisticRegression` with L2 regularization ($C = 1.0$), balanced class weights, and `random_state=42`.

Hidden-state extraction was run on an NVIDIA RTX 3060 GPU (12 GB VRAM); probing, regression, and analysis were run on CPU. The software environment used Python 3.10 with PyTorch 2.11, `transformers` 5.8, and `scikit-learn` 1.7. Full dataset and split statistics are reported in Appendix A.

4 Results

The results are organized around three research questions: L1 decodability across corpora (RQ1), robustness under topic shift and grammatical correction (RQ2), and the association between L1 separability and grammatical error profiles (RQ3).

4.1 RQ1: Cross-Corpus L1 Decodability

All five encoders recover L1-associated signals well above both control baselines(random-label, shuffled-representation) on both corpora, as shown in Table 2. On EFCAMDAT, peak macro-F1 ranges from 0.479 to 0.517; on W&I, it ranges from 0.344 to 0.370. The weaker signal on W&I is consistent with that corpus being roughly seven times smaller and more imbalanced, with label counts ranging from 152 essays for German to 834 for Spanish.

Metadata alone (essay length and proficiency) does not predict L1 label above chance on either

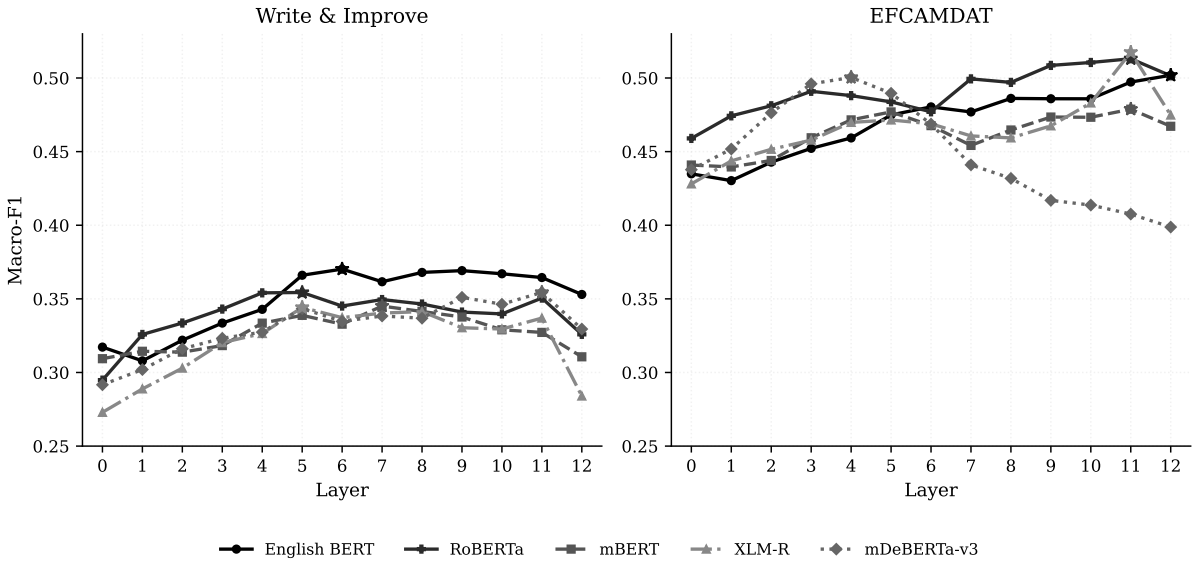

Figure 2: Layer-wise macro-F1 on W&I (left) and EFCAMDAT (right). Stars mark peak layers.

