Cross-Genre Native Language Identification with Open-Source Large Language Models

Robin Nicholls [a,*] **and Kenneth Alperin** [b,**]

[a]University of Edinburgh
[b]MIT Lincoln Laboratory

Abstract. Native Language Identification (NLI) is a crucial area within computational linguistics, aimed at determining an author's first language (L1) based on their proficiency in a second language (L2). Recent studies have shown remarkable improvements in NLI accuracy due to advancements in large language models (LLMs). This paper investigates the performance of open-source LLMs on short-form comments from the Reddit-L2 corpus compared to their performance on the TOEFL11 corpus of non-native English essays. Our experiments revealed that fine-tuning on TOEFL11 significantly improved accuracy on Reddit-L2, demonstrating the transferability of linguistic features across different text genres. Conversely, models fine-tuned on Reddit-L2 also generalised well to TOEFL11, achieving over 90% accuracy and F1 scores for the native languages that appear in both corpora. This shows the strong transfer performance from long-form to short-form text and vice versa. Additionally, we explored the task of classifying authors as native or non-native English speakers, where fine-tuned models achieve near-perfect accuracy on the Reddit-L2 dataset. Our findings emphasize the impact of document length on model performance, with optimal results observed up to approximately 1200 tokens. This study highlights the effectiveness of open-source LLMs in NLI tasks across diverse linguistic contexts, suggesting their potential for broader applications in real-world scenarios.

1 Introduction

Native Language Identification (NLI) represents a critical area of study within computational linguistics, focusing on the determination of an author's first language (L1) through their written proficiency in a second language (L2). The relevance of NLI extends across various domains, notably in forensic linguistics for authorship profiling and in educational linguistics, where it aids in the customisation of teaching materials tailored to the linguistic background of L2 learners. The significance of NLI as a computational challenge was markedly enhanced following the release of the TOEFL11 corpus [2], a comprehensive dataset of non-native English writing, which has since served as a benchmark for advancing research in this domain.

Historically, NLI research has predominantly relied on traditional supervised learning methodologies [10]. However, recent advancements have undergone a paradigm shift towards employing large language models (LLMs), particularly leveraging zero-shot [14] learning and fine-tuning strategies [8, 11], with an emphasis on long-form essay-based datasets. While preliminary findings suggest the promise of LLM-based approaches in enhancing NLI performance, further empirical exploration across diverse real-world datasets remains imperative for elucidating the practical applicability of these methods.

In this paper we explore the effectiveness of open-source LLMs on short-form comments from Reddit. In Section 2, we discuss the related works for native language identification with large language models. In Section 3, we discuss the datasets, models, and prompting techniques we used for NLI, and evaluate performance on the long-form TOEFL dataset. More specifically, we fine-tune 3-billion and 8-billion parameter Llama-3 models [7] on the TOEFL11[2] training set, following a similar regime to Ng and Markov [11]. In Section 4, we seek to measure the transfer performance of these models from long-form to short-form text. More specifically, we explore the performance of those models on a subset of the Reddit-L2 corpus [12], acting as a validation set. Next, we perform the inverse of this by fine-tuning the same foundation models on Reddit-L2 and validating on TOEFL11. This approach seeks to answer whether the models are learning general linguistic characteristics, or simply over-fitting to the training set. Additionally, we introduce the sub-task of classifying an author of the Reddit-L2 dataset as being a native or non-native English author. We conclude in Section 5 and discuss limitations and next steps in Section 6.

2 Related Work

The progression of NLI research has been notably documented through workshops such as NLI-2013 [9] and NLI-2017 [10], which predominantly utilised the TOEFL11 corpus. These collaborative efforts underscored the effectiveness of ensemble methods, incorporating various traditional machine learning classifiers. These classifiers, trained on a diverse array of features including lexical, stylistic, and syntactic elements, demonstrate superior performance. Among the participants, the ItaliaNLP Lab [4] achieved remarkable accuracy, reaching a rate of 88.18% on the TOEFL11 test set, establishing a benchmark for subsequent research endeavours.

