Investigating Linguistic Abilities of LLMs for Native Language Identification

Ahmet Yavuz Uluslu
University of Zurich
ahmetyavuz.uluslu@uzh.ch

Gerold Schneider
University of Zurich
gschneid@cl.uzh.ch

Abstract

Large language models (LLMs) have achieved state-of-the-art results in native language identification (NLI). However, these models often depend on superficial features, such as cultural references and self-disclosed information in the document, rather than capturing the underlying linguistic structures. In this work, we evaluate the linguistic abilities of open-source LLMs by evaluating their performance in NLI through content-independent features, such as POS n-grams, function words, and punctuation marks, and compare their performance against traditional machine learning approaches. Our experiments reveal that while LLM's initial performance on structural features (55.2% accuracy) falls significantly below their performance on full text (96.5%), fine-tuning significantly improves their capabilities, enabling state-of-the-art results with strong cross-domain generalization.

1 Introduction

Native Language Identification (NLI) aims to automatically determine an individual's native language (L1) based on their writing or speech in a second language (L2). This task is grounded in cross-linguistic influence theory, which posits that L1 leaves distinctive traces in the L2 production patterns (Yu and Odlin, 2016). NLI applications include providing metalinguistic feedback to language learners (Karim and Nassaji, 2020) and adapting grammatical error correction (GEC) systems based on L1 and proficiency level (Nadejde and Tetreault, 2020).

The top performing systems in the two previous shared tasks in NLI combined linguistic features with machine learning algorithms (Malmasi et al., 2017). Various feature types were investigated, including spelling errors, word and lemma n-grams, character n-grams, dependency trees, and morphosyntax (Markov et al., 2022). Recent advances in large language models (LLMs), particularly GPT-4 and LLaMA-3, demonstrate emergent metalinguistic abilities, including the capacity to process and analyze complex linguistic structures such as constituency trees of ambiguous sentences (Beguš et al., 2023). These newly acquired capabilities enabled the models to excel in downstream tasks such as NLI and GEC, which traditionally require thousands of examples to learn relatively complex linguistic relationships. Remarkably, LLMs achieve state-of-the-art performance in NLI on various benchmarks without any task-specific training (Zhang and Salle, 2023; Ng and Markov, 2024).

Despite their impressive performance, previous work has revealed that LLMs can rely on task-related shortcuts using superficial features, such as country names and cultural references, in the document rather than focusing on relevant linguistic features (Uluslu et al., 2024). Moreover, they often generate unfaithful explanations by failing to disclose their dependence on content-based hints in their reasoning process (Turpin et al., 2024). The close relationship between content and structural features makes it difficult to determine whether the models' success reflects their ability to perform genuine linguistic analysis or simply stems from pattern matching based on content cues.

The contributions of this work are the following: (i) we assess the linguistic abilities of open source LLMs through content-independent features, such as part-of-speech (POS) tags, function words, and punctuation marks, and compare their performance against traditional machine learning approaches, (ii) we demonstrate that while LLMs initially exhibit significant performance degradation when content words are replaced, indicating a strong dependence on lexical cues, fine-tuning

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Ahmet Yavuz Uluslu and Gerold Schneider. Investigating Linguistic Abilities of LLMs for Native Language Identification. *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2025)*. University of Tartu Library. 81–88.

enables them to effectively leverage structural features, achieving state-of-the-art performance with robust cross-domain generalization.

2 Related Work

Linguistic probing Probing studies provide a systematic framework to analyze the linguistic knowledge encoded in LLM representations, demonstrating their ability to capture syntactic characteristics such as POS tags and dependency structures (Waldis et al., 2024). The approach employs a linear classifier trained on top of the model's contextual representations to predict specific linguistic properties. Following the key assumption of the probing paradigm, high classifier performance signals that model representations effectively encode the targeted linguistic phenomena (Belinkov, 2022).

Metalinguistic ability A crucial aspect of understanding language is metalinguistic ability, which refers to the ability to explicitly analyze and reason about linguistic structures and properties within a formal framework (Benelli et al., 2006). Although probing studies aim to reveal what linguistic information is encoded in model representations, they do not demonstrate the model's ability to perform explicit linguistic analysis. Moving beyond probing, Beguš et al. (2023) demonstrates that GPT-4 can generate formal syntactic tree analyses for ambiguous sentences, offering a more direct assessment of metalinguistic ability. Although the entanglement between linguistic performance and competence remains an open research question, recent discussions in the literature increasingly frame this analytical process as potential evidence of metalinguistic ability (Millière, 2024).