Model	W&I	EFCAMDAT
English BERT	**0.370**	0.502
RoBERTa	0.354	0.513
mBERT	0.345	0.479
XLM-R	0.344	**0.517**
mDeBERTa-v3	0.354	0.500
TF-IDF	0.305	0.451
Metadata only	0.099	0.084
Random label	0.088–0.108	0.113–0.116
Shuffled repr.	0.099–0.106	0.108–0.112

Table 2: Peak macro-F1 on W&I and EFCAMDAT. TF-IDF and metadata-only rows report learner-grouped CV results. Random-label and shuffled-representation baselines are shown as ranges across encoders.

corpus (W&I 0.099, EFCAMDAT 0.084), showing that these covariates alone do not account for L1 decodability. A TF-IDF classifier reaches 0.305 on W&I and 0.451 on EFCAMDAT, showing that lexical patterns carry some L1 information. Encoder probes exceed TF-IDF on both corpora, and the gap widens under topic-heldout evaluation (Section 4.2), suggesting that encoder representations capture L1-associated patterns beyond topic vocabulary.

Monolingual and multilingual encoders perform comparably. On W&I, English BERT leads (0.370), with RoBERTa and mDeBERTa-v3 tied at 0.354. On EFCAMDAT, XLM-R leads marginally (0.517), followed by RoBERTa at 0.513 and English BERT at 0.502. No consistent multilingual advantage appears in either corpus, suggesting that the recoverable signal does not depend on cross-lingual pretraining.

Figure 2 shows that decodability is not confined to a single layer. For English BERT, RoBERTa, mBERT, and XLM-R, peaks fall in upper layers on EFCAMDAT (layers 11–12) and in lower-mid layers on W&I (layers 5–7). mDeBERTa-v3 shows the opposite pattern, peaking at layer 4 on EFCAMDAT and layer 11 on W&I. Because the best-performing layer varies across both encoders and corpora, we report results from the peak layer for each encoder–corpus pair rather than fixing a single layer for all models.

Per-label F1 varies substantially across L1 groups (Table 3). Arabic and Japanese rank at or near the top in both corpora across encoders, even though they contribute far fewer essays in W&I (396 and 326, respectively) than in EFCAMDAT (2,400 each). French ranks near the bottom in both corpora. The similar rankings across W&I and EFCAMDAT suggest that per-label decodability reflects group-level patterns in learner English, not only corpus-specific factors.

In EFCAMDAT, where all nine labels contribute 2,400 essays each, Portuguese has the lowest per-label F1 (0.423–0.454), so sample size alone does not determine decodability. Chinese shows the largest cross-corpus gap: F1 ranges from 0.222 to 0.301 on W&I but from 0.505 to 0.562 on EFCAMDAT. W&I Chinese is the smallest label group at 174 essays, which is consistent with the low W&I

Label	E-BERT	RoBERTa	mBERT	XLM-R	mDeB
W&I					
Arabic	**0.511**	0.479	0.494	0.492	0.491
Chinese	0.259	0.222	0.249	0.250	**0.301**
French	0.223	**0.262**	0.191	0.248	0.241
German	0.397	0.395	0.320	0.380	**0.419**
Italian	0.327	**0.351**	0.298	0.312	0.302
Japanese	**0.532**	0.508	0.517	0.464	0.508
Portuguese	**0.372**	0.368	**0.372**	0.347	0.340
Spanish	0.473	0.476	**0.484**	0.470	0.437
Turkish	**0.389**	0.341	0.326	0.297	0.311
EFCAMDAT					
Arabic	0.515	0.554	0.472	**0.567**	0.544
Chinese	0.517	0.556	0.505	**0.562**	0.553
French	0.434	0.464	0.417	**0.473**	0.458
German	0.502	0.502	0.479	**0.509**	0.501
Italian	**0.490**	0.474	0.463	0.472	0.454
Japanese	0.570	**0.583**	0.546	0.570	0.532
Portuguese	**0.454**	0.445	0.423	0.444	0.447
Spanish	0.520	0.519	0.503	**0.537**	0.503
Turkish	0.518	0.524	0.503	**0.526**	0.512

Table 3: Per-label macro-F1 on W&I (upper) and EFCAMDAT (lower) under mean pooling at peak layer. Bold indicates best encoder per label.