The first survey paper on NLI came out in 2024 [6]. Recent studies have ventured into the exploration of generative LLMs within the context of NLI, showcasing substantial advancements. An accuracy of 89.0% accuracy on TOEFL11 was achieved by fine-tuning GPT-2 [8]. Next, the application of GPT-3.5 and GPT-4 for zero-shot learning experiments on the TOEFL11 corpus set a new precedent in accuracy, achieving 91.7% with GPT-4 [14]. This exploration into the

* Corresponding Author. Email: r.nicholls@sms.ed.ac.uk.
** Corresponding Author. Email: kenneth.alperin@ll.mit.edu.

capabilities of LLMs revealed the potential for significant improvements in NLI accuracy. To our knowledge, LLMs have not been used yet on the Reddit-L2 dataset for NLI classification.

Further extending the boundaries of current methodologies, Ng and Markov [11] embarked on the approach of fine-tuning various open-source LLMs utilising 4-bit Quantization-aware training for Low Rank Adaptation (QLoRA) [5]. Their findings suggest a narrowing accuracy gap between fine-tuned LLMs and the zero-shot capabilities of GPT-4, with 8 billion parameter models nearly matching the performance of GPT-4 on the TOEFL11 dataset and surpassing it on the ICLE-NLI dataset. Such advancements underscore the rapidly evolving landscape of NLI research and its increasing reliance on the sophisticated capabilities of large language models.

3 Data and Models

The foundation of our study rests on two primary datasets: the ETS Corpus of Non-Native Written English (TOEFL11) and the Reddit-L2 corpus. To ensure a comprehensive evaluation, we utilise two LLMs of varying sizes. These were selected based on their strong performance on the TOEFL11 test set. This approach allows for an exploration of how model size impacts performance on non-native English text, providing insights into the scalability and efficiency of LLMs in handling linguistic diversity.

3.1 Data

ETS Corpus of non-Native Written English (TOEFL11) [2]: comprises 12,100 essays written by individuals across 11 L1 backgrounds (Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish) and provides a rich resource for analysing written proficiency in academic English. This corpus is divided into training, validation, and testing sets, containing 9,900, 1,100, and 1,100 essays, respectively balanced between the languages. This ensures an even distribution of L1 groups, which facilitates a balanced analysis of linguistic features across different language backgrounds.

Reddit-L2 Corpus [12]: represents a more informal register of English, encompassing approximately 250 million sentences written by over 45,000 authors. These authors were identified through the use of flairs, a metadata attribute in subreddits, which approximated the authors' L1 based on the national language of their indicated country. Although this method may introduce some inaccuracies in L1 attribution (such as with countries that have multiple national languages), due to misleading flairs or instances where an author's true L1 does not align with their country's primary language, the sheer volume of data is expected to mitigate the impact of such anomalies.

WI-LOCNESS [3]: was initially developed to support research in Grammatical Error Correction (GEC), as it comprises a total of 350 essays authored by both native English speakers and English language learners. Given that a portion of the essays originates from non-native English learners, the dataset is also suitable for tasks involving native versus non-native classification. We use approximately 150 tokens per essay to ensure consistency in document length, and a subset of 100 essays to maintain a balanced distribution between native and non-native authors.

To analyse these corpora, this study adopts a methodology [12] focusing on the masking of nouns through named entity recognition (NER) using the spaCy English core web text Transformer model.

Table 1: Comparison of TOEFL11 and Reddit-L2

Measure	**TOEFL11**	**Reddit-L2**
Average Sentence Length	24.206±15.436	15.858±3.986
Proportion Unique Tokens	0.498±0.086	0.436±0.040
First Order Coherence	0.467±0.077	0.328±0.037
Second Order Coherence	0.456±0.097	0.314±0.039
Flesch Reading Ease	72.700±18.450	81.303±7.447

This approach aims to obscure specific lexical items, thereby compelling the analytical model to emphasise semantic understanding over mere lexical recognition. We hypothesise this strategy enhances the model's robustness by reducing its reliance on identifiable and proper nouns, which may vary significantly across L1s.

