ments in Table 4 combined these variables to provide a unified performance comparison. We note that while this offers a broad overview of the three evaluation paradigms (prompting, fine-tuning, and linguistic features) across languages, future work should include dedicated modeling and evaluation by text category, which may warrant a more focused, in-depth study we assign for future work.

Language Availability and Dependency. Due to the nature of UNIVERSALCEFR being a standardized collection of open-sourced, publicly accessible CEFR data, its growth depends heavily on how the community will move forward and continuously release artifacts, including CEFR-annotated corpora for reproducibility and wider access for research purposes. We also acknowledge the efforts of researchers who work on multi-framework adoption, where CEFR descriptors and bands are overlapped with languages not within Europe (such as Hindi (Naous et al., 2024) and Arabic (Habash and Palfreyman, 2022)), and continue to open-source the annotated data.

Modalities Beyond Texts. The current data collection scope of UNIVERSALCEFR and the insights presented in this work only cover CEFR-based texts for now, specifically for reading and writing specifications. Multimodal data, such as audio and video recordings of learners associated with CEFR specifications for listening and speaking, are not yet covered. Naturally, these datasets are even more challenging to acquire and open-source, especially if they contain materials from or are created by learners under legal age and if they contain personal information.

Beyond Typical Benchmarking The rigor of analysis in this paper is not meant to be treated as a typical benchmark study, similar to recent trends in NLP papers, where the goal is to evaluate as many LLMs as possible. In this paper, we provide deeper insights into language complexities and intricacies that affect model performance in CEFR level classification across various dimensions of language, granularity, and format. Thus, within our compute budget, we carefully handpicked state-of-the-art LLMs that are worth exploring based on their properties (e.g., English-centric against massively multilingual, or linguistic features against fine-tuning and prompting). We leave the evaluation on larger, more advanced LLMs, as well as explorations in other directions to improve CEFR level classification, such as the use of high-quality synthetic datasets, for future work.

Ethics Statement

As mentioned throughout this paper, all the datasets we collected for UNIVERSALCEFR based on our criteria presented in Section 3 are already publicly accessible with permissive licenses, and can be used for non-commercial research purposes. While there are three corpora from UNIVERSALCEFR—namely APA-LHA, DEplain, EFCAMDAT—that require users to fill a short form and agree to terms, we still classified them as publicly accessible due to the quick response to access approval.

In the context of the EU AI Act, the use of AI systems for educational purposes, especially those that are intended to *"to evaluate learning outcomes, including when those outcomes are used to steer the learning process of natural persons in educational and vocational training institutions at all levels"* (European Parliament and Council, 2024), is classified under *high risk*. Thus, AI systems that will be released in the market with these goals are required to comply with obligations for high-risk systems, including data governance with high-quality, representative datasets. As a form of contribution towards meeting these requirements, the UNIVERSALCEFR is an initiative that will allow researchers and developers access to diverse, multilingual, multidimensional CEFR-labeled texts which can be used for designing systems that are representative, explainable, and fair.

Acknowledgments

JMI is supported by the National University Philippines and the UKRI Centre for Doctoral Training in Accountable, Responsible, and Transparent AI [EP/S023437/1] of the University of Bath.

HS has received funding from the European Union's Horizon Europe research and innovation program under Grant Agreement No. 101132431 (iDEM Project). HS also receives support from the Spanish State Research Agency under the Maria de Maeztu Units of Excellence Programme (CEX2021-001195-M) and from the Departament de Recerca i Universitats de la Generalitat de Catalunya (ajuts SGR-Cat 2021).

DK and FAM have received funding from the Welsh Government as part of the "Developing a CEFR Predictor for Welsh (2025-26)" project.

ER is supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (Reference: UIDB/50021/2020, DOI: 10.54499/UIDB/50021/2020) and by the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837).

References

Kais Allkivi, Pille Eslon, Taavi Kamarik, Karina Kert, Jaagup Kippar, Harli Kodasma, Silvia Maine, and Kaisa Norak. 2024. ELLE-Estonian Language Learning and Analysis Environment. *Baltic Journal of Modern Computing*, 12(4).

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Cristina Arhiliuc, Jelena Mitrović, and Michael Granitzer. 2020. Language proficiency scoring. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5624–5630, Marseille, France. European Language Resources Association.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Wilbert Berendsen and Kozea. 2025. Pyphen. `https://github.com/Kozea/Pyphen`.

Olga Blinova and Nikita Tarasov. 2022. A hybrid model of complexity estimation: Evidence from russian legal texts. *Frontiers in Artificial Intelligence*, 5:1008530.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M. Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, João Alves, Kevin El-Haddad, Manuel Faysse, Maxime Peyrard, Nuno M. Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. 2025. Eurobert: Scaling multilingual encoders for european languages. *Preprint*, arXiv:2503.05500.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1281–1288, Reykjavik, Iceland. European Language Resources Association (ELRA).

