The Impact of Editorial Intervention on Detecting Native Language Traces

Ahmet Yavuz Uluslu[1,2] **Mark Gales**[1] **Kate Knill**[1] **Gerold Schneider**[2]

[1]ALTA Institute, Department of Engineering, University of Cambridge
[2]Department of Computational Linguistics, University of Zurich

[1]{mjfg, kmk1001}@cam.ac.uk
[2]{ahmetyavuz.uluslu, gerold.schneider}@uzh.ch

Abstract

Native Language Identification (NLI) is the task of determining an author's native language (L1) from their non-native writing. With the advent of human-AI co-authorship, learner texts are routinely corrected and rewritten by large language models, fundamentally altering the linguistic features NLI approaches depend on. In this paper, we investigate the robustness of L1 traces across increasing degrees of editorial intervention. By processing 450 essays from the Write & Improve 2024 (W&I) corpus through varying levels of grammatical error correction and paraphrasing, we demonstrate that L1 attribution does not depend solely on surface-level errors. Instead, the detection models appear to leverage deeper L1-related features, including unidiomatic lexico-semantic choices and pragmatic transfer. We find that minimal edits preserve these structural traces and maintain high L1 attribution accuracy. In contrast, fluency edits and paraphrasing normalize these L1 features, leading to a sharp decline in performance.

1 Introduction

Native language identification (NLI) is the task of automatically identifying the native language (L1) of an individual based on a writing sample in a non-native language (L2). Cross-linguistic influence (CLI) is increasingly understood to be a persistent and likely inevitable characteristic of bilingual cognition (Uluslu and Murphy, 2026). Even highly proficient speakers retain unconscious traces of their L1 that are detectable across linguistic levels, from phonology to semantics to morphosyntax (Markov et al., 2022). Applications of NLI span several fields, including forensic linguistics—for example, authorship attribution in cybercrime (Perkins, 2021) and plagiarism detection (Wastl et al., 2025)—as well as second-language acquisition research focused on learner characteristics (Berti et al., 2023). While early NLI systems relied heavily on hand-crafted linguistic features and machine learning algorithms (Malmasi et al., 2017), the recent transition toward large language models (LLMs) has established a new state-of-the-art (SoTA) (Ng and Markov, 2025). This leap in zero-shot performance is often attributed to emergent capabilities from multilingual pretraining and related tasks, such as grammatical error detection (Zhang and Salle, 2023).

Widespread human-LLM co-authorship is forcing a fundamental shift in the data available for NLI (Huang et al., 2025). As non-native writers use generative AI tools to refine their work, learner errors that traditionally served as discriminative features for text analysis are becoming obsolete (Chen et al., 2025). This evolution raises an important question: as LLM assistance moves from error correction to comprehensive edits, can NLI systems still detect an author's native language? We investigate the robustness of these L1 traces across four levels of editorial intervention, ranging from unmodified learner writing to minimal GEC, unconstrained GEC incorporating fluency edits, and paraphrasing. Our findings reveal that L1 identification does not rely exclusively on grammatical errors; instead, NLI models can capture deeper linguistic phenomena, such as unidiomatic lexico-semantic choices and pragmatic transfer. While minimal-edit GEC leaves these underlying structures largely intact, paraphrasing acts as a normalizing filter that obscures L1-specific markers. Our results demonstrate that to remain viable in an era of widespread AI assistance, NLI research must develop methods to assess the degree of editorial intervention.

The rest of the paper is structured as follows. Section 2 reviews related work in NLI and the use of generative AI in writing. Section 3 describes our dataset and LLM editorial interventions. Section 4 details the experimental setup, Section 5 presents the quantitative results and qualitative feature analysis, and Section 6 concludes the paper.

2 Related Work

A recent survey of the NLI landscape indicates that many contemporary approaches have shifted toward adopting LLMs (Goswami et al., 2024). This aligns with broader trends in computational discourse analysis, where these models increasingly automate the inference of sociodemographic variables across digital platforms (Alizadeh et al., 2025). Subsequent research has investigated the performance and fine-tuning potential of various open-source models on NLI (Ng and Markov, 2025). Other investigations have introduced masking strategies to isolate specific structural elements (Uluslu and Schneider, 2025) and examined the consistency of model explanations (Uluslu et al., 2025). Additionally, studies have explored the capacity of these models to act as sophisticated feature extractors for detecting CLI—for instance, by tracing L1-induced word order or orthographic errors in L2 English directly back to L1 Arabic structures (Acharya et al., 2025).