Table 1 shows some linguistic statistics of the test sets used from the TOEFL11 and Reddit-L2 datasets. Reddit has a much lower average sentence length than TOEFL does, which indicates Reddit has simpler sentences with less syntactic complexity. It also has a lower proportion of unique tokens, which indicates Reddit also has a simpler vocabulary and simplifed content. These measures definitely effect the coherence measures for the two datasets, as having shorter/choppier text can lead to more abrupt transitions and less flow, yielding the lower coherence measures for Reddit. Due to the simpler language employed on Reddit, it is easier to read than the TOEFL essays, indicated by the higher Flesch Reading Ease. Overall, these measures support the two datasets vary significantly in their linguistic properties.

Further, to assess the influence of text length on the accuracy of the model, subsets of the Reddit-L2 dataset were curated with fixed document lengths. This aspect of the study acknowledges that text length can be a confounding variable, potentially impacting the model's performance in identifying linguistic features characteristic of non-native English writing. This was done though to remove document length as a variable of comparing TOEFL11 and Reddit-L2, so the focus is on the writing of the text.

In the comparative analysis between the Reddit-L2 corpus and the TOEFL11 dataset, it is noteworthy that the Reddit-L2 corpus includes authors from an extensive array of 50 countries, whereas only a subset consisting of seven countries corresponds to the language backgrounds represented in the TOEFL11 corpus. Among these seven countries, the representation from China was deemed insufficient for meaningful analysis, thus necessitating its exclusion from certain comparative studies within this research. Hindi and Telugu, the two predominant languages in India, are both represented in TOEFL11. However, in Reddit-L2, Indian authors do not affiliate with a specific language. To address this discrepancy for validation tests involving the TOEFL11 dataset, predictions made by the model that assign either Hindi or Telugu to an author of Indian origin are considered accurate.

3.2 Dataset Topics

We conduct a qualitative analysis using Latent Dirichlet Allocation (LDA) topic modelling to evaluate the potential overlap of lexical-based features between the two datasets. We use LDA both on the overall datasets and on the L1 groups within them. Such an analysis is critical to ensure that topic bias does not influence feature selection. Prior to examining the topics identified through LDA, it is important to highlight the fundamental differences in the content of the two datasets. The TOEFL11 dataset consists of responses to eight standardised writing prompts, designed to ensure balanced representation of topics across the dataset. These prompts reflect the diversity of themes typically encountered in the TOEFL writing sec-

tion. In contrast, the Reddit-L2 dataset consists of comments made in response to posts within specific subreddits related to Europe.

Generally, the TOEFL11 dataset includes topics such as education, travel, social life, and transportation, with similar thematic patterns observed across individual L1 groups. This consistency is likely attributable to the standardised nature of the prompts: while authors may draw upon their personal experiences to respond, their answers are constrained by the predefined topic of the questions. On the other hand, the Reddit-L2 dataset predominantly features discussions of public issues popular at the time the comments were posted. For instance, topics such as the Greek economy, Turkish-European relations, and terrorism were frequently observed. Importantly, within Reddit-L2, the topics vary significantly across individual L1 groups, aligning closely with issues that might be expected to be of particular interest to speakers of those languages. For example, German authors frequently discussed political ideologies, refugees, and the German language, whereas Turkish authors often focused on Islam, history, and Europe.

These differences in topic distributions across the two datasets support the hypothesis that classification is not solely driven by lexical-based features. However, the variation in topics within the Reddit-L2 L1 groups raises the concern that models trained on this dataset may become overly reliant on lexical features during fine-tuning. This highlights the importance of applying noun masking to the Reddit-L2 data, as this technique prevents the model from biasing toward topic-specific lexical features.