Topic-heldout evaluation (EFCAMDAT, 25 held-out topics)			
Model	**Standard F1**	**Held-out F1**	**Δ**
XLM-R	0.517	0.448	−0.069
RoBERTa	0.513	0.440	−0.073
English BERT	0.502	0.434	−0.068
mDeBERTa-v3	0.500	0.434	−0.066
mBERT	0.479	0.409	−0.070
TF-IDF	0.451	0.360	−0.091
Correction robustness (EFCAMDAT, $N = 13{,}746$)			
Model	**Original F1**	**Corrected F1**	**Δ**
English BERT	0.504	0.473	−0.030
XLM-R	0.510	0.464	−0.046
RoBERTa	0.508	0.463	−0.045
mBERT	0.467	0.444	−0.023
mDeBERTa-v3	0.490	0.439	−0.052

Table 4: Robustness results on EFCAMDAT. The upper panel compares standard and topic-heldout evaluation. The lower panel compares original and grammatically corrected essays in a matched subset.

result, but the gap relative to other small-sample labels is wide. Chinese label-specific estimates in Section 4.3 carry greater uncertainty than other labels and should be interpreted cautiously.

4.2 RQ2: Robustness under Topic Shift and Correction

We test robustness in two settings, both using EFCAMDAT. The topic-heldout setting asks whether the signal persists when prompt-specific lexical cues from the test topics cannot be learned directly from training examples. The correction setting asks how much of the signal remains after overt grammatical errors are corrected.

Topic-heldout macro-F1 ranges from 0.409 to 0.448 across encoders, remaining 3.7–4.1 times the shuffled-representation baseline (Table 4). Performance drops by 0.066–0.073 from the full-topic results, indicating that topic-correlated cues contribute to L1 decodability. TF-IDF drops by 0.091 under the same shift, a larger decline than any encoder, suggesting that encoder representations retain L1-associated structure more robustly than surface lexical patterns when prompt topics change.

On the correction-matched subset ($N = 13{,}746$), corrected-text macro-F1 ranges from 0.439 to 0.473, approximately four times above the shuffled-representation baseline. The drop from original to corrected text ranges from 0.023 to 0.052, showing that surface grammatical errors carry some L1-associated information but do not account for the signal. Appendix B.3 shows that original and corrected representations remain highly similar across encoders.

In both robustness settings, decodability remains well above baseline. Topic-correlated vocabulary and overt grammatical errors each contribute to L1 decodability, but neither fully explains it. This motivates the next analysis, which asks whether particular grammatical error profiles are associated with stronger L1 separability.

4.3 RQ3: Error-Profile Association

Sections 4.1 and 4.2 show that the signal is recoverable across corpora and stable under surface controls. We now ask whether its strength varies with the grammatical error profile of the essay: do essays with higher rates of a particular error category receive stronger probe separability for their L1 group?

To measure probe separability at the essay level, we define the probe margin m_i as the difference between the logit the probe assigns to the correct L1 and the highest logit it assigns to any other L1:

$$m_i = s_i(y_i) - \max_{l \neq y_i} s_i(l) \qquad (1)$$

where $s_i(l)$ is the probe's logit for label l and y_i is the essay's true L1. A positive m_i means the probe ranked the correct L1 first; a larger value reflects stronger separability.

For each error category c, we regress m_i on the error rate (errors per 100 words, z-scored across essays), controlling for proficiency level and log essay length. The key parameter is the L1-specific slope $\beta_{y_i,c}$: within a given L1 group, does a higher

rate of category-c errors predict a higher probe margin? Formally:

$$m_i = \alpha_{y_i} + \beta_{y_i,c}\, z_{i,c} + \boldsymbol{\gamma}^\top \mathbf{p}_i + \delta \log w_i + \varepsilon_i \quad (2)$$

The intercept α_{y_i} captures the baseline margin for each L1 group. A positive $\beta_{y_i,c}$ means that essays with more category-c errors receive higher probe margins for that L1, so the probe separates them more strongly toward their true group. $\boldsymbol{\gamma}^\top \mathbf{p}_i$ is a vector of proficiency level fixed effects, w_i is essay word count, and ε_i is the error term.