While LLMs have advanced the analytical capabilities of NLI systems, their concurrent rise as writing assistants challenges the assumptions underlying the task. Non-native writers are increasingly adopting generative AI tools across a spectrum of editorial intervention, ranging from targeted grammar checks to paraphrasing (Yang and Li, 2024). From an NLI perspective, this shift can effectively neutralize the orthographic and morphosyntactic errors that historically served as the most discriminative features for L1 identification (Malmasi et al., 2017). To formalize this intervention, modern GEC is typically defined by two primary annotation paradigms (Bryant et al., 2023). Minimal edits are strictly constrained to enforcing basic grammaticality while maintaining faithfulness to the author's original meaning and syntax. Fluency edits extend beyond targeted corrections, introducing broader structural changes to improve overall sentence flow and native-like phrasing. As texts advance along this continuum toward paraphrasing, the author's lexical and syntactic fingerprints are systematically overwritten (Richburg et al., 2024). The exact trajectory of NLI performance across this range of L1 feature retention remains underexplored. This paper addresses that gap by quantifying the impact of textual modification on NLI accuracy across editorial interventions, from unmodified learner texts to minimal edits, fluency edits, and paraphrasing.

3 Data

3.1 W&I 2024 NLI Dataset

We evaluate our approach using a curated subset of the Write & Improve (W&I) 2024 corpus (Nicholls et al., 2024), which features manual minimal-edit GEC and CEFR-level annotations provided by human experts. To ensure our results remain directly comparable with existing literature, we construct our subset to mirror the linguistic diversity of the TOEFL11 dataset (Blanchard et al., 2013), the de facto standard benchmark for NLI. Specifically, we select essays from nine L1 backgrounds available in W&I that intersect with the TOEFL11 languages: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Spanish, and Turkish. We sample 50 examples from each L1, yielding a balanced evaluation corpus of 450 texts. We restrict our selection to essays containing at least 100 words, aligning with established thresholds in stylometric and authorship studies (Koppel et al., 2011). Furthermore, we only include essays assessed at CEFR level A2 or above, ensuring the texts exhibit sufficient structural coherence and vocabulary for L1 feature analysis (Knill et al., 2025). After filtering, the final evaluation corpus has a mean essay length of 193 ± 61 words.

Stage	Sentence
original	*Every English learner searches the way how to learn english while now a days most of learners searches online method.
manual-min-gec	Every English learner searches for the way to learn English while nowadays most learners search for an online method.
llm-gec	Every English learner searches for ways to learn English, and nowadays, most learners look for online methods.
llm-paraphrase	Every English learner seeks ways to improve their skills, and nowadays, most turn to online methods.

Table 1: Transformation of a W&I sentence from a Hindi L1 writer across editorial conditions. The auto-min-gec output is identical to the manual-min-gec output and is therefore omitted.

3.2 Editorial Interventions

To investigate how varying degrees of textual modification affect NLI performance, we evaluate four levels of editorial intervention, instantiated by five experimental conditions. The minimal-edit