The impact on native language identification The emergence of metalinguistic performance in LLMs has significant implications for authorship analysis. Contemporary LLMs can explicitly analyze cross-linguistic influence, enabling them to identify word order influences and grammatical patterns that reveal an author's native language (Zhang and Salle, 2023). Earlier approaches using smaller models such as GPT-2 were based on learning probability distributions of various learner innovations specific to each L1 through the fine-tuning process (Uluslu and Schneider, 2022). The attempts to use a single GPT-2 model for the direct classification of L1 yielded suboptimal results. In these approaches, the surprisal of the model served as a proxy for L1 rather than an explicit analysis of the relevant linguistic features. Formally, for a text sequence $X = (x_1, ..., x_n)$, the surprisal S is defined as:

$$S(X) = -\sum_{i=1}^{n} \log P(x_i|x_{<i}) \quad (1)$$

To identify the native language L_1, separate language models M_{L_1} are trained for each potential linguistic background in $\mathcal{L}$. For a given text X, we compute surprisal scores using each language-specific model, and the final prediction is determined by identifying the model that yields the minimum surprisal:

$$L_1^* = \arg\min_{L_1 \in \mathcal{L}} S_{M_{L_1}}(X) \quad (2)$$

Although this approach showed strong results in the benchmarks, it suffered from poor cross-domain generalization (Vian, 2023), which we attribute to its strong lexical dependency, an inherent limitation of not being able to target specific linguistic features in the text. The most recent LLMs such as GPT-4 represent a fundamental shift in this regard, as they can explicitly identify and analyze task-relevant linguistic features out-of-the-box. This capability has been highlighted in previous studies, in which LLMs demonstrate state-of-the-art performance in zero-shot settings, eliminating the need for large training sets while simultaneously providing natural language explanations for their predictions (Zhang and Salle, 2023; Ng and Markov, 2024).

However, research indicates that such explanations, while superficially convincing, do not accurately represent the actual reasoning processes of the model (Turpin et al., 2024). Instead, the models frequently generate L1 predictions first and then construct plausible explanations, creating the illusion of metalinguistic analysis. This disconnect is problematic, as it allows findings to be selectively framed to support or refute any given authorship hypothesis (Ishihara et al., 2024). Evaluating linguistic abilities poses significant challenges in tasks like NLI, where distinguishing between cause and effect creates a chicken-and-egg problem: Does proficiency in grammatical error detection enable L1 identification, or does L1 identification reveal grammatical patterns? Although prompts can instruct models to attend to

specific features, this constrained behavior still results in unfaithful explanations that often underestimate the influence of content-dependent features (Agarwal et al., 2024).

Content-independent features, widely adopted in authorship analysis research (Nini et al., 2024; Markov et al., 2022), offer a more robust approach by retaining structural patterns while minimizing the influence of topical and contextual cues. Markov et al. (2022) employs POS n-grams, function words, and punctuation marks with SVM, while Nini et al. (2024) constructs author-specific grammatical representations using n-gram models and compares disputed texts against these models using log-likelihood ratios. Both studies demonstrate that content-independent features can effectively capture structural patterns that generalize across domains. We investigate whether LLMs' claimed linguistic abilities reflect genuine analytical processes by examining their exploitation of structural features such as POS tags, function words, and punctuation patterns.

3 Dataset

The **TOEFL11** dataset (Blanchard, 2013) contains 1,100 essays in English, written by native speakers (L1) of 11 different languages. In total, there are 12,100 essays with an average of 348 tokens per essay. The essays were written in response to eight different writing topics, all of which appear in the 11 L1 groups, by authors with low, medium or high English proficiency. For our experiments, we focus on TOEFL4, a four-language subset of TOEFL11 (n=4400) that includes only essays written by native French, German, Italian, and Spanish speakers (Markov et al., 2022).

The **ICLE4-NLI** dataset, drawn from the ICLEv2 corpus (Granger et al., 2009), serves as our cross-domain evaluation benchmark. It contains 400 essays written by medium to high proficiency English learners, evenly distributed across four first languages of TOEFL4: French, German, Italian, and Spanish.