Standard errors are clustered by learner (Colin Cameron and Miller, 2015) and Holm–Bonferroni correction is applied across label-specific slope tests within each encoder (Holm, 1979). Results are reported for English BERT, RoBERTa, and XLM-R.

The regression yields four label-specific patterns that survive Holm correction consistently across all three encoders (Figure 3, Table 5). The associations are selective: error rates predict stronger L1 separability not for essays with more errors overall, but for particular grammatical profiles within particular L1 groups. Arabic essays show positive associations with verb morphology, article/determiner, and preposition error rates; Turkish essays show a positive association with verb morphology errors. These patterns align with specific English–L1 contrasts discussed in prior work. Arabic differs from English in article use, including the absence of an indefinite article, and in many prepositional mappings; Arabic verbal morphology also differs from English (Chrabaszcz and Jiang, 2014; Öksüz et al., 2025; Jarvis and Pavlenko, 2008). Turkish verbal morphology encodes tense, agreement, and aspect through suffixal concatenation, contrasting with English verbal morphology (Jarvis and Pavlenko, 2008; Odlin, 1989). Word-order errors yielded no Holm-significant label-specific associations in W&I. The four positive patterns are tied to specific L1 groups rather than reflecting global error rates; adding prompt fixed effects does not remove them (Appendix B.5).

Several negative slopes are also significant after Holm correction, most notably for Chinese. Unlike the positive Arabic and Turkish patterns, these indicate lower true-L1 margins rather than stronger Chinese-specific separability: Chinese essays with more preposition and article errors are predicted less confidently as Chinese, with predictions dispersed across several competing labels rather than shifting toward a single alternative. Because Chinese is also the smallest W&I label group at 174 essays, we treat these effects as descriptive rather than primary evidence. Full label-specific results for all nine L1s are in Appendix B.4.

Pattern	E-BERT	RoBERTa	XLM-R
Arabic × verb morph	0.576	0.592	0.518
Arabic × article/det	0.448	0.509	0.559
Arabic × preposition	0.608	0.564	0.465
Turkish × verb morph	0.612	0.739	0.536

Table 5: Label-specific slopes ($\hat{\beta}$) surviving Holm correction across all three encoders. Confidence intervals are shown in Figure 3.

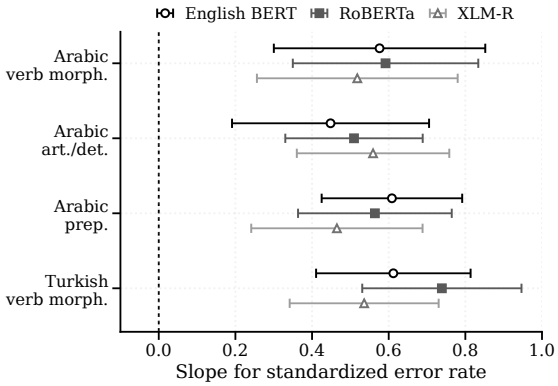

Figure 3: Label-specific associations between grammatical error rates and probe margin in W&I. Points show slopes for standardized error rates with 95% confidence intervals. Patterns shown are those consistent in sign across all three encoders; full results are in Appendix B.4.

5 Discussion

The three analyses together show that L1-associated variation is recoverable in learner English representations. NLI research has established that learner writing contains L1-predictive information, but classification accuracy alone does not indicate whether this signal is stable across corpora or robust to surface controls. By probing frozen encoder representations, we show that the signal remains linearly recoverable across two corpora and under topic and correction conditions.