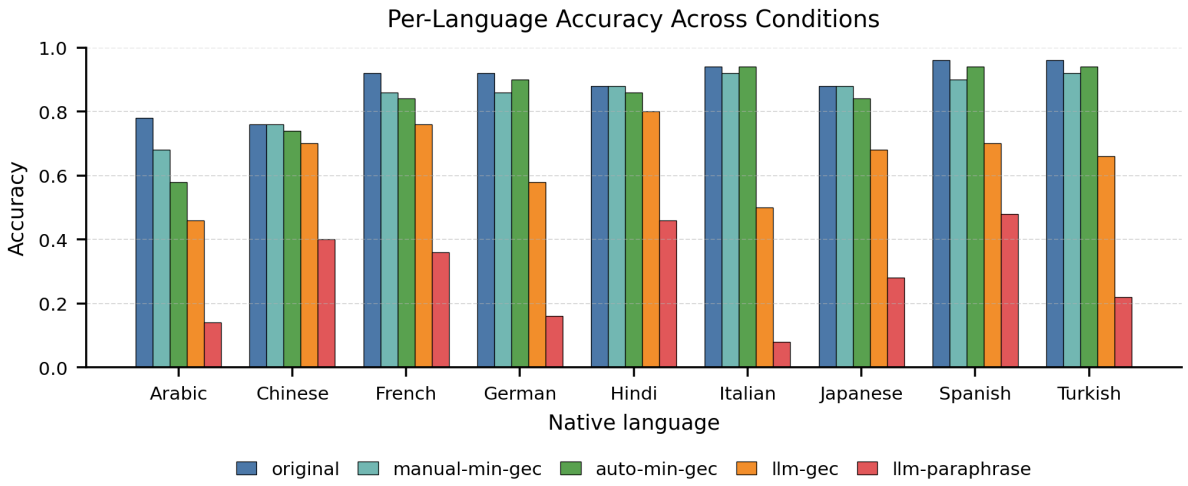

Figure 1: NLI accuracy scores across nine L1 backgrounds under varying degrees of editorial intervention.

level includes two parallel conditions—human expert correction and automatic minimal correction—allowing us to compare the two implementations directly. For our LLM-based interventions, we employ `gpt-4o-mini` as a representative baseline for a typical, widely accessible daily assistant capable of both GEC (Kovalchuk et al., 2025) and paraphrasing (Wang et al., 2025b). The exact prompts used for each condition are provided in Appendix A. The five experimental conditions are defined as follows (see Table 1 for an illustrative example):

- **original:** The unmodified learner texts extracted directly from the W&I corpus, serving as our baseline for maximum L1 feature retention.
- **manual-min-gec:** Expert-annotated minimal corrections provided with the W&I corpus (Nicholls et al., 2024).
- **auto-min-gec:** The learner texts processed by a SoTA minimal-edit system (Staruch et al., 2025) to determine if automated correction achieves parity with human expert annotations.
- **llm-gec:** A simple GEC setup in which the AI assistant is prompted to correct the grammatical errors in the text without edit restrictions. LLMs in this setting routinely over-correct, blending necessary fixes with subjective fluency edits (Zhan et al., 2026).
- **llm-paraphrase:** An unconstrained setup where the AI assistant is instructed to rewrite the text while preserving the original meaning and tone. This simulates real-world usage in which authors use AI assistants to refine their writing without completely abandoning its original style (Wang et al., 2025a).

4 Methodology

4.1 L1 Classification

We utilize OpenAI's `gpt-4o` for the NLI task, as it represents current SoTA performance (Goswami et al., 2025) and provides access to token-level probabilities. We additionally evaluate our approach using Qwen3.6-Plus (Qwen Team, 2026) to assess whether the observed pattern generalizes to another model. We employ a zero-shot prompting paradigm, with the exact prompt templates provided in Appendix A. By formulating the task as a multiple-choice selection that maps nine candidate languages plus an 'Other' category to letter identifiers, we apply constrained decoding to extract log probabilities and investigate model confidence (detailed in Appendix B). To mitigate position bias and obtain a more robust estimate of predictive uncertainty, we evaluate each instance across three randomized label orderings (Liusie et al., 2024), applying a softmax over the logits of the valid candidate tokens and averaging the resulting distributions; the class with the highest aggregate probability is selected as the final prediction. Furthermore, to mitigate topic bias and prevent reliance on content-based shortcuts, we mask named entities (including locations, nationalities, and personal names) and non-English words prior to classification (Uluslu et al., 2025).

4.2 Evaluation Metrics

We evaluate performance on the NLI task using standard classification accuracy. To characterize the extent and nature of textual modification across conditions, we compare each output with the human minimal-edit reference. Specifically, we report the MaxMatch (M^2) $F_{0.5}$ score (Dahlmeier and Ng, 2012), computed using the ERRANT toolkit (Bryant et al., 2017), together with Word Error Rate (WER). Additionally, we utilize BERTScore (Zhang et al., 2019) with a `microsoft/deberta-xlarge-mnli` backbone to assess semantic faithfulness. We use this metric to calculate the similarity between the manually corrected texts and the LLM-generated outputs (`llm-gec` and `llm-paraphrase`) and thereby assess the extent to which the original meaning is preserved.