3.3 Models

As shown in Table 2, we compare the results of Ng and Markov [11] to our own study of similar open-source LLMs: Llama-3.2 (3B), Llama-3.1 (8B), Llama-3.1 (70B) [7], and Mixtral (8x7B) [13]. For all results, we provide the average accuracy score and standard deviation over three runs. All model temperatures and top-p values were set to 0.95 and 0.7. We used Llama-Factory [15] for fine-tuning and validation of all models.

Table 2: Comparative analysis of foundation and fine-tuned open-source LLM performance on TOEFL11 in terms of classification accuracy (%).

Model	TOEFL11 (11 L1s, test set) Closed-set
foundation	
Llama-3.2 (3B)	14.4 ± 0.0
Llama-3 (8B) Ng and Markov [11]	56.8 ± 1.1
Llama-3.1 (8B)	59.2 ± 0.0
Llama-3.1 (70B)	84.0 ± 0.0
Mixtral (8x7B)	67.2 ± 0.8
Gemma (7B) Ng and Markov [11]	13.6 ± 0.0
Phi-3 (3.8B) Ng and Markov [11]	18.2 ± 0.3
fine-tuned	
Llama-3 (8B) Ng and Markov [11]	85.3 ± 0.1
Llama-3.2 (3B)	86.8 ± 0.2
Llama-3.1 (8B)	90.0 ± 0.3
Gemma (7B) Ng and Markov [11]	90.3 ± 1.2

The selection criteria for advancing specific models to the fine-tuning phase are based on both baseline performance scores and practical considerations regarding model deployment. We observe that the 8B and 70B variants of the Llama model exhibit strong baseline performance. However, the decision to proceed with the fine-tuning of the 3B and 8B models, while excluding the 70B variant, is informed by a strategic preference for models that are compatible with typical consumer-grade systems. This choice reflects a pragmatic approach to model selection, aiming to balance the pursuit of high accuracy with the constraints imposed by the computational resources commonly available to end-users. Although our fine-tuned model does not achieve the same performance level as the fine-tuned Gemma 7B model [11], the advantage of using models from the same family is a significant consideration in our decision-making process.

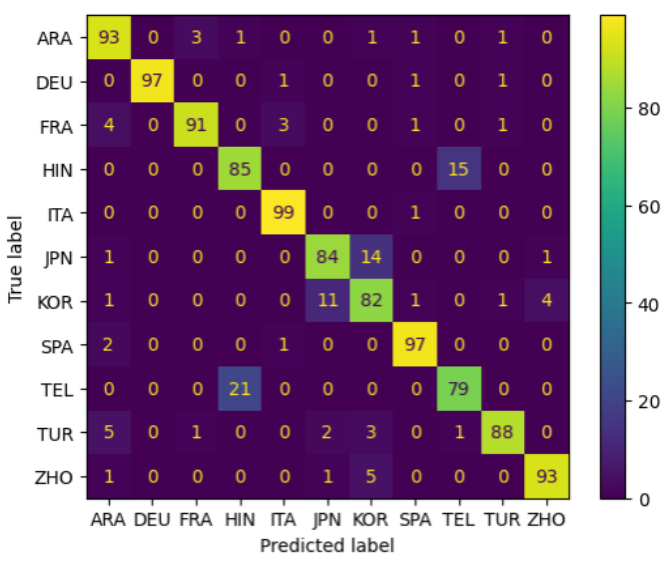

Figure 1: Confusion matrix for L1 accuracy per L1 on TOEFL11

3.4 L1 Analysis

Figure 1 shows the confusion matrix of the classification results for each language in the TOEFL test set on the fine-tuned Llama-3.1 8B model. For most of the languages, the model does a good job of correctly predicting the L1. The model performances lower on Telugu (79 %) than the others. This may be due to high confusion of Telugu with Hindi since they are both languages primarily in India. We also observe some minor overlap of Japanese and Korean, which makes sense as they are both East Asian languages and have a lot of similar syntactic features.

3.5 Prompting Technique

For the closed-set NLID task, we chose to use the prompts provided by Zhang and Salle [14]. For the native vs. non-native English classification, we modify their prompt and provide it in Figure 2. For all experiments, we employ iterative prompting. This allows us to continue to prompt the model, until the LLM returns an answer within the accepted criteria, or a maximum of five attempts have been made.