4 Methodology

To investigate the syntactic capabilities of LLMs, we adopted a methodology inspired by content-independent features of authorship analysis (Markov et al., 2022; Nini et al., 2024). Among various possible masking configurations shown in Table 1, we use *Ex. FW-Punct* approach where each content word in the dataset is replaced (masked) with its corresponding POS tag, while retaining function words and punctuation marks to preserve structural patterns. The function words and phrases were identified using the POSNoise word list, which aims to mask topic-related words while preserving as much structural information as possible (Halvani and Graner, 2021). This approach allows certain delexicalised verbs, such as "make", to remain in their original form, compared to stop-word lists available in various open-source packages (Nothman et al., 2018).

Step	Sentence
Original	make products seem much better!
All POS	VB NNS VB RB JJ PUNCT
Ex. FW	make N seem much J PUNCT
Ex. FW-Punct	make N seem much J !

Table 1: The original sentence, its transformation into POS tags, POS tags except for function words, and the final form where both function words and punctuations are preserved.

4.1 Machine Learning with Linguistic Features

In our experiments, we use the liblinear implementation of support vector machines (SVM) from the scikit-learn library, using a one-vs-rest (OvR) strategy for multiclass classification (Pedregosa et al., 2011). To optimize hyperparameters, we perform a search over a range of regularization values $C \in 0.1, 1, 2.5, 5, 10$ and determine that the optimal value lies within the range of 2 to 2.5. Previous research has established the optimal POS n-gram range for this task as 1-3 (Malmasi and Dras, 2018). The experiments were evaluated using ten-fold cross-validation. This traditional machine learning baseline enables direct comparison of how effectively LLMs can exploit the same linguistic features.

4.2 LLM Analysis

For our experiments with LLMs, we employed LLaMA-3.1-8B-Instruct[1] and LLaMA-3.1-70B-Instruct to evaluate their zero-shot performance. The 70B model is particularly notable for being the most comparable to GPT-4 in NLI tasks under zero-shot settings (Uluslu et al., 2024). It was also

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

able to complete the previously mentioned metalinguistic tasks from Beguš et al. (2023). We further fine-tuned the 8B model using 4-bit QLoRA (Dettmers et al., 2024) with the following hyperparameters: a learning rate of 7e-4, a batch size of 16, three epochs, and the AdamW optimizer. Due to computational constraints, we evaluated the 70B model using a few-shot in-context learning with 4, 8, and 16 examples. These examples were randomly sampled from the test set and presented to the model in random order to prevent potential sequence-related biases. We assess both models' performance when asked to assume two distinct roles: a language teacher and a forensic linguist. LLM experiments employed three-fold cross-validation due to computational constraints. Based on previous studies on the TOEFL dataset, lower-fold cross-validation typically yields more variable accuracy scores, though differences are generally small (1-3%) (Malmasi and Dras, 2018). Details of the fine-tuning setup and the system prompt are provided in Appendix A.1.

5 Results and Discussion

We present our key experimental findings through two main analyses. First, Table 2 shows the performance of LLMs on TOEFL4 in different configurations. Second, we assess cross-domain generalization by evaluating these TOEFL4-trained models on ICLE4-NLI, with results presented in Table 3.

System + Features	Acc (%)
POS 1-3 grams $_{ML}$	76.2
LLaMA-3.1-8B$_{full}$	64.3
LLaMA-3.1-70B$_{full}$	96.5
LLaMA-3.1-8B$_{POS-ZS}$	26.0
LLaMA-3.1-70B$_{POS-ZS}$	55.2
LLaMA-3.1-70B$_{POS-FS@4}$	58.4
LLaMA-3.1-70B$_{POS-FS@8}$	63.2
LLaMA-3.1-70B$_{POS-FS@16}$	63.5
LLaMA-3.1-8B$_{POS-FT}$	**89.2**

Table 2: The accuracy score for systems, comparing syntactic features (POS n-grams, DT-grams), combined features (POS + DT-grams) and LLM analysis of (LLaMA 3.1 8B and 70B). ZS: Zero-shot, FS: Few-shot, FT: Fine-tuned, ML: Machine Learning (SVM)

5.1 Performance of LLMs

We demonstrate that the impressive zero-shot performance of LLMs (96.5%) on the TOEFL4 benchmark (LLaMA-3.1-70B$_{full}$) drops to 55.2% when the content words are masked with their POS tags (LLaMA-3.1-70B$_{POS-ZS}$) excluding function words and punctuation marks. While few-shot prompting with L1-specific examples improves performance, we observe diminishing returns beyond 16 examples. This setup also significantly increases the computational overhead, as transformer models' memory and computation requirements scale quadratically with input sequence. Furthermore, zero-shot performance ($POS-ZS$) falls short of the traditional machine learning baseline trained on POS n-grams (ML), suggesting that the model struggles to fully capture the linguistic patterns present in the text. However, after fine-tuning, the 8B parameter model achieves an accuracy of 89.2%, a performance that proves robust even in cross-domain evaluation. This represents a considerable advance over previous approaches, which typically showed substantial performance degradation when tested with out-of-domain data, as shown in Table 3 (TOEFL4→ICLE). Our analysis revealed that the prompt for the role of language teacher achieved higher performance in zero-shot settings (53.1%) compared to the role of forensic linguist (55.2%), with this difference being statistically significant ($p < 0.03$).