The comparison between monolingual and multilingual encoders shows no consistent advantage for multilingual pretraining. Prior work has shown that multilingual models retain language identity when inputs come from typologically diverse languages (Libovický et al., 2020; Zheng and Liu, 2022). Our setting is different: all inputs are English learner

essays. The comparable performance of English-only and multilingual encoders suggests that the recoverable signal is accessible from systematic variation within learner English itself.

The signal persists after both topic shift and grammatical correction, showing that it is not fully explained by prompt vocabulary or overt grammatical errors. Topic-heldout evaluation lowers performance, and so does grammatical correction, but in both settings decodability remains well above baseline.

The correction and error-profile results should be read together. Correction lowers L1 decodability but does not remove it, showing that overt grammatical errors contribute to the signal without accounting for it. The W&I error-profile analysis identifies where part of that contribution is linguistically interpretable: for Arabic and Turkish learners, essays with higher rates of particular grammatical errors also receive stronger probe margins, in ways that correspond to known structural differences between English and those L1s. The association is observational, and the mechanism connecting representation geometry to surface error patterns remains to be established.

6 Conclusion

We examined whether L1-associated variation in learner English is recoverable from pretrained encoder representations, and whether it remains stable under topic shift and grammatical correction. Across five encoders and two corpora, L1-associated signals are decodable well above baseline, and monolingual encoders perform comparably to multilingual ones. The signal withstands topic-heldout evaluation and grammatical correction, showing that it is not reducible to prompt vocabulary or overt grammatical errors. In Write & Improve, stronger L1 separability is linked to specific grammatical error profiles consistent with English–L1 structural contrasts.

Future work can test the mechanisms behind these associations using interventional approaches, and extend the analysis to discourse-level features such as connective use and clause structure.

These findings establish that L1-associated variation leaves a stable, cross-corpus trace in learner-English representations, one that persists under surface controls and aligns with interpretable grammatical patterns in learner production.

Limitations

Three aspects of the present study bound the scope of its findings. First, the error-profile analysis in RQ3 is observational; the regression design does not distinguish between alternative causal explanations of the co-variation between representation geometry and learner error profiles, and Chinese estimates should be interpreted with particular caution given the small label group of 174 essays. Second, EFCAMDAT labels are corpus-defined proxies based on the dominant language of each nationality group, not self-reported individual L1; these results should be read as population-level corpus evidence. Third, more recent encoder-only architectures such as ModernBERT (Warner et al., 2025) were not included; whether the patterns we report extend to these models is a natural direction for future work.

Ethical Considerations

This study uses two publicly available learner corpora, Write & Improve 2024 and EFCAMDAT, both of which are anonymized and released for research purposes. No new data collection was conducted. The findings concern population-level patterns in learner writing and are not intended to profile or identify individual learners. Although the present analysis is conducted at the population level for research purposes, techniques for inferring demographic properties from text carry risks of misuse in assessment or surveillance contexts, and should not be applied to individual learners without appropriate safeguards.

Acknowledgments

Anonymized for review.

References

Poorvi Acharya, J Elizabeth Liebl, Dhiman Goswami, Kai North, Marcos Zampieri, and Antonios Anastasopoulos. 2025. Tracing l1 interference in english learner writing: A longitudinal corpus with error annotations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15157–15178.

Ananth Agarwal, Jasper Jian, Christopher D Manning, and Shikhar Murty. 2025. Mechanisms vs. outcomes: Probing for syntax fails to explain performance on targeted syntactic evaluations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33737–33757, Suzhou, China. Association for Computational Linguistics.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 1: Long Papers)*, pages 793–805.

Anna Chrabaszcz and Nan Jiang. 2014. The role of the native language in the use of the english nongeneric definite article by l2 learners: A cross-linguistic comparison. *Second Language Research*, 30(3):351–379.

A Colin Cameron and Douglas L Miller. 2015. A practitioner's guide to cluster-robust inference. *Journal of human resources*, 50(2):317–372.