4 Reddit-L2 for NLI

4.1 Effect of document length

For our initial set of experiments, we utilise the subset of Reddit-L2 data described in the data section, which include German, French, Italian, Turkish, Indian (Hindi/Telugu), and Spanish. The dataset consists of 3,632 training documents and 909 testing documents. As illustrated in Figure 3, we conducted a comprehensive series of

<system>
You are a forensic linguistics expert that reads English texts written by native and non-native authors in order to classify the authors as either:
"NATIVE: native English author
or
"NONNATIVE": Non native English author

Use clues such as spelling errors, word choice, syntactic patterns, and grammatical errors to decide.
DO NOT USE ANY OTHER CLASS.

Valid output formats:
Class: "NATIVE"
Class: "NONNATIVE"

<user>
[document]

<response>
[predicted label]

Figure 2: Prompting template for native vs. non-native English classification.

experiments across varying document lengths to elucidate the significant influence that document length exerts on the model's proficiency in accurate L1 classification. The findings indicate that document length strongly correlates with an enhanced likelihood of feature manifestation. This correlation remains consistent up to an approximate threshold of 1200 tokens, beyond which the benefits of increased document length begin to exhibit diminishing returns.

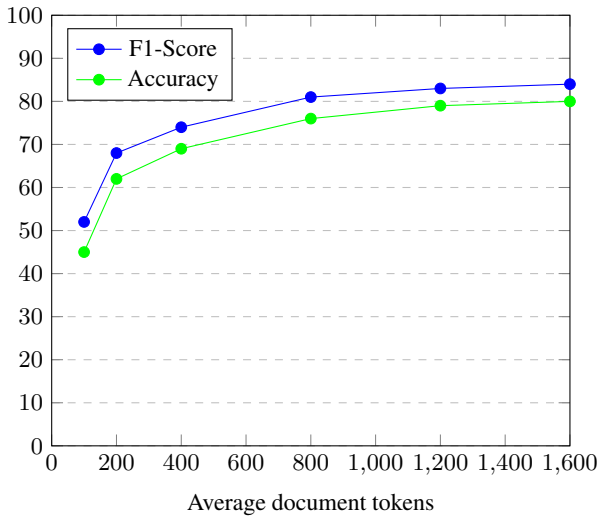

Figure 3: Effect of document length on Reddit-L2

4.2 *LLMs fine-tuned on TOEFL11 and tested on Reddit-L2*

When considering the results for document lengths of approximately 1200 tokens, the fine-tuned models exhibit a commendable ability to generalise to the Reddit authors. This achievement is particularly notable given the differences in genre and the application of noun masking. Specifically, the TOEFL dataset comprises short essays, whereas our construction of the Reddit data involves concatenated short-form comments. These comments, due to their nature, do not typically conform to a single coherent conversation, making them harder to follow as a unified passage. As noted in the data description, aggressive noun masking was applied to the Reddit data to ensure that semantic understanding, rather than mere lexical recognition, was required. This approach is particularly critical given the method used to identify Reddit authors. As many authors were sourced from country-specific subreddits, this frequently led to discussions about their home countries, potentially revealing their presumed native language to the language model.

Table 3: Comparative analysis of foundation and TOEFL11 fine-tuned models on Reddit-L2, 6 L1s, average 1200 tokens.

Model	Accuracy (%)	F1-score (%)
Llama-3.2 (3B)	19.3 ± 0.4	17.9 ± 0.5
Llama-3.1 (8B)	46.3 ± 1.7	53.1 ± 1.7
Llama-3.2 (3B) (fine-tuned)	66.7 ± 0.8	71.6 ± 0.8
Llama-3.1 (8B) (fine-tuned)	78.7 ± 0.2	83.1 ± 0.1

Given the limitation of using six L1s for this analysis, a random guess of the classification would yield an accuracy of approximately 16.7%. Table 3 shows that while the zero-shot Llama-3.1 (8B) model comfortably surpasses this baseline, the fine-tuned models improve on this by an additional 20% to 30%. This substantial enhancement clearly demonstrates that the linguistic features present in the TOEFL11 documents are also discernible in the Reddit-L2 data.

