Can We Still Hear the Accent? Investigating the Resilience of Native Language Signals in the LLM Era

Nabelanita Utami[1] Ryohei Sasano[1]
[1]Nagoya University
utami.nabelanita.v7@s.mail.nagoya-u.ac.jp sasano@i.nagoya-u.ac.jp

Abstract

The evolution of writing assistance tools driven by advances in large language models (LLMs) has changed how researchers write. This study investigates whether this shift is homogenizing research papers by analyzing native language identification (NLI) trends in ACL Anthology papers across three distinct eras: pre-neural network (NN), pre-LLM, and post-LLM. Due to the lack of data available, we first construct a labeled dataset using a semi-automated framework that includes abstract texts written by researchers with various linguistic backgrounds. Using the dataset, we further evaluate the presence of the homogenization by fine-tuning a general classifier to detect linguistic fingerprints of author backgrounds. Our analysis shows a decline in overall NLI performance over time that points toward a "native-like" pattern. Interestingly, the post-LLM era shows specific anomalies, including a rise in the distinctiveness of Chinese- or French-authored papers.

1 Introduction

The landscape of academic writing has changed over the past decade, driven by the rapid evolution of writing assistance tools. Researchers have moved from relying on simple dictionary lookups to neural machine translation (NMT). In the current era of large language models (LLMs), unlike previous tools, LLMs are capable of rewriting entire sections, smoothing out syntax, and suggesting vocabulary. While LLM-based writing assistants help non-native speakers to sound more "native-like," it has also raised the possibility of writing homogenization in academia. This phenomenon is similar to the "translationese" effect observed in machine-translated text, where output becomes simplified and standardized [1]. Recent studies have already noted the spread of "LLM-specific" vocabulary, such as the overuse of words that include "delve" or "showcases" [2]. This suggests that the unique linguistic voices of the authors are being replaced by a standardized English.

Native language identification (NLI), the task of predicting an author's native language (L1) based on their writing, serves as useful a metric to examine this shift. It was first formalized as a text classification problem demonstrated that an author's L1 could be predicted from their writing style with high accuracy [3]. Historically, NLI has focused on learner corpora (e.g., TOEFL essays) that relies on explicit grammatical errors to detect L1 interference [4]. However, applying NLI to academic writing presents a unique challenge since researchers are typically highly fluent in English. The "fingerprints" of native language are not obvious errors in this domain, but are subtle preferences often tied to the author's native language linguistical rules [5, 6]. In high-proficiency academic writing, this often manifests as "rhetorical transfer," a culturally distinct preference, rather than simple syntactic errors [7, 8].

In this work, we investigate the impact of writing assistance tools on the diversity of academic texts by analyzing NLI performance across three distinct technological eras: pre-neural network (NN), pre-LLM, and post-LLM. We hypothesize that as writing tools become more advanced, the signals of an author's native language will weaken. This indicates a shift towards a standardized English in modern academic writing. Our contributions are as follows:

1. We construct two datasets specifically for high-fluency academic writing, extracted from papers published in arXiv and the ACL Anthology, and then mapped to author demographics across three eras.
2. We analyze how the widespread use of LLMs is changing academic writing, specifically by measuring whether the unique traces of an author's native

language are disappearing or being replaced by a standardized English.

2 Dataset Construction

To analyze the evolution of scientific writing styles, we first constructed two datasets: a large-scale training set derived from arXiv, and a high-quality evaluation set derived from the ACL Anthology that we use as a testbed for the pre- and post-LLM analysis.

2.1 Semi-Automated Labeling

Given the lack of large-scale scientific paper datasets labeled with author native languages, we developed a semi-automated framework to produce high-confidence labels. Our pipeline utilizes an LLM-augmented labeling strategy that combines metadata with LLM-predicted name origins. Specifically, our labeling workflow proceeds in three stages. The first two stages involve estimating the author's country of origin, followed by a third stage in which the estimated country is mapped to a corresponding language label.

1) Author-Level Verification. For each paper, we retrieved author names and institutional affiliations via the OpenAlex API [9]. To resolve ambiguities caused by researcher migrations, we employed a cross-verification step:

- We prompted Qwen3-8B [10] to predict the top-2 most likely countries of origin based solely on the author's name. If certain, the model is allowed to output the same language twice.
- We intersected this prediction with the author's affiliation country. An author is assigned a country label *if and only if* their affiliation appears within the LLM's top-2 candidate list.
- Anglosphere exclusion: To minimize the influence of English immersion on L1 signals, we strictly excluded any non-native candidate who held a dual affiliation with an institution in an English-speaking country (e.g., an author affiliated with both *Tsinghua University* and *Harvard University*). We assume such authors possess high fluency that may conceal their native linguistic fingerprints.

