Leveraging Open-Source Large Language Models for Native Language Identification

Yee Man Ng
CLTL, Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
y.m.ng@student.vu.nl

Ilia Markov
CLTL, Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
i.markov@vu.nl

Abstract

Native Language Identification (NLI) – the task of identifying the native language (L1) of a person based on their writing in the second language (L2) – has applications in forensics, marketing, and second language acquisition. Historically, conventional machine learning approaches that heavily rely on extensive feature engineering have outperformed transformer-based language models on this task. Recently, closed-source generative large language models (LLMs), e.g., GPT-4, have demonstrated remarkable performance on NLI in a zero-shot setting, including promising results in open-set classification. However, closed-source LLMs have many disadvantages, such as high costs and undisclosed nature of training data. This study explores the potential of using open-source LLMs for NLI. Our results indicate that open-source LLMs do not reach the accuracy levels of closed-source LLMs when used out-of-the-box. However, when fine-tuned on labeled training data, open-source LLMs can achieve performance comparable to that of commercial LLMs.

1 Introduction

Native Language Identification (NLI) is the task of automatically identifying an author's native language (L1) based on texts written in their second language (L2). The task is based on the language transfer hypothesis, the phenomenon in which characteristics of L1 influence the production of texts in L2 to the degree that L1 is identifiable (Odlin, 1989). NLI is useful for educational purposes, forensic applications in the context of author profiling, and to inform second language acquisition research (Goswami et al., 2024).

From a machine learning (ML) perspective, NLI is commonly framed as a supervised multi-class classification task, where NLI systems are trained to assign an author's L1. While the task has been proven difficult to perform by humans

(Malmasi et al., 2015), automated methods have shown remarkable results using conventional ML approaches based on extensive feature engineering, e.g., (Cimino and Dell'Orletta, 2017; Markov, 2018). Such methods rely on features that capture L1-indicative linguistic patterns in L2 writing, e.g., spelling errors (Koppel et al., 2005; Chen et al., 2017; Markov et al., 2019), word choice (Brooke and Hirst, 2012), and syntactic patterns (Wong and Dras, 2011).

Transformer-based encoder models, like BERT (Devlin et al., 2019), on the other hand, have yielded poorer performance than conventional ML approaches for the NLI task (Markov et al., 2022; Steinbakken and Gambäck, 2020; Goswami et al., 2024). Previous research suggests that this is likely because NLI concerns very specific linguistic features that models trained on general corpora cannot capture (Markov et al., 2022). Recent research has shown that generative large language models (LLMs) demonstrate promising results for NLI. Lotfi et al. (2020) presented the first study addressing NLI using fine-tuned GPT-2 models, which outperformed previous traditional ML approaches and achieved state-of-the-art results on the NLI benchmark TOEFL11 and ICLE datasets. Zhang and Salle (2023) explored the ability of GPT-3.5 (Brown et al., 2020) and GPT-4 (OpenAI, 2023) to perform NLI. Their results indicate that out-of-the-box GPT models demonstrate outstanding performance, with GPT-4 setting a new performance record of 91.7% accuracy on the TOEFL11 benchmark dataset, and achieve promising results for open-set classification (without a predefined set of L1s), a useful setting for real-world NLI applications.

While Zhang and Salle's results indicate that LLMs achieve state-of-the-art performance on NLI, they only evaluate the performance of GPT-3.5 and GPT-4. The closed-source nature of these models presents a multitude of limitations to research.

Providers of closed-source models often disclose minimal information regarding the training data or procedure, hindering the evaluation of results achieved with these models and obscuring biases in training data and models (Balloccu et al., 2024). The undisclosed nature of the training data has also raised concerns among researchers about data contamination risks, as it is challenging to determine whether a model's high performance on a task can be attributed to the model's effective generalization or potential data leakage (Yu et al., 2023). In addition, closed-source models are typically only accessible via an API, causing lack of control over model updates, which are often communicated poorly to users (Yu et al., 2023; Pozzobon et al., 2023). In turn, the reproducibility of experiments cannot be guaranteed. The usage of closed-source LLMs is also highly costly, which negatively impacts the accessibility of LLMs (Bender et al., 2021).