Mark Breuker. 2022. CEFR labelling and assessment services. In *European Language Grid: A Language Technology Platform for Multilingual Europe*, pages 277–282. Springer International Publishing Cham.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Andrew Caines and Paula Buttery. 2020. REPROLANG 2020: Automatic proficiency scoring of Czech, English, German, Italian, and Spanish learner essays. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5614–5623, Marseille, France. European Language Resources Association.

Davide Colla, Matteo Delsanto, and Elisa Di Nuovo. 2023. EliCoDe at MultiGED2023: fine-tuning XLM-RoBERTa for multilingual grammatical error detection. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, pages 24–34, Tórshavn, Faroe Islands. LiU Electronic Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ–IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, Scotland, UK. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

European Parliament and Council. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). `https://eur-lex.europa.eu/eli/reg/2016/679/oj`. OJ L 119, 4.5.2016, p. 1–88.

European Parliament and Council. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts. https://eur-lex.europa.eu/eli/reg/2024/1689/oj. OJ L 2024/1689, 12.7.2024, p. 1–88.

Neus Figueras. 2012. The impact of the CEFR. *ELT journal*, 66(4):477–485.

Jeroen Geertzen, Theodora Alexopoulou, Anna Korhonen, and 1 others. 2013. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project*, pages 240–254.

Gemma Team. 2024. Gemma: Open Models Based on Gemini Research and Technology. *arXiv preprint arXiv:2403.08295*.

Gemma Team. 2025. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, Magali Paquot, and 1 others. 2009. *International corpus of learner English*, volume 2. UCL, Presses Univ. de Louvain.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Nizar Habash and David Palfreyman. 2022. ZAEBUC: An annotated Arabic-English bilingual writer corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 79–88, Marseille, France. European Language Resources Association.

Claudia Harsch. 2014. General Language Proficiency Revisited: Current and Future Issues. *Language Assessment Quarterly*, 11(2):152–169.

Junyi He and Xia Li. 2024. Zero-shot cross-lingual automated essay scoring. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17819–17832, Torino, Italia. ELRA and ICCL.

Yan Huang, Jeroen Geertzen, Rachel Baker, Anna Korhonen, Theodora Alexopoulou, and EF Education First. 2017. The EF Cambridge open language database (EFCAMDAT): Information for users.

Joseph Marvin Imperial and Ekaterina Kochmar. 2023a. Automatic readability assessment for closely related languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5371–5386, Toronto, Canada. Association for Computational Linguistics.

Joseph Marvin Imperial and Ekaterina Kochmar. 2023b. BasahaCorpus: An expanded linguistic resource for readability assessment in Central Philippine languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6302–6309, Singapore. Association for Computational Linguistics.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2024. SpeciaLex: A benchmark for in-context specialized lexicon learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 930–965, Miami, Florida, USA. Association for Computational Linguistics.

O Jantunen, Sisko Brunni, and University of Oulu, Department of Finnish Language. 2013. International Corpus of Learner Finnish.

Matias Jentoft and David Samuel. 2023. NoCoLA: The Norwegian corpus of linguistic acceptability. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 610–617, Tórshavn, Faroe Islands. University of Tartu Library.

Nikola Ljubešić. 2018. Concreteness and imageability lexicon MEGA.HR-crossling. Slovenian language resource repository CLARIN.SI.

Louis Martin, Samuel Humeau, Pierre-Emmanuel Mazaré, Éric de La Clergerie, Antoine Bordes, and Benoît Sagot. 2018. Reference-less quality estimation of text simplification systems. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 29–38, Tilburg, the Netherlands. Association for Computational Linguistics.

Cristina Martins, T Ferreira, M Sitoe, C Abrantes, M Janssen, A Fernandes, A Silva, I Lopes, I Pereira, and J Santos. 2019. Corpus de produções escritas de aprendentes de PL2 (PEAPL2): Subcorpus Português língua estrangeira. *Coimbra: CELGA-ILTEC*.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, and 1 others. 2024. EuroLLM: Multilingual Language Models for Europe. *arXiv preprint arXiv:2409.16235*.

Arianna Masciolini, Andrew Caines, Orphée De Clercq, Joni Kruijsbergen, Murathan Kurfalı, Ricardo Muñoz Sánchez, Elena Volodina, Robert Östling, Kais Allkivi, Špela Arhar Holdt, Ilze Auzina, Roberts Darģis, Elena Drakonaki, Jennifer-Carmen Frey, Isidora Glišić, Pinelopi Kikilintza, Lionel Nicolas, Mariana Romanyshyn, Alexandr Rosen, and 11 others. 2025. Towards better language representation in Natural Language Processing: A multilingual dataset for text-level Grammatical Error Correction. *International Journal of Learner Corpus Research*.

Amália Mendes, Sandra Antunes, Maarten Janssen, and Anabela Gonçalves. 2016. The COPLE2 corpus: a learner corpus for Portuguese. In *Proceedings*