4.3 *Reddit-L2 as a training set*

Next, we perform the inverse of the first experiment to evaluate how a model fine-tuned on Reddit would perform on the TOEFL11 documents. For this purpose, we choose to fine-tune the model on five languages (French, Italian, Turkish, German, and Spanish), excluding Indian languages. This exclusion is due to the insufficient representation of Indian authors, with fewer than 100 authors, which is not adequate for fine-tuning.

As shown in Table 4, the baseline models struggle to classify accurately, with the Llama-3.2 (3B) model performing no better than a random guess (20%). Upon fine-tuning, both models significantly improve one their performance, with the 3B model nearly matching the baseline scores of the 8B model. Most notably, the fine-tuned 8B model achieves accuracy and F2 scores exceeding 90% on both the Reddit-L2 and TOEFL11 datasets, with the TOEFL scores being higher than those for Reddit-L2. This indicates that when trained on the Reddit authors, the model generalises exceptionally well to the TOEFL data. One plausible explanation for this is that TOEFL authors are generally intermediate learners, and as such, they may make more discernible errors more frequently. This characteristic makes TOEFL11 a relatively easier dataset for NLI tasks.

The results of the analysis suggest that noun masking is effective in mitigating the potential lexical-based bias. Additionally, the findings suggest that models trained on Reddit-L2 are able to classify TOEFL11 authors with a high degree of accuracy, relying on more than merely lexical topic-based features. This demonstrates the overall robustness of the proposed approach.

It is important to note that while our results exceed 90%, they should not be directly compared with previous works [14] and [11] since their research included the full set of TOEFL11 L1s. To make a proper comparison, we would need to source authors from the missing L1s.

4.4 *Reddit-L2 for native vs nonnative classification*

The identification of an author as either native or non-native can be beneficial across various fields, including educational strategies

Table 4: Comparative analysis of foundation and Reddit-L2 fine-tuned models on Reddit-L2 average 1200 tokens and TOEFL11 (test-set).

Model	Reddit-L2 (5 L1s)		TOEFL11 (5 L1s)	
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
Llama-3.2 (3B)	15.6 ± 0.0	15.2 ± 0.0	21.8 ± 0.0	19.5 ± 0.0
Llama-3.1 (8B)	48.4 ± 0.2	58.1 ± 0.1	68.4 ± 0.0	70.9 ± 0.0
Llama-3.2 (3B) (fine-tuned)	79.3 ± 0.1	85.5 ± 0.1	67.6 ± 0.1	65.5 ± 0.1
Llama-3.1 (8B) (fine-tuned)	92.0 ± 0.1	92.0 ± 0.1	93.8 ± 0.1	93.6 ± 0.1

Table 5: Comparative analysis of foundation and Reddit-L2 fine-tuned models on Reddit-L2 for native vs non-native, average 1200 tokens.

Model	Accuracy (%)	F1-score (%)
Llama-3.2 (3B)	51.5 ± 0.0	38.3 ± 0.0
Llama-3.1 (8B)	50.8 ± 0.2	39.8 ± 0.1
Llama-3.2 (3B) (fine-tuned)	97.6 ± 0.1	97.6 ± 0.1
Llama-3.1 (8B) (fine-tuned)	98.7 ± 0.1	98.7 ± 0.1

and security. The Reddit-L2 dataset comprises over 40,000 authors, with approximately 12,000 originating from native English-speaking countries. This substantial representation enables us to curate a subset of the dataset containing 24,000 authors, evenly divided between native English authors and a random sample of non-native English authors. We subsequently partition the data into training and testing sets with a 70:30 split.