Providers of open-source LLMs, on the other hand, often release more information regarding training data and procedures. As model weights are released openly, open-source LLMs can be fine-tuned for a down-stream task, which is often highly costly or not supported for closed-source models. Despite these advantages, employing open-source LLMs for NLI remains unexplored, and it is therefore important to investigate the difference in performance between open-source and proprietary LLMs on this task. Hence, the research question addressed in this study is: *Can open-source LLMs be used for effective Native Language Identification?*

The contributions of this work are the following: (i) we are the first to explore the performance of open-source LLMs on NLI and quantify the difference in performance with closed-source models, and (ii) we investigate the impact of fine-tuning open-source LLMs on NLI performance.

2 Data and Models

To comprehensively evaluate the ability of current LLMs to perform NLI, we compare the performance of two closed-source commercial LLMs (i.e., GPT-3.5 and GPT-4) with five open-source LLMs (§2.2), used out-of-the-box and after fine-tuning, on two NLI benchmark datasets.

2.1 Data

TOEFL11 (Blanchard et al., 2013): the ETS Corpus of Non-Native Written English (TOEFL11) consists of 12,100 essays, with 1,100 essays per L1, written by English learners with low, medium, or high proficiency levels. The 11 L1s covered in the data are Arabic (ARA), Chinese (CHI), French (FRE), German (GER), Hindi (HIN), Italian (ITA), Japanese (JPN), Korean (KOR), Spanish (SPA), Telugu (TEL), and Turkish (TUR). We use the TOEFL11 test set for evaluation, which contains 100 essays per L1. The average length of essays in TOEFL11 is 348 words.

ICLE-NLI (Granger et al., 2009): a 7-language subset of the ICLEv2 dataset commonly used for NLI (Tetreault et al., 2012). The data contains 770 essays, with 110 essays per L1, written by highly-proficient English learners. The L1s represented in the dataset are Bulgarian (BUL), Chinese (CHI), Czech (CZE), French (FRE), Japanese (JPN), Russian (RUS), and Spanish (SPA). We evaluate the models on the complete ICLE-NLI dataset. The average length of essays in this corpus is 747 words.

2.2 Models

Baselines We compare the performance of LLMs to several baseline approaches: the best-performing feature-engineered approach (SVM) (Markov, 2018), a simple SVM approach with bag-of-words (BoW) features, BERT and GPT-2 approaches, with all scores directly cited from the original paper (Lotfi et al., 2020).

Closed-source LLMs We rely on the results reported by Zhang and Salle (2023) for GPT-3.5 (gpt-3.5-turbo) (Brown et al., 2020) and GPT-4 (gpt-4-0613) (OpenAI, 2023) on TOEFL11 and evaluate their performance on the ICLE-NLI dataset.

Open-source LLMs We conduct a comparative study of five recent open-source LLMs: LLaMA-2 (7B) (Touvron et al., 2023), LLaMA-3 (8B) (Meta, 2024), Gemma (7B) (Mesnard et al., 2024), Mistral (7B) (Jiang et al., 2023), and Phi-3 (3.8B) (Microsoft, 2024). While there is an ongoing debate surrounding the definition of 'open-source' with the rise of LLMs (Liesenfeld and Dingemanse, 2024), for the purpose of our experiments, we consider open-source models that are open in weights. Following Zhang and Salle (2023), we carry out experiments in a zero-shot setup, both for the closed-set and open-set NLI tasks.

We run inference on the selected open-source LLMs using the same prompt as Zhang and Salle (2023), with the only difference that we instruct each model to respond using JSON dictionaries to