5 Results

5.1 Quantitative Performance

Our experimental results, detailed in Table 2, show that NLI performance declines as the degree of editorial intervention increases. The model (`gpt-4o`) achieves an initial accuracy of 88.9% on the unmodified baseline texts. Applying minimal edits results in only a modest performance decline: accuracy decreases to 85.1% for human corrections and 84.2% for automatic corrections. This indicates that overt grammatical and orthographic errors are not the only cues supporting L1 identification.

Condition	M^2	WER%	Accuracy%
original	—	13.0	88.9
auto-min-gec	0.54	9.7	84.2
llm-gec	0.25	26.3	64.9
llm-paraphrase	0.11	54.2	28.7

Table 2: Comparison of text outputs with the human minimal-edit reference. M^2 reports $F_{0.5}$. Lower Word Error Rate (WER) indicates fewer edits relative to the reference. The `manual-min-gec` condition is omitted because it serves as the reference.

In the `llm-gec` condition, where the model applies subjective fluency edits alongside grammatical corrections, accuracy drops to 64.9%. In the unconstrained `llm-paraphrase` setting, accuracy decreases to 28.7%. Our supplementary evaluation using Qwen3.6-Plus demonstrates the same downward trend, with classification accuracy falling from 75.6% on the original learner texts to 25.3% under the paraphrasing condition. The results for this model are detailed in Table D.1. The decline in the performance of our primary model is consistent with the text-similarity metrics in Table 2 and the model-confidence scores in Table B.1. To further investigate the performance degradation in `llm-gec` and `llm-paraphrase` conditions, we provide the corresponding confusion matrices in Figure C.1. As the Word Error Rate (WER) increases and the M^2 score decreases, indicating heavier deviation from the reference text, NLI accuracy falls correspondingly. Importantly, our similarity analysis suggests that the outputs retain strong semantic similarity to the manually corrected references; BERTScore values remain high in both the `llm-gec` ($F_1 = 0.924$) and `llm-paraphrase` ($F_1 = 0.847$) conditions.

5.2 Qualitative Feature Analysis

As illustrated in Table 1, the original (unmodified) sentence from an L1 Hindi speaker contains several potential manifestations of cross-linguistic influence. At the surface level, it exhibits morphosyntactic errors, such as missing prepositions and articles (e.g., "searches online method") and subject–verb agreement problems ("most of learners searches"), which have been observed among English learners with an L1 Hindi background (Usama and Tarai, 2024). The minimal-edit systems (`manual-min-gec` and `auto-min-gec`) correct these grammatical errors while preserving much of the learner's syntactic structure. The non-native-like phrasing `[the way] (+) [how to learn english]` may reflect structural transfer from Hindi correlative constructions (Lipták, 2009). The `llm-gec` output performs a broader revision, replacing this structure with "searches for ways to learn English."

To enable a closer examination of the features affected by editorial intervention, Appendix E presents selected sentences from the same essay. The author repeatedly employs the discourse marker "As we know." This framing presents the following proposition as shared knowledge; its repeated use may reflect pedagogical templates found in some South Asian educational contexts and is compatible with discourse patterns described for Indian English (Valentine, 1991). Similarly, phrases such as "to live well in this world" preserve literal or conceptually extended formulations that may derive from underlying L1 expressions. These features remain visible in the `original`,

auto-min-gec, and llm-gec versions, and the model predicts Hindi for all three versions.

The llm-paraphrase condition, however, applies more extensive normalization. It reduces the repeated use of "As we know," replaces literal formulations with more idiomatic alternatives, and introduces new lexical choices such as "improve their skills" and "online methods." This process does not simply correct grammatical errors; it also removes the discourse framing and structural features that contribute to the text's profile. As a result, the classifier no longer predicts Hindi, instead predicting Chinese with low confidence.