Table 5 presents the results obtained using the base models and after fine-tuning. The base models perform no better than random guessing, predominantly classifying the majority of documents as non-native. However, once fine-tuned, the models exhibit remarkable performance on the testing set, achieving near-perfect accuracy.

Table 6: Comparative analysis of Reddit-L2 and WI-LOCNESS for native vs non-native. Model used: Llama3.1 (8B) fine-tuned on Reddit-L2.

Dataset	Accuracy	F1-score
Reddit-L2	$87.6\,\% \pm 0.0\,\%$	$87.6\,\% \pm 0.0\,\%$
WI-LOCNESS	$75.0\,\% \pm 0.0\,\%$	$73.3\,\% \pm 0.0\,\%$

To ensure that the models do not overfit on the data or merely identify lexical features, we employed the WI-LOCNESS [3] dataset as an evaluation set. For a fair assessment of the models' capabilities, we compared the validation results with those of the Reddit-L2 test set, limiting the document length to 150 tokens. This constraint was applied to ensure all documents were of similar length, providing the models with an equivalent number of tokens to analyse. As Table 6 shows, although there was a performance drop, the models still achieve reasonable scores on the WI-LOCNESS dataset.

This observation is particularly noteworthy given that, similar to the TOEFL11 corpus, the document style of WI-LOCNESS differs significantly from the concatenated short-form comments of Reddit-L2.

5 Conclusion

In this study, we explored the effectiveness of open-source LLMs in identifying native languages from short-form comments on Reddit, using both the TOEFL11 and Reddit-L2 corpora. Our findings highlight several key insights that contribute to the ongoing research in NLI.

Firstly, our experiments demonstrate that fine-tuning smaller Llama models (3B and 8B) on TOEFL11 can yield significant improvements in accuracy when applied to Reddit-L2 data. This suggests that the linguistic features captured from structured, academic English texts can generalise well to the more informal and varied language use on social media platforms like Reddit. The fine-tuned models significantly outperform baseline models, with accuracy improvements of 20% to 30%, indicating the transferability of learned linguistic characteristics across different text genres.

The inverse experiment of fine-tuning on Reddit-L2 and validating on TOEFL11 shows that models trained on informal text can also generalise effectively to more structured academic writing. The fine-tuned 8B model achieves accuracy and F1 scores exceeding 90% on both datasets, with higher performance on TOEFL11. This outcome underscores the robustness of the model in handling diverse linguistic contexts and suggests that models trained on a wide range of informal texts can successfully adapt to more formal writing styles.

Additionally, our study on classifying authors as native or non-native English speakers reveals that fine-tuned models could achieve near-perfect accuracy on the Reddit-L2 dataset. This classification task is crucial for various applications, including educational strategies and security measures. The models retain reasonable performance on the WI-LOCNESS dataset, further validating their generalisation capability.

One notable observation from our experiments is the influence of document length on model performance. We find that longer documents tend to provide more linguistic features that aid in accurate L1 classification, up to a threshold of approximately 1200 tokens. Beyond this point, the benefits of increased document length diminish. This finding is critical for future NLI research and practical applications, as it emphasises the need to balance document length with computational efficiency.

Our research contributes to the field of NLI by demonstrating the potential of open-source LLMs in handling diverse and informal text genres while maintaining high accuracy. The ability of these models to generalise across different datasets and writing styles highlights their versatility and applicability in real-world scenarios.

6 Limitations and Next Steps

While our study demonstrates the effectiveness of open-source LLMs for NLI, several limitations persist. Firstly, the Reddit-L2 dataset may overlap with the Llama models' pre-training data, potentially influencing the observed performance improvements. A curated dataset collected post-Llama release could mitigate this issue. Additionally, our exploration of document length effects is dataset-specific, requiring further validation across diverse text genres.

The scope of native languages (L1s) in our study is limited, restricting the generality of findings. Expanding the range of L1s, particularly underrepresented ones, is essential for broader applicability. Lastly, our focus on English as the target L2 leaves open the challenge of extending NLI to other L2s, particularly those with fewer resources and greater linguistic variation.