| Model | TOEFL11 (11 L1s, test set) | | ICLE-NLI (7 L1s, 5FCV/entire) | |
	Closed-set	Open-set	Closed-set	Open-set
Baselines				
BoW SVM (Lotfi et al., 2020)	71.1	–	80.6	–
Feature-engineered SVM (Markov, 2018)	88.6	–	93.4	–
BERT (Lotfi et al., 2020)	80.8	–	76.8	–
GPT-2 (fine-tuned) (Lotfi et al., 2020)	89.0	–	94.2	–
GPT-3.5 (Zhang and Salle, 2023)	74.0	73.4	81.2	84.2
GPT-4 (Zhang and Salle, 2023)	**91.7**	**86.7**	95.5	**89.1**
Open-source LLMs				
LLaMA-2 (7B) (zero-shot)	$29.2_{\pm 0.9}$	$22.1_{\pm 0.7}$	$29.2_{\pm 1.0}$	$15.5_{\pm 0.3}$
LLaMA-2 (7B) (fine-tuned)	$78.7_{\pm 1.0}$	–	$42.9_{\pm 2.0}$	–
LLaMA-3 (8B) (zero-shot)	$56.8_{\pm 1.1}$	$56.4_{\pm 0.7}$	$75.8_{\pm 0.4}$	$71.0_{\pm 0.9}$
LLaMA-3 (8B) (fine-tuned)	$85.3_{\pm 0.1}$	–	$78.5_{\pm 2.5}$	–
Gemma (7B) (zero-shot)	$13.6_{\pm 0.0}$	$7.0_{\pm 0.0}$	$28.2_{\pm 0.1}$	$13.1_{\pm 0.0}$
Gemma (7B) (fine-tuned)	$90.3_{\pm 1.2}$	–	$\mathbf{96.6}_{\pm 0.2}$	–
Mistral (7B) (zero-shot)	$35.6_{\pm 1.6}$	$24.2_{\pm 0.1}$	$53.1_{\pm 1.1}$	$41.5_{\pm 0.1}$
Mistral (7B) (fine-tuned)	$89.8_{\pm 0.8}$	–	$83.2_{\pm 9.4}$	–
Phi-3 (3.8B) (zero-shot)	$18.2_{\pm 0.3}$	$21.6_{\pm 1.6}$	$33.6_{\pm 0.4}$	$40.9_{\pm 2.1}$
Phi-3 (3.8B) (fine-tuned)	$65.6_{\pm 0.4}$	–	$51.4_{\pm 1.7}$	–

Table 1: Comparative analysis of the performance of the baseline methods and closed- and open-source LLMs on the TOEFL11 and ICLE-NLI datasets in terms of classification accuracy (%).

restrict the model output to one L1 classification label. For the closed-set task, we include the set of possible L1s in the prompt. If the model classifies an L1 outside of the provided set of classes, we apply iterative prompting up to 5 times. For the open-set task, the prompt does not include a set of possible L1s. For both closed- and open-set tasks, we adapt the prompt to each model's prompt template. If a prediction cannot be extracted after 5 attempts, the predicted label is set to 'other'. The prompts for closed-set and open-set tasks are provided in appendices C.1 and C.2, respectively. We use 4-bit quantized instruction-fine-tuned versions of the open-source LLMs when prompting out-of-the-box.

In addition, we fine-tune the 4-bit quantized models on the TOEFL11 training set and under 5-fold cross-validation (5FCV) on ICLE-NLI[1] with QLoRA (Dettmers et al., 2023), using the Hugging Face framework and Unsloth library[2]. The prompts used for fine-tuning are provided in Appendix C.3.

---

[1]We used 5-fold cross-validation for a direct comparison with previous studies, e.g., (Lotfi et al., 2020; Markov, 2018).

[2]https://unsloth.ai/

3 Results

Table 1 shows the results in terms of classification accuracy (%) for the baseline approaches and LLMs, both out-of-the-box and after fine-tuning, in closed-set and open-set settings. For open-source LLMs, we provide the average score and standard deviation over three runs to account for stochasticity in model inference and training.

3.1 Closed-Source LLMs

We observe high accuracy scores on the ICLE-NLI dataset in our experiments using the GPT-3.5 and GPT-4 models. The results are in line with the state-of-the-art results on the TOEFL11 dataset reported in (Zhang and Salle, 2023) and indicate that GPT-4 is able to identify the L1s of highly-proficient English learners both in closed-set and open-set classification experiments.

3.2 Open-Source LLMs Out-of-the-Box

We note a surprisingly low performance of open-source LLMs when used out-of-the-box in a closed-set setting, with the exception of LLaMA-3 on ICLE-NLI. While GPT-4 achieves an accuracy of 91.7% and 95.5% on TOEFL11 and ICLE-

NLI, respectively, the five open-source models obtain accuracy scores ranging between 13.6% and 75.8%. All open-source LLMs also perform worse than the baseline approaches, including the simple SVM model with BoW features. Some open-source LLMs tend to predict mostly one or two languages, e.g., Gemma predicting mostly French and LLaMA-2 mostly Chinese, which partially explains such low results. The large performance gap raises the concern that closed-source LLMs might have seen the NLI benchmark datasets in training. Additional research is required to explore the possibility of data leakage, e.g., by examining whether a model has memorized a given text using perplexity measurements (Carlini et al., 2021).