5.3 Error Analysis

As the texts lose cues associated with the nine target classes, the model might be expected to assign an increasing proportion of texts to the 'Other' category. However, the confusion matrices (Appendix C) indicate that this shift happens gradually. Under the intermediate llm-gec condition, the model primarily re-ranks the existing L1 classes. It is only under the llm-paraphrase condition that we observe a marked increase in 'Other' predictions across all backgrounds. While this anticipated convergence is supported by the overall decrease in model confidence (Appendix B), the texts do not uniformly shift to this classification. Instead, gpt-4o exhibits a fallback behavior: when faced with high predictive uncertainty, the model frequently misclassifies the paraphrases as Spanish. Interestingly, a parallel evaluation using Qwen3.6-Plus demonstrates a similar fallback mechanism, but with Chinese as the most frequent prediction. We hypothesize that this phenomenon is driven by the models' pretraining corpora, which implicitly associate generic learner English with their most prevalent L2 demographics. This behavior may help explain why accuracy for Spanish remains relatively high at 48% even under the llm-paraphrase condition. The residual accuracy for Chinese (40%) and Hindi (46%) may have a different explanation. As discussed in the previous section, LLM-based rewriting does not uniformly alter every linguistic feature. In some outputs, syntactic structures, lexical choices, and discourse patterns remain intact. These artifacts likely extend beyond syntax to include cultural and pragmatic elements that escape entity redaction. Anecdotal passages detailing a writer's perspective or cultural practices are difficult to entirely neutralize.

6 Conclusion

In this paper, we investigated the resilience of L1-related features under increasing levels of editorial intervention. By evaluating NLI performance across a continuum of textual modifications, we demonstrate that L1 attribution with LLMs does not rely solely on grammatical errors. Our findings reveal that while GEC remediates morphosyntactic violations, it often leaves lexico-semantic and conceptual traces, enabling high attribution accuracy. However, as the degree of intervention increases toward fluency edits and paraphrasing, these persistent features are increasingly normalized. This process effectively overwrites the author's linguistic profile, reducing NLI accuracy from 88.9% to 28.7%. As a result, future NLI research must establish methods to assess editorial intervention and prioritize the detection of persistent patterns—such as conceptual and pragmatic transfer—that survive the automated editing process.

Limitations

One-Pass Intervention vs. True Co-Authorship Our experiments evaluate AI assistance as a single-pass, essay-level editorial intervention. This differs from authentic human-LLM co-authorship, which is highly iterative. In real-world scenarios, users often engage in a back-and-forth dialogue with the AI—rejecting specific edits, prompting for localized adjustments, or manually injecting their own phrasing and structural choices back into the revised text.

Restriction to L2 English While our dataset covers a diverse set of nine L1 backgrounds, our evaluation is limited to English as the target language. The capabilities of both GEC and NLI systems are often weaker for target languages other than English (Goswami et al., 2025). The dynamics of L1 feature erasure and the resilience of structural traces may manifest quite differently in non-English L2 contexts.

Ethics Statement

Our research exclusively utilized a publicly available L2 English learner corpus: the pseudonymized W&I corpus (Nicholls et al., 2024), which contains no personally identifiable information. We acknowledge the broad societal implications of authorship analysis, including potential risks to the security and privacy of individuals (Saxena et al., 2025). Our results are presented in a controlled experimental setting, primarily to study the impact of editorial intervention on existing NLI systems and explore approaches for enhancing their robustness. This work is not intended for deployment in critical real-world applications, such as forensic linguistics or automated academic integrity enforcement, without extensive further safeguards. As detailed in our Limitations section, we also recognize that our efforts to mitigate bias are not exhaustive, and further research is needed.

Acknowledgments

This work was supported by the collaboration between the University of Zurich and PRODAFT as part of the Innosuisse innovation project 103.188 IP-ICT conducted at Linguistic Research Infrastructure (LiRI). The project also received support from Cambridge University Press & Assessment, a department of The Chancellor, Masters, and Scholars of the University of Cambridge.

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 15146–15167, Suzhou, China. Association for Computational Linguistics.

Meysam Alizadeh, Fabrizio Gilardi, Zeynab Samei, and Mohsen Mosleh. 2025. Web-browsing LLMs can access social media profiles and infer user demographics. *arXiv preprint arXiv:2507.12372*.

Barbara Berti, Andrea Esuli, and Fabrizio Sebastiani. 2023. Unravelling interlanguage facts via explainable machine learning. *Digital Scholarship in the Humanities*, 38(3):953–977.

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.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, pages 643–701.