System	ICLE	TOEFL4→ICLE
POS 1-3 grams $_{ML}$	-	62.1
LLaMA-3.1-8B$_{ZS}$	30.1	-
LLaMA-3.1-70B$_{ZS}$	64.3	-
LLaMA-3.1-8B$_{FT}$	95.6	**90.3**

Table 3: The accuracy scores (%) on ICLE4 dataset, comparing in-domain performance with cross-domain transfer from TOEFL4. ZS: Zero-shot, FT: Fine-tuned, ML: Machine Learning (SVM)

For zero shot settings, the primary challenge was differentiating between the pair of French and Italian languages, as detailed in the confusion matrix presented in Figure 1. The model exhibits a systematic bias toward Italian predictions under uncertainty, resulting in a notably low prediction accuracy for both French and Spanish L1 texts. Although few-shot prompting partially mitigates this limitation by improving French L1 identification, the confusion between Romance languages persists. This pattern can be attributed to two key factors: first, language learners from these

Romance language backgrounds exhibit similar error patterns in their English writing; second, our content-masking approach prevents the model from leveraging distinctive lexical cues such as false friends. In contrast, German L1 texts were consistently the most accurately identified in all approaches. This superior performance can be potentially explained by several linguistic factors: The transfer effects of German L1 learners are more structurally distinct. Importantly, our prompt design, which explicitly mentions "German" as a classification label, may guide the model to search for these distinctive features even under content-masked conditions. For instance, German L1 writers show characteristic patterns in their use of function words (e.g., unique placement of "that" in subordinate clauses) and delexicalized verbs (e.g., distinct usage patterns of "make" and "do" influenced by German "machen"). These systematic differences, particularly visible in the syntactic structures preserved by our replacement approach, possibly make German L1 texts more readily distinguishable from Romance language backgrounds. To verify this hypothesis about the role of function words in L1 identification, future work could extend the masking approach to replace function words with their corresponding POS tags.

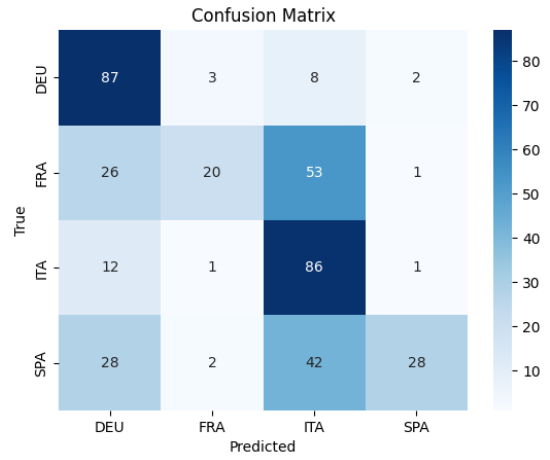

Figure 1: Confusion matrix of the model LLaMA3.1-70B under zero-shot settings. The results highlight the distribution of predictions across different L1.

5.2 Model Interpretation

One of the main criticisms of LLM is their lack of explainability, as their output often does not transparently reflect their underlying reasoning processes (Turpin et al., 2024). For exploratory analysis, we provide an example NLI analysis completed by LLM in Appendix A.3. The linguistic features most frequently mentioned in these analyses are summarized in Table 4. Assessing the significance of these features involves studying the self-consistency of the model by removing or perturbing specific features and evaluating their impact on classification (Parcalabescu and Frank, 2024), which we leave for future work.