3.3 Fine-Tuned Open-Source LLMs vs. Closed-Source LLMs

The results indicate that the performance of open-source LLMs improves substantially after task-specific fine-tuning. Fine-tuned Gemma achieves an accuracy score of 90.3% (±1.2) on the TOEFL11 dataset, nearly matching the results of GPT-4 as reported in (Zhang and Salle, 2023), and a near-perfect accuracy score of 96.6% (±0.2) on the ICLE-NLI dataset, outperforming GPT-4 by 1.1%. We also observe that the open-source models that perform best out-of-the-box do not necessarily demonstrate the best performance after fine-tuning.

Previous studies comparing closed-source and fine-tuned open-source LLMs provide contradictory findings, with some researchers reporting a drop in accuracy of 16% on sentiment classification for fine-tuned smaller language models (Flan-T5, 770M) compared to ChatGPT (Zhang et al., 2024), while others report that fine-tuned open-source LLMs (Qwen, 7B; LLaMA-3, 8B) outperform closed-source LLMs (GPT-3.5, GPT-4) on text classification tasks (Bucher and Martini, 2024; Edwards and Camacho-Collados, 2024; Wang et al., 2024). The results presented in this study provide evidence that fine-tuned open-source LLMs can achieve comparable performance to closed-source LLMs.

We also observe that LLaMA-3 stands out with a high result on ICLE-NLI compared to TOEFL11. While out-of-the-box LLaMA-3 obtains 56.6% accuracy on TOEFL11, it achieves a higher score of 75.8% on ICLE-NLI. In addition, while all other open-source LLMs gain a large boost in performance after fine-tuning on both datasets, LLaMA-3's accuracy after fine-tuning

on ICLE-NLI increases by 2.7 percentage points only. LLaMA-3's relatively high performance out-of-the-box and marginal performance boost after fine-tuning are inconsistent with the results for other open-source LLMs, possibly indicating that LLaMA-3 has seen the ICLE data in training.

Comparing the confusion matrices for GPT-4 and fine-tuned Gemma, the best-performing closed-source and open-source LLMs (Appendix B), we note that both models tend to misclassify Hindi texts as Telugu in the TOEFL11 dataset. Hindi and Telugu have been considered a problematic language pair in previous studies on TOEFL11 (Malmasi et al., 2013). Fine-tuned Gemma has a tendency to misclassify Japanese essays as Korean. The high degree of confusion between Korean and Japanese has also been observed in previous research (Markov et al., 2022). On ICLE-NLI, GPT-4 erroneously classifies Bulgarian as Russian, both Slavic languages. Gemma misclassifies 14 Czech and Russian samples as Bulgarian. In line with previous research, we note that the confused L1s are either related through geographical location or belong to the same language family.

3.4 Closed-Set and Open-Set Settings

We observe a drop in performance for most open-source LLMs from a closed-set to open-set setting, similarly to closed-source LLMs. Surprisingly, some of the models, i.e., GPT-3.5 and Phi-3, perform better in the open-set than in the closed-set setup. Further research is required to understand the reasons for this behaviour.

4 Conclusion

We explored the performance of a variety of open-source LLMs for the NLI task. Our results indicate that open-source LLMs achieve lower performance than closed-source LLMs for this task when used out-of-the-box, while domain-specific fine-tuning of open-source LLMs allows these models to achieve comparable results to the proprietary LLMs, such as GPT-4, on the benchmark TOEFL11 and ICLE-NLI datasets. We believe that our work opens up avenues for future research on LLM-based Native Language Identification. Future research could explore few-shot prompting and different prompt variations as a way to potentially boost the performance of open-source LLMs.

Limitations

Multilingual NLI   Our study focuses purely on native language identification in English, which is the most well-studied L2 in the NLI task (Goswami et al., 2024). It would be interesting to explore whether the high performance of LLMs on NLI holds for L2s other than English.

Fine-tuned LLMs in cross-corpus setting While fine-tuning drastically improves the performance of open-source LLMs, the prerequisite of fine-tuning for optimal performance is a disadvantage for open-source LLMs compared to closed-source LLMs. Previous research has shown that NLI models suffer from performance degradation in a cross-corpus setting, and thus cannot be applied directly to different corpora (Markov et al., 2022; Malmasi and Dras, 2015). Future research could explore the use of fine-tuned open-source LLMs for NLI in a cross-corpus setup.