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.

Rod Ellis, Yoshihiro Tanaka, and Asako Yamazaki. 1994. Classroom interaction, comprehension, and the acquisition of l2 word meanings. *Language learning*, 44(3):449–491.

Rena Wei Gao, Xuetong Wu, Tatsuki Kuribayashi, Mingrui Ye, Siya Qi, Carsten Roever, Yuanxing Liu, Zheng Yuan, and Jey Han Lau. 2025. Can llms simulate l2-english dialogue? an information-theoretic analysis of l1-dependent biases. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4355–4379.

Jeroen Geertzen, Theodora Alexopoulou, Anna Korhonen, and 1 others. 2013. Automatic linguistic annotation of large scale l2 databases: The ef-cambridge open language database (efcamdat). In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project*, pages 240–254.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

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.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Scott Jarvis and Aneta Pavlenko. 2008. *Crosslinguistic influence in language and cognition*. Routledge.

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shervin Malmasi and Mark Dras. 2015. Large-scale native language identification with cross-corpus evaluation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1403–1409.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75.

Diane Nicholls, Andrew Caines, and Paula Buttery. 2024. The write & improve corpus 2024: Error-annotated and cefr-labelled essays by learners of english.

Terence Odlin. 1989. *Language transfer*. Cambridge University Press.

Doğuş Öksüz, Theodora Alexopoulou, Kateryna Derkach, and Ianthi Maria Tsimpli. 2025. The influence of l1 typology on the acquisition of the l2 english article: A large-scale corpus study. *Second Language Research*, page 02676583251395876.

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*, pages 3982–3992.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the association for computational linguistics*, 8:842–866.

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.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 48–57.

Ahmet Yavuz Uluslu, Tannon Kew, Tilia Ellendorff, Gerold Schneider, and Rico Sennrich. 2025. Robust native language identification through agentic decomposition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8409–8425.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

Wei Zhang and Alexandre Salle. 2023. Native language identification with large language models. *arXiv preprint arXiv:2312.07819*.

Jianyu Zheng and Ying Liu. 2022. Probing language identity encoded in pre-trained multilingual models: a typological view. *PeerJ Computer Science*, 8:e899.

A Dataset and Split Statistics

A.1 Write & Improve 2024

The aligned-9 subset of Write & Improve 2024 contains 3,077 essays from 410 learners across 50 prompts. Labels are self-reported by learners at submission time. Table 6 reports the essay and learner counts per L1 label. The distribution is uneven, ranging from 152 essays for German to 834 for Spanish; class weights are set inversely proportional to label frequency in all probing experiments.

Label	Essays	Learners
Spanish	834	59
Portuguese	596	55
Arabic	396	44
Japanese	326	48
Italian	253	33
French	192	38
Chinese	174	44
Turkish	154	40
German	152	49
Total	3,077	410

Table 6: W&I aligned-9 label distribution.

A.2 EFCAMDAT

The EFCAMDAT aligned-9 sample contains 21,600 essays from 14,479 learners across 120 topics. Labels are nationality-based corpus-defined L1 proxies. The Chinese label uses the `cn` nationality group only; `tw` is excluded. Table 7 shows the level-stratified sampling design: 600 essays are sampled per label per proficiency band, yielding 2,400 essays per label.

Level band	Essays
L1–3 (A1)	5,400
L4–6 (A2)	5,400
L7–9 (B1)	5,400
L10–15 (B2/C1)	5,400
Total	21,600

Table 7: EFCAMDAT level-stratified sampling.

A.3 Topic-Heldout Split

Table 8 summarizes the topic-heldout evaluation split used in Section 4.2. Twenty-five EFCAMDAT topics are held out for testing. Training examples are drawn from the remaining 95 topics after removing learners who appear in the held-out test set, ensuring no learner overlap between train and test.