2) Paper-Level Consensus. To ensure the text has a consistent native language signal, we require strict background coherence across co-authors. We restricted the dataset to papers with five or fewer authors. A final L1 label is assigned to the paper only if the key authors (first, second, and last) share the same verified country label.

3) Language Mapping. Finally, the verified country labels were mapped to their primary official languages (e.g., 'US' → `english_american`, 'CN' → `chinese`) using information obtained from Wikidata [11] references. In this study, to avoid ambiguity in the mapping process, we restrict language mapping to papers assigned a country label that has a primary official language and for which sufficient experimental data can be collected. The languages used in this study are described in Section 2.3.

2.2 Data Curation

Training Data (arXiv). To train a general classifier, we applied the framework to the arXiv dataset [12]. From the filtered results, we sampled a balanced subset across target languages to prevent class imbalance. To prevent the model from overfitting to the writing style of a specific time era, we applied a strict sampling cap per publication year. This number differs per language caused by the difference in quantity of samples.

Evaluation Data (ACL Anthology). To test our hypothesis regarding writing homogenization, we compiled a dataset from the ACL Anthology corpus [13] and divided it into three technological eras: pre-NN (≤2015), pre-LLM (2016–2022), and post-LLM (2023–2025). For the pre-NN and pre-LLM eras, we used the framework with additional manual verification. For the post-LLM era, due to the lack of updated dumps nor datasets, we performed a complete manual collection and verification of author backgrounds.

2.3 Data Statistics

We focused on eight target languages: American English, British English, French, German, Italian, Chinese, Japanese, and Korean. For the training set, which is derived from arXiv, we selected a balanced subset of 1,600 samples (200 per language) spanning the years 1999–2021.

As described above, we use data collected from the ACL Anthology as evaluation data. Since most of these papers have been published through a peer-review process, they exhibit a standardized level of writing quality. This allows our model to capture stylistic characteristics specific to authors' native languages (L1), rather than simple grammatical errors or differences in fluency. To construct a class-balanced dataset, we select 50 papers for each combination of three eras and eight languages, resulting in a total of 1,200 papers. For combinations with fewer than 50 available papers, such as Korean in the pre-NN era, we address this by duplicating a subset of the collected papers. We manually verified this dataset and found that almost all instances were assigned correct language labels.

3 Methodology

We introduce two NLI models: one based on few-shot prompting and the other based on fine-tuning. These models are applied to evaluation data from three eras. A relative performance drop on post-LLM-era data suggests that the advent of LLMs has reduced the presence of L1-specific stylistic traces in the English of non-native speakers.

3.1 Few-Shot Prompting

We prompted the models for classifying native languages of authors by restricting the output to a closed set of language labels. Since peer-reviewed academic text is highly fluent, standard models frequently default to predicting “native English.” As a solution, our system prompt explicitly directs the model to identify subtle L1-interference patterns and to avoid assigning English labels without any strong evidence.

3.2 Fine-Tuning

We fine-tuned two state-of-the-art open-weights models to evaluate their capability in detecting subtle stylistic fingerprints: Qwen3-14B and Gemma-3-12B-it [14]. We employed quantized low-rank adaptation (QLoRA) [15] for fine-tuning and both models were quantized to 4-bit precision (NormalFloat4) with double quantization. We froze the base model parameters and attached low-rank adapters (LoRA) [16] to the linear layers.

The training was performed on the balanced arXiv dataset constructed in Section 2. We used a maximum sequence length of 1024 tokens that properly covers the length of standard research paper abstracts. The specific hyperparameters used for each model are detailed in Table 1.

Table 1 Hyperparameter settings used in our experiment.