Defining open-source LLMs   More broadly, in our study, we define open-source and closed-source relatively loosely, treating the terms 'open' and 'closed' as a binary feature to perform a comparative analysis between open-source and closed-source LLMs for NLI. However, there are various dimensions of openness, as a model release involves different components ranging from the disclosure of training datasets to model access (Solaiman, 2023; Liesenfeld and Dingemanse, 2024). Most providers of proclaimed open-source LLMs release little to no information regarding their training data and procedure, despite framing them as being open-source. In turn, it is difficult to determine whether an open-source model's performance can be attributed to the model's learning or possible data contamination. The lack of insights into the training data of proclaimed open-source LLMs also hindered our evaluation of LlaMA-3 on the ICLE-NLI dataset.

References

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual Event, Canada. Association for Computing Machinery, Inc.

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.

Julian Brooke and Graeme Hirst. 2012. Robust, lexicalized native language identification. In *Proceedings of COLING 2012*, pages 391–408, Mumbai, India. The COLING 2012 Organizing Committee.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned 'small' llms (still) significantly outperform zero-shot generative ai models in text classification. *arXiv*, arXiv:2406.08660.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Lingzhen Chen, Carlo Strapparava, and Vivi Nastase. 2017. Improving native language identification by using spelling errors. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 542–546, Vancouver, Canada. Association for Computational Linguistics.

Andrea Cimino and Felice Dell'Orletta. 2017. Stacked sentence-document classifier approach for improving native language identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 430–437, Copenhagen, Denmark. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv*, arXiv:2305.14314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies), pages 4171–4186, Minneapolis, USA. Association for Computational Linguistics.

Aleksandra Edwards and Jose Camacho-Collados. 2024. Language models for text classification: Is in-context learning enough? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10058–10072, Torino, Italia. ELRA and ICCL.

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, Mexico City, Mexico. Association for Computational Linguistics.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English v2 (ICLE)*. Presses Universitaires de Louvain, Louvain-la-Neuve, Belgium.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv*, arXiv:2310.06825.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. 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*, KDD '05, page 624–628, New York, NY, USA. Association for Computing Machinery.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv*, arXiv:1910.09700.

Andreas Liesenfeld and Mark Dingemanse. 2024. Rethinking open source generative ai: open-washing and the eu ai act. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1774–1787, New York, NY, USA. Association for Computing Machinery.

Ehsan Lotfi, Ilia Markov, and Walter Daelemans. 2020. A deep generative approach to native language identification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1778–1783, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Shervin Malmasi, Joel Tetreault, and Mark Dras. 2015. Oracle and human baselines for native language identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 172–178, Denver, Colorado. Association for Computational Linguistics.

Shervin Malmasi, Sze-Meng Jojo Wong, and Mark Dras. 2013. NLI shared task 2013: MQ submission. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–133, Atlanta, Georgia. Association for Computational Linguistics.

Ilia Markov. 2018. *Automatic Native Language Identification*. Ph.D. thesis, Instituto Politécnico Nacional, Mexico City, Mexico.

Ilia Markov, Vivi Nastase, and Carlo Strapparava. 2019. Anglicized words and misspelled cognates in native language identification. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 275–284, Florence, Italy. Association for Computational Linguistics.

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

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv*, arXiv:2403.08295.

Meta. 2024. Llama 3. *Meta Blog*. Accessed: 20 May 2024.

Microsoft. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv*, arXiv:2404.14219.

Terence Odlin. 1989. *Language Transfer: cross-linguistic influence in language learning*. Cambridge University Press, Cambridge, UK.

OpenAI. 2023. Gpt-4 technical report. *arXiv*, arXiv:2303.08774.

Luiza Pozzobon, Beyza Ermis, Patrick Lewis, and Sara Hooker. 2023. On the challenges of using black-box APIs for toxicity evaluation in research. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7595–7609, Singapore. Association for Computational Linguistics.

Irene Solaiman. 2023. The gradient of generative ai release: Methods and considerations. *arXiv*, arXiv:2302.04844.

Stian Steinbakken and Björn Gambäck. 2020. Native-language identification with attention. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 261–271, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Joel Tetreault, Daniel Blanchard, Aoife Cahill, and Martin Chodorow. 2012. Native tongues, lost and found: Resources and empirical evaluations in native language identification. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 2585–2602, Mumbai, India. The COLING 2012 Organizing Committee.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv*, arXiv:2307.09288.

Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2024. Smart expert system: Large language models as text classifiers. *arXiv*, arXiv:2405.10523.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Hao Yu, Zachary Yang, Kellin Pelrine, Jean Francois Godbout, and Reihaneh Rabbany. 2023. Open, closed, or small language models for text classification? *arXiv*, arXiv:2308.10092.

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

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

A Hyperparameters and Computation Time

We fine-tuned the open-source LLMs with the following hyperparameters: a learning rate of 1e-4, batch size of 16, 3 epochs, and optimization via AdamW optimizer. The experiments were conducted on Google Colaboratory Pro with the A100 GPU (40 GB RAM). The models were loaded with 4-bit NF-quantization and QLoRA adapters were added and fine-tuned using the bitsandbytes library[3]. The total computation time was roughly 120 hours. Total emissions are estimated to be 17.1 kgCO$_2$eq of which 100% was directly offset by the cloud provider[4].

B Confusion Matrices

The confusion matrices are provided in Figure 1.

---

[3]https://huggingface.co/docs/bitsandbytes
[4]Estimations were conducted using the Machine Learning Impact calculator (Lacoste et al., 2019).

C LLM Prompts

C.1 Closed-Set Prompts

For the closed-set experiments on the TOEFL11 dataset, we used the prompts below. For ICLE-NLI, we used exactly the same prompts, with the only difference being the set of possible L1s covered in the dataset.

> You are a forensic linguistics expert that reads English texts written by non-native authors to classify the native language of the author as one of:
>
> "ARA": Arabic
> "CHI": Chinese
> "FRE": French
> "GER": German
> "HIN": Hindi
> "ITA": Italian
> "JPN": Japanese
> "KOR": Korean
> "SPA": Spanish
> "TEL": Telugu
> "TUR": Turkish
> Use clues such as spelling errors, word choice, syntactic patterns, and grammatical errors to decide on the native language of the author.
>
> DO NOT USE ANY OTHER CLASS.
> IMPORTANT: Do not classify any input as "ENG" (English). English is an invalid choice.
>
> Valid output formats:
> Class: "ARA",
> Class: "CHI",
> Class: "FRE",
> Class: "GER"
>
> You ONLY respond in JSON files. The expected output from you is: json {"native_lang": The chosen class, ARA, CHI, FRE, GER, HIN, ITA, JPN, KOR, SPA, TEL, or TUR}

When possible, the prompt above was entered as a System prompt. If the system role was not supported by the prompt formatter, the prompt was entered as part of the User prompt. We input the given text and used the prompt below as a User prompt:

> <TOEFL11 ESSAY TEXT>
> Classify the text above as one of ARA, CHI, FRE, GER, HIN, ITA, JPN, KOR, SPA, TEL, or TUR. Do not output any other class - do NOT choose "ENG" (English). What is the closest native language of the author of this English text from the given list?