Property	Value
Total topics	120
Held-out topics	25
Train essays	12,622
Test essays	4,970
Learner overlap	0
Test label range	534–566
Train label range	1,025–1,755
Empty label × level-band cells	0
Essays used in split	17,592

Table 8: Topic-heldout split statistics for EFCAMDAT.

A.4 Correction-Matched Subset

The correction robustness analysis uses a matched subset of EFCAMDAT essays for which both original and grammatically corrected versions are available. Table 9 summarizes the subset. Label counts range from 1,250 for Chinese to 1,786 for Arabic; this imbalance reflects uneven annotation coverage across labels (54.9–78.8%).

Property	Value
Essays	13,746
Labels	9
Min label count	1,250 (Chinese)
Max label count	1,786 (Arabic)
Split	learner-level group

Table 9: Correction-matched subset statistics for EFCAMDAT.

B Additional Results

B.1 CLS Pooling Sensitivity

Table 10 reports macro-F1 under CLS and mean pooling for all five encoders on both corpora. Mean pooling outperforms CLS pooling in all ten encoder–corpus combinations. The gap is largest for mDeBERTa-v3, with gains of 0.176 on EFCAMDAT and 0.121 on W&I.

Corpus	Model	CLS	Mean	Δ
EFCAMDAT	English BERT	0.451	0.502	+0.051
	RoBERTa	0.493	0.513	+0.020
	mBERT	0.425	0.479	+0.054
	XLM-R	0.470	0.517	+0.047
	mDeBERTa-v3	0.325	0.500	+0.176
W&I	English BERT	0.311	0.370	+0.059
	RoBERTa	0.301	0.354	+0.053
	mBERT	0.281	0.345	+0.064
	XLM-R	0.271	0.344	+0.073
	mDeBERTa-v3	0.233	0.354	+0.121

Table 10: CLS vs. mean pooling macro-F1. Δ = mean − CLS.

B.2 Per-Label F1

Tables 11 and 12 report per-label macro-F1 for all nine L1 labels on W&I and EFCAMDAT. Results are reported at the peak layer under mean pooling. Bold indicates the best encoder per label.

Label	E-BERT	RoBERTa	mBERT	XLM-R	mDeB
Arabic	**0.511**	0.479	0.494	0.492	0.491
Chinese	0.259	0.222	0.249	0.250	**0.301**
French	0.223	**0.262**	0.191	0.248	0.241
German	0.397	0.395	0.320	0.380	**0.419**
Italian	0.327	**0.351**	0.298	0.312	0.302
Japanese	**0.532**	0.508	0.517	0.464	0.508
Portuguese	**0.372**	0.368	**0.372**	0.347	0.340
Spanish	0.473	0.476	**0.484**	0.470	0.437
Turkish	**0.389**	0.341	0.326	0.297	0.311

Table 11: Per-label macro-F1 on W&I under mean pooling at the peak layer.

Label	E-BERT	RoBERTa	mBERT	XLM-R	mDeB
Arabic	0.515	0.554	0.472	**0.567**	0.544
Chinese	0.517	0.556	0.505	**0.562**	0.553
French	0.434	0.464	0.417	**0.473**	0.458
German	0.502	0.502	0.479	**0.509**	0.501
Italian	**0.490**	0.474	0.463	0.472	0.454
Japanese	0.570	**0.583**	0.546	0.570	0.532
Portuguese	**0.454**	0.445	0.423	0.444	0.447
Spanish	0.520	0.519	0.503	**0.537**	0.503
Turkish	0.518	0.524	0.503	**0.526**	0.512

Table 12: Per-label macro-F1 on EFCAMDAT under mean pooling at the peak layer.