Hyperparameter	Qwen3-14B	Gemma-3-12B-it
Epochs	2	3
Batch Size	8	16
Gradient Accumulation	4	2
Learning Rate	1.0×10^{-3}	2.0×10^{-4}
Lora Rank (r)	16	16
Lora Alpha (α)	64	32
Lora Dropout	0.0001	0.1
Weight Decay	0	0.01

4 Experiments and Results

4.1 Experimental Setup

We evaluated our models on the ACL Anthology evaluation set described in Section 2. This set consists of balanced subsets for three eras: pre-NN (≤2015), pre-LLM (2016–2022), and post-LLM (2023–2025). As a baseline, we first evaluated the base models (Qwen3-14B and Gemma-3-12B-it) in an 8-shot setting (one example per language) without fine-tuning. We then evaluated our fine-tuned variants. We report both accuracy and F1-score. Given that the test sets are balanced, these metrics provide a direct measure of the model’s ability to differentiate L1 signals.

4.2 Results: Evidence of Homogenization

Table 2 presents the overall performance across the three eras. Although there is a substantial performance gap between the few-shot prompting–based model and the fine-tuning–based model, both models exhibit the same trend: the highest scores are obtained on pre-NN-era data, followed by pre-LLM-era data, with the lowest scores observed on post-LLM-era data.

Focusing on the results of the few-shot prompting–based model, when using Qwen3-14B, the accuracy drops from 37.8% on pre-NN-era data to 14.5% on post-LLM-era data. Similarly, when using Qwen2-7B, the accuracy drops from 30.4% to 19.1%. The models exhibited a strong bias toward American English, often collapsing into a single-class prediction, with an interesting choice of British English by Gemma 3 in post-LLM. This confirms that untuned LLMs cannot reliably detect subtle native language signals without targeted training.

Our fine-tuned models demonstrated significantly higher performance, achieving over 70% accuracy in the pre-NN era by both models. However, consistent with our hypoth-

Table 2 Performance comparison across three eras.

Metric	Qwen3-14B			Gemma-3-12B-it		
	pre-NN	pre-LLM	post-	pre-NN	pre-LLM	post-
	Few-Shot Prompting					
Accuracy	**0.378**	0.181	0.145	**0.304**	0.258	0.191
F1-Score	**0.393**	0.137	0.067	**0.304**	0.222	0.111
	Fine-Tuning					
Accuracy	**0.728**	0.650	0.633	**0.718**	0.628	0.590
F1-Score	**0.726**	0.637	0.623	**0.715**	0.614	0.598

Table 3 Detailed F1-scores per language for fine-tuned models. Bolded scores indicate the highest performance across three eras.

Lang.	Qwen3-14B			Gemma-3-12B-it		
	pre-NN	pre-LLM	post-	pre-NN	pre-LLM	post-
Eng-US	**0.648**	0.574	0.593	**0.679**	0.522	0.576
Eng-UK	**0.602**	0.438	0.406	**0.565**	0.435	0.235
French	0.703	**0.720**	0.690	0.667	0.654	**0.723**
German	**0.686**	0.604	0.612	**0.694**	0.559	0.581
Italian	**0.752**	0.732	0.703	**0.788**	0.739	0.688
Chinese	**0.876**	0.815	0.869	0.812	0.737	**0.885**
Japanese	0.758	0.553	**0.462**	**0.757**	0.615	0.508
Korean	**0.784**	0.657	0.628	**0.761**	0.651	0.590

esis, performance degraded significantly over time, i.e., from 72.8% to 63.3% for Qwen3 and from 71.8% to 59.0% for Gemma 3.

This monotonic decline indicates that the "fingerprints" of native language are becoming weaker in modern academic writing. Native language signals are disappearing because AI tools rewrite English phrasing unique to certain languages into standard American English. Detailed results including confusion matrices are provided in Appendix A.

4.3 Discussion

To understand more of this erosion, we analyzed class-specific performance. Table 3 shows the F1-scores per language for fine-tuned models. Three distinct patterns were observed.

Collapse of British English. The most noticeable drop occurred in British English. For Gemma 3, detection accuracy fell from 56.5% (pre-NN) to just 23.5% (post-LLM). Confusion matrices reveals that these misclassified papers are largely predicted as American English, suggesting a standardization of spelling and vocabulary toward US norms.

Loss of Asian L1 Signals. Japanese and Korean used to be distinct, but their classification accuracy dropped significantly in the post-LLM era. Japanese accuracy dropped from 75.8% to 46.2% for Qwen3, suggesting that LLMs are bridging the linguistic distance between Asian languages and English.