In the closed-set experiments, if the L1 was incorrectly predicted as English, we prompted the model again using the prompt below:
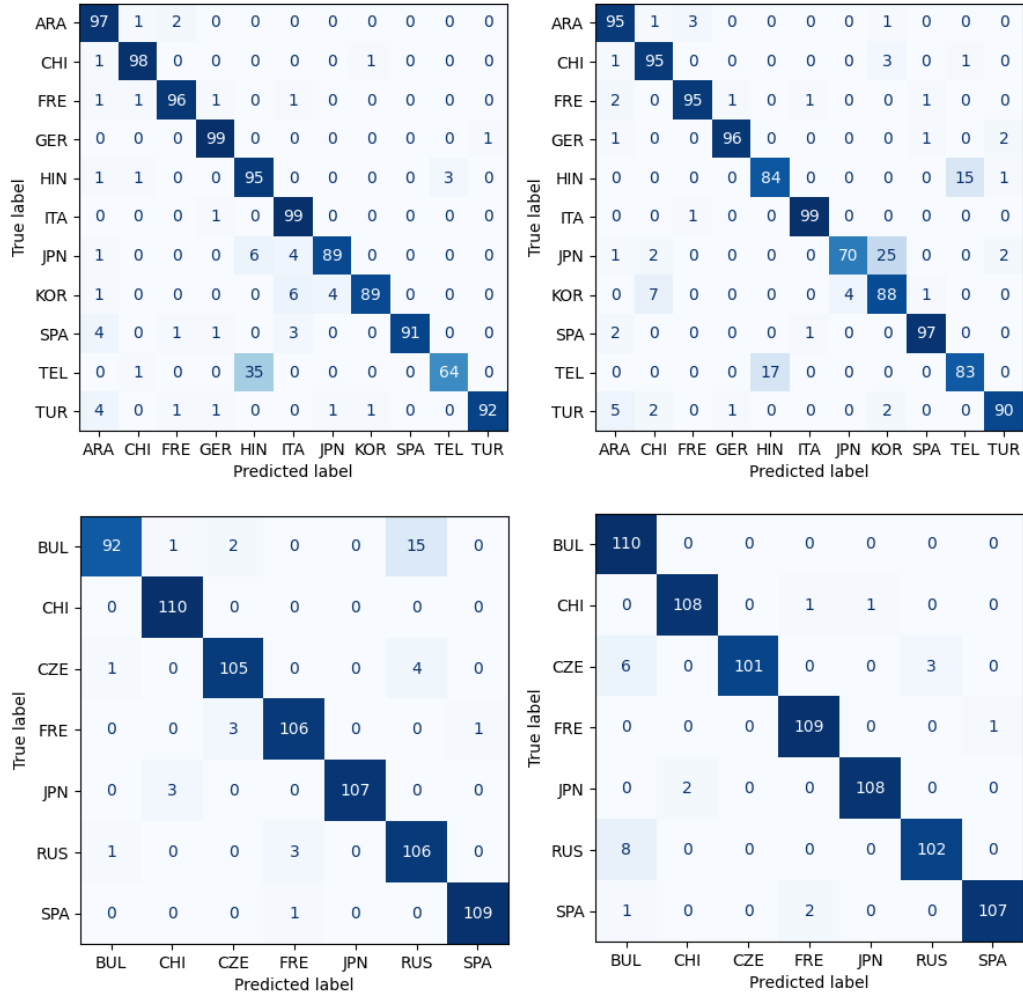
Figure 1: Confusion matrices for GPT-4 on TOEFL (Zhang and Salle, 2023) (top left), Gemma (7B) (fine-tuned) on TOEFL (top right). GPT-4 on ICLE-NLI (bottom left), Gemma (7B) (fine-tuned) on ICLE-NLI (bottom right).

> You previously mistakenly predicted this text as "ENG" (English). The class is NOT English. Please classify the native language of the author of the text again.

If we were unable to parse the prediction or the predicted L1 was not in the set of possible classes, we prompted the model again. For the TOEFL11 experiments, we used the prompt below:

> Your classification is not in the list of possible languages.
> Please try again and choose only one of the following classes: ARA, CHI, FRE, GER, HIN, ITA, JPN, KOR, SPA, TEL, or TUR

C.2 Open-Set Prompts

For the open-set experiments, we used the prompt below as an input prompt for all the models:

> You are a forensic linguistics expert that reads texts written by non-native authors in order to identify their native language.
> Analyze each text and identify the native language of the author.
> Use clues such as spelling errors, word choice, syntactic patterns, and grammatical errors to decide.
>
> You ONLY respond in JSON files. The expected output from you has to be: "json {"native_lang": ""}"

If the predicted L1 could not be extracted from the generated output, we used the prompt below to apply iterative prompting to get a valid prediction:

> Your previous classification was not in the correct format. Please only respond in the following JSON format:
> "json {"native_lang": ""}"

C.3 Fine-Tuning Prompts

We used the following prompt for the fine-tuning experiments:

Instruction:
You are a forensic linguistics expert that reads English texts written by non-native authors to classify the native language of the author as one of:

"ARA": Arabic
"CHI": Chinese
"FRE": French
"GER": German
"HIN": Hindi
"ITA": Italian
"JPN": Japanese
"KOR": Korean
"SPA": Spanish
"TEL": Telugu
"TUR": Turkish
Use clues such as spelling errors, word choice, syntactic patterns, and grammatical errors to decide on the native language of the author.

DO NOT USE ANY OTHER CLASS.
IMPORTANT: Do not classify any input as "ENG" (English). English is an invalid choice.

Valid output formats:
Class: "ARA",
Class: "CHI",
Class: "FRE",
Class: "GER"

Classify the text below as one of ARA, CHI, FRE, GER, HIN, ITA, JPN, KOR, SPA, TEL, or TUR. Do not output any other class - do NOT choose "ENG" (English). What is the closest native language of the author of this English text from the given list?

Input:
<TOEFL11 ESSAY TEXT>

Response:
<L1 LABEL>