B.3 Representation Similarity after Correction

We compare mean-pooled representations of original and corrected essays using cosine similarity. Table 13 reports mean cosine similarity at layer 0, at the peak probing layer, and at the final layer for each encoder. At layer 0 and the peak layer, original and corrected representations remain highly similar across all encoders (0.963–0.997). The drop in probe performance reported in Section 4.2 therefore occurs despite high overall representation similarity. High overall similarity, however, does not identify which representation dimensions carry the L1-associated signal.

B.4 RQ3 Full Label-Specific Results

Table 14 reports all label-specific slopes surviving Holm–Bonferroni correction ($p < 0.05$) in the W&I error-profile regression, across all three encoders and all four error categories. Positive values indicate that higher error rates are associated with stronger true-L1 margins for that label; negative values indicate the opposite. Patterns reported in the main text (Table 5) are those consistent in sign across all three encoders.

Model	L0 cos.	Peak L	Peak cos.	Final cos.
English BERT	0.963	12	0.978	0.978
RoBERTa	0.985	11	0.996	0.996
mBERT	0.966	11	0.985	0.956
XLM-R	0.970	11	1.000	1.000
mDeBERTa-v3	0.974	4	0.991	0.980

Table 13: Mean cosine similarity between original and corrected essay representations on the EFCAMDAT correction-matched subset. Peak L denotes the peak probing layer for each encoder.

Model	Category	Label	$\hat{\beta}$	Holm p
E-BERT	article/det	Arabic	+0.448	0.020
E-BERT	preposition	Arabic	+0.608	<0.001
E-BERT	preposition	Chinese	−0.620	0.001
E-BERT	verb morph	Arabic	+0.576	0.001
E-BERT	verb morph	Turkish	+0.612	<0.001
RoBERTa	article/det	Arabic	+0.509	<0.001
RoBERTa	article/det	Chinese	−0.369	0.038
RoBERTa	preposition	Arabic	+0.564	<0.001
RoBERTa	verb morph	Arabic	+0.592	<0.001
RoBERTa	verb morph	Turkish	+0.739	<0.001
XLM-R	article/det	Arabic	+0.559	<0.001
XLM-R	article/det	French	−0.347	0.046
XLM-R	preposition	Arabic	+0.465	0.002
XLM-R	verb morph	Arabic	+0.518	0.004
XLM-R	verb morph	Turkish	+0.536	<0.001

Table 14: Holm–Bonferroni corrected label-specific slopes ($p < 0.05$) from the W&I error-profile regression. All four error categories are included; word-order errors yielded no surviving slopes in any encoder.

B.5 Prompt Fixed-Effects Sensitivity for RQ3

Table 15 compares the base W&I error-profile regression with a sensitivity model that adds prompt fixed effects. The four focal label–category associations remain positive and significant after Holm correction in both models.

Model	**Category**	**Label**	**Base** $\hat{\beta}$	**FE** $\hat{\beta}$	**FE Holm** p
E-BERT	article/det	Arabic	$+0.448$	$+0.451$	0.017
E-BERT	preposition	Arabic	$+0.608$	$+0.592$	<0.001
E-BERT	verb morph	Arabic	$+0.576$	$+0.567$	0.001
E-BERT	verb morph	Turkish	$+0.612$	$+0.581$	<0.001
RoBERTa	article/det	Arabic	$+0.509$	$+0.514$	<0.001
RoBERTa	preposition	Arabic	$+0.564$	$+0.559$	<0.001
RoBERTa	verb morph	Arabic	$+0.592$	$+0.596$	<0.001
RoBERTa	verb morph	Turkish	$+0.739$	$+0.725$	<0.001
XLM-R	article/det	Arabic	$+0.559$	$+0.571$	<0.001
XLM-R	preposition	Arabic	$+0.465$	$+0.459$	0.003
XLM-R	verb morph	Arabic	$+0.518$	$+0.530$	0.002
XLM-R	verb morph	Turkish	$+0.536$	$+0.510$	<0.001

Table 15: Prompt fixed-effects sensitivity for the four focal W&I RQ3 patterns. FE denotes the model with prompt fixed effects.