Category	Occurrences (%)
Word Order and Sentence Structure	491 (30.66)
Prepositions	359 (22.42)
Grammatical Errors	260 (16.24)
Syntactic Patterns	195 (12.18)
Idiomatic Expressions and Phrasing	177 (11.06)
Articles and Determiners	132 (8.24)
Error Patterns and Miscellaneous	107 (6.68)
Pronouns	104 (6.50)
Function Words	80 (5.00)
Vocabulary and Lexical Choices	29 (1.81)

Table 4: Frequency Distribution of Linguistic Features in Generated Analysis (Llama-3.1-70B)

Traditional SLA research has relied heavily on inherently interpretable models, such as the linear SVM baseline used in our study. These models allow researchers to directly examine the coefficients to identify the most significant features for classification, providing clear insights into cross-linguistic influences from different backgrounds of L1 (Berti et al., 2023). In particular, many features identified in LLM outputs align closely with the findings of these traditional models and are well documented in the SLA literature. This suggests that model behavior may resemble a form of approximate retrieval, where the models reference documents containing these linguistic structures to derive their classifications.

Our analysis of the best-performing zero-shot model's results for German L1 writers (the most accurately identified background) illustrates this alignment through three distinctive features. German writers employ the expletive construction ("there is") more frequently than writers from other L1 backgrounds in the TOEFL-4 dataset. They demonstrate a clear preference for complex sentences, characterized by frequent use of relative clauses ("N which") and generally more intricate syntactic structures. They show a distinctive pattern in their use of impersonal expressions, particularly through the use of "one," often appearing

in fixed expressions like "man kann sagen" (translated as "one can say," commonly used to mean "in conclusion" or "I think"). These patterns not only align with the findings of traditional SLA research, but also emerge consistently in LLM outputs, suggesting an ability to identify and leverage meaningful linguistic features.

6 Conclusion

Our study makes significant contributions to understanding LLMs' linguistic abilities in native language identification. Although LLMs have shown impressive performance in NLI tasks on various benchmarks, our investigation reveals a more nuanced picture when evaluating their ability to analyze structural linguistic features in isolation. The dramatic performance drop when content words are masked (from 96.5% to 55.2% for LLaMA-3.1-70B) suggests that these models heavily rely on lexical and content-based cues in their initial predictions. However, through further fine-tuning in these controlled settings, models can achieve an accuracy of 89.2% with strong generalization across domains. These findings reveal that LLMs can acquire structural analysis capabilities through fine-tuning. Our results contribute to the growing body of evidence that LLMs can exhibit metalinguistic abilities, as demonstrated not only through their performance on formal linguistic tasks but also through their capabilities in downstream applications such as NLI.

Limitations

Experimental Design: Our evaluation is focused solely on NLI as a proxy of metalinguistic competence, which may not capture the full spectrum of linguistic abilities and understanding. The controlled setup using POS tags and function words cannot fully represent the complex interactions between syntax, morphology, semantics, and pragmatics in natural language.

Dataset Coverage: The study's reliance on TOEFL11 and ICLE4-NLI datasets with only four L1 backgrounds (French, German, Italian, Spanish) limits the generalizability of our findings across different languages.

Practical Applications: While our findings demonstrate promising results in controlled settings, their applicability to real-world forensic linguistics or educational applications requires further investigation.

Ethics Statement

Our project only processes information from publicly available learner corpora. No sensitive personal data was accessed, stored, or processed at any stage of the project. The study was granted an ethics board review exemption under the University of Zurich guidelines.

Acknowledgments

This work was supported by the collaboration between the University of Zurich and PRODAFT as part of the Innosuisse innovation project AUCH 103.188 IP-ICT (Author profiling to automatize attribution in cybercrime investigations). The authors thank Prof. Rico Sennrich from the University of Zurich for his exchange of ideas.

References

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*.

Gašper Beguš, Maksymilian Dąbkowski, and Ryan Rhodes. 2023. Large linguistic models: Analyzing theoretical linguistic abilities of llms. *arXiv preprint arXiv:2305.00948*.

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

Beatrice Benelli, Carmen Belacchi, Gianluca Gini, and Daniela Lucangeli. 2006. 'to define means to say what you know about things': the development of definitional skills as metalinguistic acquisition. *Journal of Child Language*, 33(1):71–97.

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

D Blanchard. 2013. TOEFL11: A Corpus of Non-native English. *Educational Testing Service*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English. Version 2. Handbook*.

Oren Halvani and Lukas Graner. 2021. Posnoise: An effective countermeasure against topic biases in authorship analysis. In *Proceedings of the 16th International Conference on Availability, Reliability and Security*, pages 1–12.

Shunichi Ishihara, Sonia Kulkarni, Michael Carne, Sabine Ehrhardt, and Andrea Nini. 2024. Validation in forensic text comparison: Issues and opportunities. *Languages*, 9(2):47.