To address these limitations, we propose several directions for future work:

Curate a Post-Llama Reddit-L2 Dataset: Collect Reddit comments posted after Llama's release using the same collection method as the Reddit-L2 dataset to eliminate pre-training data overlap.

Expand Short-Form Sources: Evaluate models on short-form text from platforms like X or Discord to test robustness to different styles of writing (e.g. formality, target audience).

Increase L1 Diversity: Include low-resource native languages to improve multilingual applicability. Additionally, focus on collecting Reddit comments from a more diverse set of languages beyond primarily European languages.

Extend to Non-English L2s: Focus on NLI tasks for other L2s with limited data and greater linguistic diversity to see how methods generalize to other L2s.

Refine Document Length Insights: Investigate the impact of document length across varied text genres for optimal input design.

Conduct Native Language Style Transfer: Use a style transfer pipeline such as the one from Alperin et al. [1] to evaluate the quality of generating text to look like particular L2s (i.e. make non-native Spanish look native or vice versa).

These steps collectively aim to enhance the generalisation, robustness, and applicability of LLMs in NLI tasks.

References

[1] K. Alperin, R. Leekha, A. Uchendu, T. Nguyen, S. Medarametla, C. L. Capote, S. Aycock, and C. Dagli. Masks and mimicry: Strategic obfuscation and impersonation attacks on authorship verification. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 102–116, 2025.

[2] D. Blanchard, J. Tetreault, D. Higgins, A. Cahill, and M. Chodorow. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15, 2013. doi: https://doi.org/10.1002/j.2333-8504.2013.tb02331.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.2013.tb02331.x.

[3] C. Bryant, M. Felice, Ø. E. Andersen, and T. Briscoe. The BEA-2019 shared task on grammatical error correction. In H. Yannakoudakis, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, and T. Zesch, editors, *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4406. URL https://aclanthology.org/W19-4406/.

[4] A. Cimino and F. Dell'Orletta. Stacked sentence-document classifier approach for improving native language identification. In J. Tetreault, J. Burstein, C. Leacock, and H. Yannakoudakis, editors, *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 430–437, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5049. URL https://aclanthology.org/W17-5049/.

[5] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023. URL https://arxiv.org/abs/2305.14314.

[6] D. Goswami, S. Thilagan, K. North, S. Malmasi, and M. Zampieri. 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, 2024.

[7] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[8] E. Lotfi, I. Markov, and W. Daelemans. A deep generative approach to native language identification. In D. Scott, N. Bel, and C. Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1778–1783, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.159. URL https://aclanthology.org/2020.coling-main.159/.

[9] S. Malmasi, S.-M. J. Wong, and M. Dras. NLI shared task 2013: MQ submission. In J. Tetreault, J. Burstein, and C. Leacock, editors, *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–133, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-1716/.

[10] S. Malmasi, K. Evanini, A. Cahill, J. Tetreault, R. Pugh, C. Hamill, D. Napolitano, and Y. Qian. A report on the 2017 native language identification shared task. In J. Tetreault, J. Burstein, C. Leacock, and H. Yannakoudakis, editors, *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5007. URL https://aclanthology.org/W17-5007/.

[11] Y. M. Ng and I. Markov. Leveraging open-source large language models for native language identification. In Y. Scherrer, T. Jauhiainen, N. Ljubešić, P. Nakov, J. Tiedemann, and M. Zampieri, editors, *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 20–28, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.vardial-1.3/.

[12] E. Rabinovich, Y. Tsvetkov, and S. Wintner. Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, 6:329–342, 2018. doi: 10.1162/tacl_a_00024. URL https://aclanthology.org/Q18-1024/.

[13] M. A. team. Mixtral of experts. https://mistral.ai/news/mixtral-of-experts, 2023. Accessed: 2025-04-22.

[14] W. Zhang and A. Salle. Native language identification with large language models, 2023. URL https://arxiv.org/abs/2312.07819.

[15] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, Z. Feng, and Y. Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models, 2024. URL https://arxiv.org/abs/2403.13372.