Fengchao Chen, Tingmin Wu, Van Nguyen, and Carsten Rudolph. 2025. SoK: Exposing the generation and detection gaps in LLM-generated phishing through examination of generation methods, content characteristics, and countermeasures. *arXiv preprint arXiv:2508.21457*.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 568–572.

Dhiman Goswami, Sharanya Thilagan, Kai North, Shervin Malmasi, and Marcos Zampieri. 2024. Native Language Identification in Texts: A Survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3149–3160.

Dhiman Goswami, Marcos Zampieri, Kai North, Shervin Malmasi, and Antonios Anastasopoulos. 2025. Multilingual native language identification with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas*

Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 193–199.

Baixiang Huang, Canyu Chen, and Kai Shu. 2025. Authorship attribution in the era of LLMs: Problems, methodologies, and challenges. *ACM SIGKDD Explorations Newsletter*, 26(2):21–43.

Kate M Knill, Diane Nicholls, Mark JF Gales, Mengjie Qian, and Pawel Stroinski. 2025. Introducing the Speak & Improve Corpus 2025: an L2 English speech corpus for language assessment and feedback. In *10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 167–171.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2011. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94.

Roman Kovalchuk, Mariana Romanyshyn, and Petro Ivaniuk. 2025. Introducing OmniGEC: A silver multilingual dataset for grammatical error correction. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 162–178, Vienna, Austria (online). Association for Computational Linguistics.

Anikó Lipták. 2009. The landscape of correlatives: An empirical and analytical survey. In *Correlatives Cross-Linguistically*, pages 1–46. John Benjamins Publishing Company.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151.

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.

Ilia Markov, Vivi Nastase, and Carlo Strapparava. 2022. Exploiting native language interference for native language identification. *Natural Language Engineering*, 28(2):167–197.

Yee Man Ng and Ilia Markov. 2025. Leveraging open-source large language models for native language identification. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 20–28.

Diane Nicholls, Andrew Caines, and Paula Buttery. 2024. *The Write & Improve Corpus 2024: Error-annotated and CEFR-labelled essays by learners of English*. Cambridge University Press & Assessment.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13991–14001.

Ria C Perkins. 2021. The application of forensic linguistics in cybercrime investigations. *Policing: A Journal of Policy and Practice*, 15(1):68–78.

Qwen Team. 2026. Qwen3.6-Plus: Towards real world agents.

Aquia Richburg, Calvin Bao, and Marine Carpuat. 2024. Automatic authorship analysis in human-AI collaborative writing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1845–1855.

Vageesh Saxena, Aurelia Tamò-Larrieux, Gijs Van Dijck, and Gerasimos Spanakis. 2025. Responsible guidelines for authorship attribution tasks in NLP. *Ethics and Information Technology*, 27(2):16.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Ryszard Staruch, Filip Gralinski, and Daniel Dzienisiewicz. 2025. Adapting LLMs for minimal-edit grammatical error correction. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 118–128, Vienna, Austria. Association for Computational Linguistics.

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 8398–8414, Suzhou, China. Association for Computational Linguistics.

Ahmet Yavuz Uluslu and Elliot Murphy. 2026. Neurocomputational mechanisms of syntactic transfer in bilingual sentence production. *arXiv preprint arXiv:2601.18056*.

Ahmet Yavuz Uluslu and Gerold Schneider. 2025. Investigating linguistic abilities of LLMs for native language identification. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 81–88, Tallinn, Estonia. University of Tartu Library.

Mohammad Usama and Shashikanta Tarai. 2024. A comparative study of errors in esl learners' writing: A cross-linguistic analysis. *Language in India*, 24(6).

Tamara M Valentine. 1991. Getting the message across: discourse markers in indian english. *World Englishes*, 10(3):325–34.

Zhengxiang Wang, Nafis Irtiza Tripto, Solha Park, Zhenzhen Li, and Jiawei Zhou. 2025a. Catch me if you can? not yet: LLMs still struggle to imitate the implicit writing styles of everyday authors. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10040–10055.

Zhilin Wang, Yafu Li, Jianhao Yan, Yu Cheng, and Yue Zhang. 2025b. Unveiling attractor cycles in large language models: A dynamical systems view of successive paraphrasing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12740–12755.