Persistence of L1 Signals in Chinese Text. Contrary to our predictions, French and Chinese remain highly detectable, with accuracy staying stable or even increasing. Unlike other languages, French and Chinese L1 signals appear resistant to homogenization. While a thorough examination of these factors is left for future work, the limited accessibility of Western APIs (e.g., OpenAI' s GPT series) in China may constitute one contributing factor affecting the Chinese language results.

5 Conclusion

In this work, we presented a study quantifying the impact of LLMs on the linguistic diversity of scientific writing. We constructed two native language identification datasets specifically tailored for high-fluency academic writing. By fine-tuning state-of-the-art open models (Qwen3 and Gemma 3), we were able to detect subtle linguistic shifts in academic writings. Our results provide an evidence that academic writing is undergoing a homogenization into standardized English. We confirmed that the accuracy of NLI on post-LLM era data is substantially lower than that on pre-NN era data. This suggests that the unique L1 signals that reflect a researcher's background are being faded by the effects of LLM-based writing assistance. However, when examining language-specific trends, the results are not entirely consistent; for French and Chinese, NLI exhibits higher performance on post-LLM data, in contrast to the overall trend.

References

[1] Vered Volansky, Noam Ordan, and Shuly Wintner. On the features of translationese. **Digital Scholarship in the Humanities**, Vol. 30, No. 1, pp. 98–118, 2015.

[2] Tom S Juzek and Zina B. Ward. Why does ChatGPT "delve" so much? Exploring the sources of lexical overrepresentation in Large Language Models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, **Proceedings of the 31st International Conference on Computational Linguistics**, pp. 6397–6411, January 2025.

[3] Moshe Koppel, Jonathan Schler, and Kfir Zigdon. Determining an author's native language by mining a text for errors. In **Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining**, pp. 624–628, 2005.

[4] Joel Tetreault, Daniel Blanchard, and Aoife Cahill. A report on the first native language identification shared task. In Joel Tetreault, Jill Burstein, and Claudia Leacock, editors, **Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 48–57, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[5] Yulia Tsvetkov, Naama Twitto, Nathan Schneider, Chris Dyer, and Karën Fort. Identifying the L1 of non-native writers: the CMU-Haifa system. In **Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 279–287, 2013.

[6] Scott Jarvis and Scott. A Crossley. **Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach**. Multilingual Matters Channel View Publications, 2012.

[7] Robert B Kaplan. Cultural thought patterns in inter-cultural education. **Language Learning**, Vol. 16, No. 1-2, pp. 1–20, 1966.

[8] Ken Hyland. Authority and invisibility: Authorial identity in academic writing. **Journal of Pragmatics**, Vol. 34, No. 8, pp. 1091–1112, 2002.

[9] Jason Priem, Heather Piwowar, and Richard Orr. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, 2022.

[10] An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 Technical Report, 2025.

[11] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. **Commun. ACM**, Vol. 57, No. 10, pp. 78–85, September 2014.

[12] Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. On the use of ArXiv as a dataset, 2019.

[13] Shaurya Rohatgi. ACL Anthology Corpus with Full Text. Github, 2022.

[14] Gemma Team, Aishwarya Kamath, et al. Gemma 3 Technical Report, 2025.

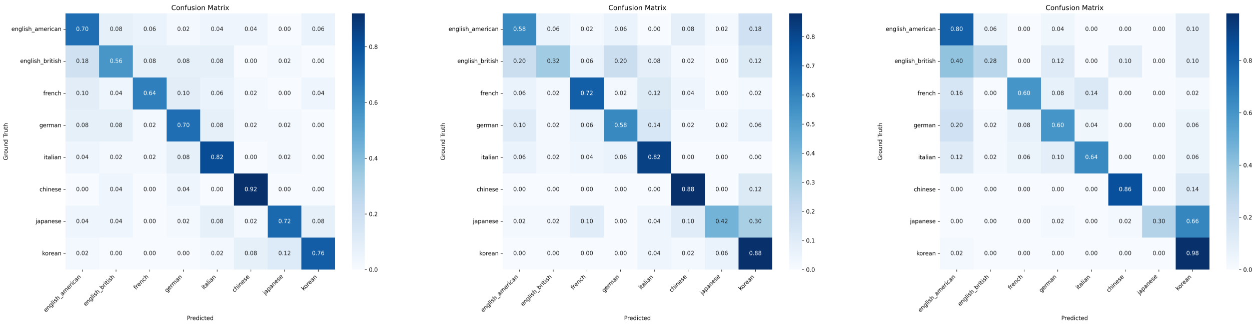
[15] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. Vol. 36, pp. 10088–10115, 2023.