Khaled Karim and Hossein Nassaji. 2020. The revision and transfer effects of direct and indirect comprehensive corrective feedback on esl students' writing. *Language Teaching Research*, 24(4):519–539.

Shervin Malmasi and Mark Dras. 2018. Native language identification with classifier stacking and ensembles. *Computational Linguistics*, 44(3):403–446.

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.

Raphaël Millière. 2024. Language models as models of language. *arXiv preprint arXiv:2408.07144*.

Maria Nadejde and Joel Tetreault. 2020. Personalizing grammatical error correction: Adaptation to proficiency level and l1. *arXiv preprint arXiv:2006.02964*.

Yee Man Ng and Ilia Markov. 2024. Leveraging open-source large language models for native language identification. *arXiv preprint arXiv:2409.09659*.

Andrea Nini, Oren Halvani, Lukas Graner, Valerio Gherardi, and Shunichi Ishihara. 2024. Authorship verification based on the likelihood ratio of grammar models. *arXiv preprint arXiv:2403.08462*.

Joel Nothman, Hanmin Qin, and Roman Yurchak. 2018. Stop word lists in free open-source software packages. In *Proceedings of workshop for NLP open source software (NLP-OSS)*, pages 7–12.

Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Ahmet Yavuz Uluslu and Gerold Schneider. 2022. Scaling native language identification with transformer adapters. *arXiv preprint arXiv:2211.10117*.

Ahmet Yavuz Uluslu, Gerold Schneider, and Can Yildizli. 2024. Native language identification improves authorship attribution. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 297–303.

Matias Johansen Vian. 2023. A study of transformers for cross-corpus native language identification. Master's thesis, NTNU.

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes: Benchmark the linguistic competence of language models. *arXiv preprint arXiv:2404.18923*.

Liming Yu and Terence Odlin. 2016. *New perspectives on transfer in second language learning*, volume 92. Multilingual Matters.

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

A Appendix – Supplementary Material

A.1 LLM Prompts and Outputs

You are a forensic linguist/**language teacher** that reads English texts written by non-native speakers in order to identify their native language. Use clues such as word choice patterns, syntactic structures, and grammatical errors to determine the author's native language. You will analyze texts where all content words have been replaced with specific POS tags:
- NOUN -> N
- PROPN (Proper Noun) -> P
- VERB -> V
- ADJ (Adjective) -> J
- ADV (Adverb) -> B
- NUM (Number) -> D
- SYM (Symbol) -> S

All function words remain unchanged. Analyze this new text and identify the native language of the author from among four possibilities: Italian (ITA), Spanish (SPA), French (FRA), or German (DEU).

Figure A.1: System Prompts 1-2

```
### Instruction:
<system prompt>
### Input:
<TOEFL11 ESSAY TEXT>
### Response:
<L1 LABEL>
```

Figure A.2: Fine-tuning Prompt

Analyzing the provided text, I've identified several clues that suggest the author's native language is German (DEU). Here are the key indicators:

1. **Word order:** German often uses a subject-object-verb word order, especially in subordinate clauses. In the text, we see phrases like "i think it is an J N and in the following N you will V my N about this" and "when i think about all this it would be the J to make now V what we should do with the N". This word order is more typical of German than the other options.

2. **Use of modal verbs:** German often uses modal verbs like "können" (can) and "müssen" (must) in a way that's different from the other languages. In the text, we see phrases like "N can also go by N" and "it would be a J N for the N and N J when less N V the N with N and N". This use of modal verbs is more typical of German.

3. **Clause structure:** German often uses complex clause structures with multiple subordinate clauses. In the text, we see phrases like "my J N is that it would be a J N for the N and N J when less N V the N with N and N" and "when i think about all this it would be the J to make now V what we should do with the N". This complex clause structure is more typical of German.

4. **Lack of articles:** German often omits articles in certain contexts, especially in informal writing. In the text, we see phrases like "i think it is an J N" and "N can also go by N". This lack of articles is more typical of German.

5. **Use of prepositions:** German often uses prepositions like "mit" (with) and "von" (from) in a way that's different from the other languages. In the text, we see phrases like "N can also go by N, N and N or by N or V B with other N in one N" and "B it could be a N in D N for us to V N with N from N or N". This use of prepositions is more typical of German.

Based on these clues, I conclude that the author's native language is German (DEU).

Figure A.3: An example linguistic analysis generated by LLaMA-3.1-70B.