Michelle Wastl, Jannis Vamvas, and Rico Sennrich. 2025. Machine translation models are zero-shot detectors of translation direction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1054–1074.

Lu Yang and Rui Li. 2024. ChatGPT for L2 learning: Current status and implications. *System*, 124:103351.

Yuhao Zhan, Yuqing Zhang, Jing Yuan, Qixiang Ma, Zhiqi Yang, Yu Gu, Zemin Liu, and Fei Wu. 2026. JELV: A judge of edit-level validity for evaluation and automated reference expansion in grammatical error correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 41, pages 34611–34619.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

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

A Experimental Prompts

For the generative editorial interventions (**llm-paraphrase** and **llm-gec**), the model was initialized with the standard system prompt: "You are a helpful assistant." The L1 classification task utilized a specialized forensic expert persona.

A.1 LLM-Paraphrase Prompt

For the **llm-paraphrase** condition, the model was provided with the learner text and the following instruction to induce rewriting:

> "Rewrite the text to sound natural and fluent, preserving meaning and tone. Don't add new information. Return only the rewritten text. `\n\n{text}`"

A.2 LLM-GEC Prompt

For the **llm-gec** condition, the model was instructed to act as an unconstrained grammar corrector using the following prompt:

> "Improve my English grammar and fix my errors. Return only the corrected text, with no explanation `\n\n{text}`."

A.3 L1 Classification Prompt

For the L1 attribution task, the model was cast in an expert persona and provided with a multiple-choice selection of the nine candidate languages (represented by *{options}*). The exact prompt template was:

> "You are a forensic linguistics expert that reads essays from non-native speakers of English in order to identify their native language. Use clues such as word choice, syntactic patterns, and grammatical errors to decide. Analyze the text and choose the most likely native language from the following options:
>
> {options}
>
> (X) Other
>
> Reply with only the single letter (A, B, C, etc.) representing your choice."

B Model Confidence and Uncertainty

We characterize model confidence and uncertainty using three metrics: mean true-class probability, mean predicted-class confidence, and predictive entropy (Ovadia et al., 2019).

Condition	Acc%	$P(y)$	$P(\hat{y})$	Entropy
original	88.9	86.0	92.4	0.22
manual-min-gec	85.1	79.4	86.9	0.37
auto-min-gec	84.2	80.5	88.3	0.33
llm-gec	64.9	58.0	73.5	0.76
llm-paraphrase	28.7	25.1	56.4	1.20

Table B.1: Model performance and predictive uncertainty across experimental conditions. Metrics include Accuracy (Acc.), Mean True Probability $P(y)$, Mean Confidence $P(\hat{y})$, and Entropy (Ent.). Accuracy and probability scores are presented as percentages (%).

B.1 Mean Probability and Confidence

For our classification task, we average the predicted probabilities across option-order permutations. Let x represent the input text, y the correct target class, and $\hat{y}$ the model's predicted class. For a given instance x, we calculate the expected probability assigned to y by averaging across its three permutations. We define the mean true probability, $P(y)$, as this expected probability averaged across the entire test set. Similarly, we define the mean confidence, $P(\hat{y})$, as the expected probability assigned to $\hat{y}$ given x, also averaged across the test set. These metrics allow us to determine how confident the model remains in its predictions and in the correct attribution after varying levels of GEC and paraphrasing.

B.2 Predictive Entropy

To capture the overall uncertainty across the entire predictive distribution, we calculate the Shannon entropy (Shannon, 1948) over the averaged output probabilities. For a set of possible classes $\mathcal{Y}$, using the expected probability $\hat{p}(y \mid x)$ derived from the permutations, the predictive entropy H is defined as:

$$H(Y \mid x) = -\sum_{y \in \mathcal{Y}} \hat{p}(y \mid x) \log \hat{p}(y \mid x) \quad (1)$$

We compute entropy for each instance and report the mean across the evaluation set. Higher entropy values signify a flatter expected probability distribution, indicating that the model is highly uncertain and struggles to attribute the text to a specific L1.