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models., 2022.

A Fine-Tuning Full Results

Table 4 Detailed per-class performance metrics for Qwen3-14B (fine-tuned).

Language	Precision			Recall			F1-score		
	pre-NN	pre-LLM	post-LLM	pre-NN	pre-LLM	post-LLM	pre-NN	pre-LLM	post-LLM
English (US)	0.603	0.569	0.471	0.700	0.500	0.800	**0.648**	0.574	0.593
English (UK)	0.651	0.696	0.737	0.560	0.320	0.280	**0.602**	0.438	0.406
French	0.781	0.720	0.812	0.640	0.720	0.600	0.703	**0.720**	0.690
German	0.673	0.630	0.625	0.700	0.580	0.600	**0.686**	0.604	0.612
Italian	0.695	0.661	0.781	0.820	0.820	0.640	**0.752**	0.732	0.703
Chinese	0.836	0.759	0.878	0.920	0.880	0.860	**0.876**	0.815	0.869
Japanese	0.800	0.808	1.000	0.720	0.420	0.300	**0.758**	0.553	0.462
Korean	0.809	0.524	0.462	0.760	0.880	0.980	**0.784**	0.657	0.628



(a) pre-NN (Qwen3 fine-tuned)

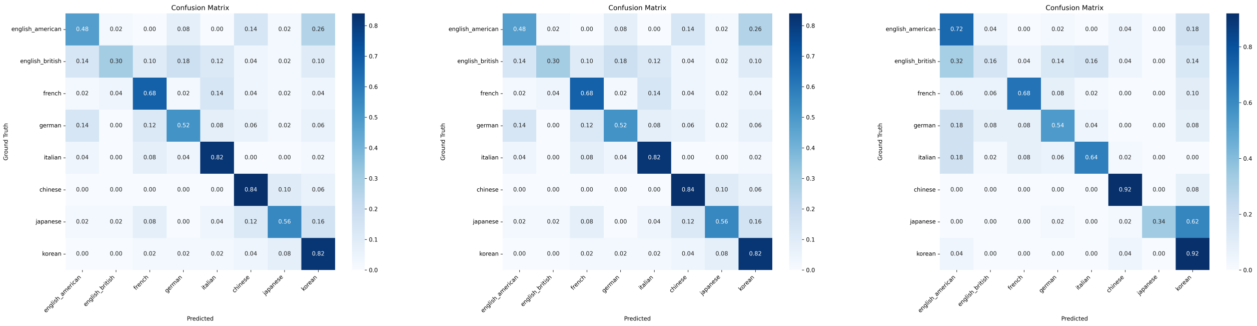
(b) pre-LLM (Qwen3 fine-tuned)

(c) post-LLM (Qwen3 fine-tuned)

Figure 1 Confusion matrices for Qwen3-8B (fine-tuned).

Table 5 Detailed per-class performance metrics (precision, recall, F1-score) across all three eras for Gemma-3-14B-it (fine-tuned).

Language	Precision			Recall			F1-score		
	pre-NN	pre-LLM	post-LLM	pre-NN	pre-LLM	post-LLM	pre-NN	pre-LLM	post-LLM
English (US)	0.643	0.571	0.480	0.720	0.480	0.720	**0.679**	0.522	0.576
English (UK)	0.619	0.790	0.444	0.520	0.300	0.160	**0.565**	0.435	0.235
French	0.674	0.630	0.773	0.660	0.680	0.680	0.667	0.654	**0.723**
German	0.708	0.605	0.628	0.680	0.520	0.540	**0.694**	0.559	0.581
Italian	0.796	0.672	0.744	0.780	0.820	0.640	**0.788**	0.739	0.688
Chinese	0.738	0.656	0.852	0.900	0.840	0.920	0.812	0.737	**0.885**
Japanese	0.736	0.683	1.000	0.780	0.560	0.340	**0.757**	0.615	0.508
Korean	0.833	0.540	0.434	0.700	0.820	0.920	**0.761**	0.651	0.590



(a) pre-NN (Gemma 3 fine-tuned)

(b) pre-LLM (Gemma 3 fine-tuned)

(c) post-LLM (Gemma 3 fine-tuned)

Figure 2 Confusion matrices for Gemma-3-12B-it (fine-tuned). Similar to Qwen3, the model struggles to differentiate L1 styles without fine-tuning.