C Confusion Matrices

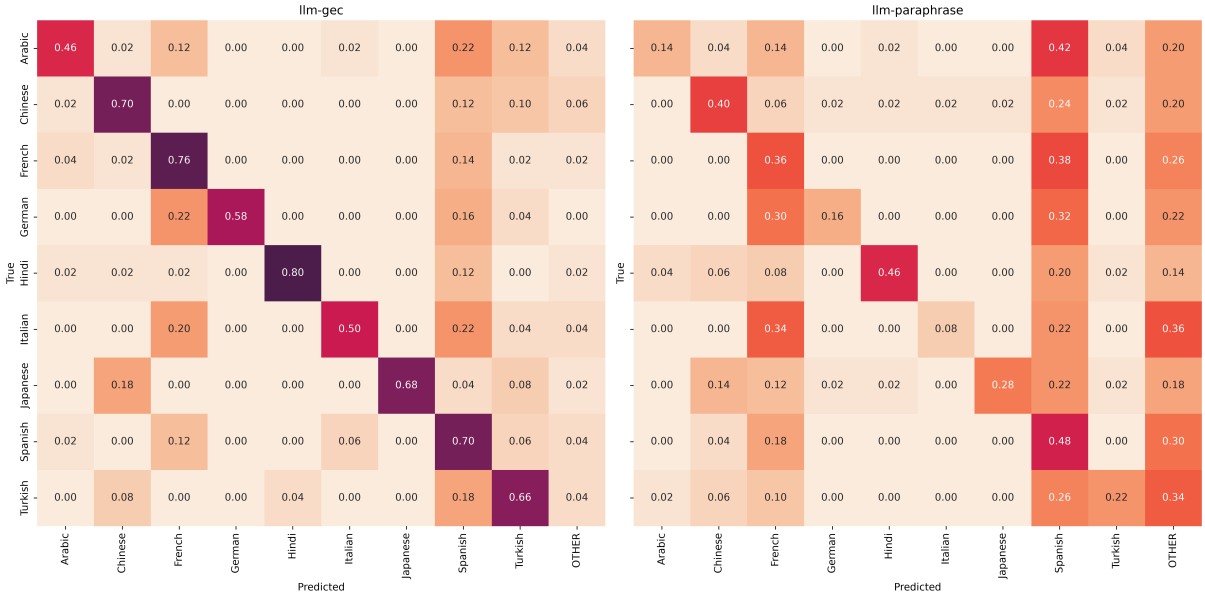

Figure C.1: Confusion matrices for GPT-4o L1 classification under the `llm-gec` and `llm-paraphrase` conditions.

D Supplementary Results

	Accuracy (%)	
Condition	**Qwen3.6-Plus**	**GPT-4o**
original	75.6	88.9
manual-min-gec	65.6	85.1
auto-min-gec	65.8	84.2
llm-gec	44.0	64.9
llm-paraphrase	25.3	28.7

Table D.1: Performance of Qwen3.6-Plus (calibrated) and GPT-4o across the five experimental conditions on the W&I 2024 dataset.

E Model Outputs

original *Classification: hindi*

(1) As we know English as secondary language is known as across word where it is not as first the language.

(2) As a language English is crucial to learn to live well in this world. A lot of issues which could be solved by knownig only English language.

(3) As we know, to learn any language, there are four parts, reading, writing, listening and speaking.

auto-min-gec *Classification: hindi*

(1) As we know, English as a secondary language is known as across word where it is not as the first language.

(2) As a language, English is crucial to learn to live well in this world. A lot of issues which could be solved by knowing only the English language.

(3) As we know, to learn any language, there are four parts: reading, writing, listening and speaking.

llm-gec *Classification: hindi*

(1) As we know, English as a secondary language is referred to as a foreign language when it is not the first language.

(2) As a language, English is crucial to learn in order to live well in this world. Many issues can be solved by knowing only the English language.

(3) As we know, to learn any language, there are four components: reading, writing, listening, and speaking.

llm-paraphrase *Classification: chinese*

(1) As we know, English is often referred to as a secondary language, meaning it is not the first language for many people.

(2) Learning English is essential for navigating the world effectively, as many issues can be resolved simply by knowing the language.

(3) To learn any language, we typically focus on four key areas: reading, writing, listening, and speaking.

Figure E.1: Selected sentences authored by a Hindi L1 speaker from the W&I 2024 corpus. The examples illustrate how various editorial interventions change the L1 classification (displayed in the top right of each panel).