

L2-Bench: An Evaluation Benchmark for Measuring LLM Capabilities in Second Language Education (Extended Abstract)

James Edgell^{*1}, Wm. Matthew Kennedy^{*2, 3}, Ben Knight¹,
Danielle Carvalho¹, Martin Ku¹, Isaac Pattis¹

¹Oxford University Press

²Oxford Internet Institute, University of Oxford

³King's College London
elt-bench@oup.com

Extended Abstract

Despite rapid adoption of AI systems in educational spaces, very few evaluations for AI in educational (AIED) exist. Those that purport to cover this space often suffer from poor construct validity, are underpowered, or focus more on broad effects rather than instance-level performativity. The result is a situation in which deployment is far outpacing evidence even of basic validation of AIED systems.

This AIED evaluations gap reflects the ongoing crisis in AI evaluations in general. However, it is more acute: although prior validation and efficacy studies of conventional education technologies suggest that these predecessor systems have not meaningfully improved learning outcomes (Cuban 2001; Zawacki-Richter et al. 2019; Selwyn 2012, 2019), many AIED systems replicate or extend the same conventional system design patterns (Jurenka et al. 2024). Such a headlong rush on the part of AIED developers and adopters without parallel development of novel evaluation methodologies creates the conditions for myriad interaction-, learning-group-, systemic-, and compounding harms (Bastani et al. 2024; Holmes 2024; Kennedy and Campos 2025). In the meantime, a generation of learners must endure laboratory conditions (Alcaras and Ricci 2025).

Drawing inspiration from allied efforts to improve the state of the evaluations ecosystem (Eriksson et al. 2025; Weidinger et al. 2025; Schwartz et al. 2025; Reuel et al. 2025), our work hastens progress in one particularly challenging subdomain of AI for education: second language (L2) education. The use of LLMs for language learning tasks represents one of the most common applications of AI models, yet it is among the least evaluated. To work towards closing this gap, we propose an evaluation benchmark for **second language learning experience design** for L2 US/UK English. Our benchmark makes three distinct contributions:

1) It provides a validated taxonomy of the 12 competencies and 31 subcompetencies that compose second language learning design, one which operationalizes discipline-specific practices and frameworks into a sociotechnical boundary object that promotes the uptake of domain knowl-

edge into current AI systems design. A large sample of $n = 221$ expert practitioner raters scored task authenticity and criteria adequacy highly (task authenticity = 4.42/5.00, criteria adequacy = 4.18/5.00).

2) It offers an evaluation methodology for assessing model performance across each of these competences. We advance, as a research agenda item rather than a validated claim, the conjecture that this rubric-based methodology could—with substantial domain adaptation—be extended to other educational domains that also feature well-developed assessment and pedagogical frameworks within SHAPE disciplines (i.e. the social sciences and humanities); establishing this generalization empirically is left to future work and discussed in Appendix C.

3) It produces a benchmark dataset of 1000+ task-response item pairs structured by our taxonomy, ultimately allowing stakeholders as diverse as instructors, administrators, policymakers, learning scientists, AI evaluations scientists, AI application developers, and AI model developers to better understand the current and near-future state of AI model capabilities in this “high risk” social sector.

We evaluate nine popular models at the time of writing, finding that, overall, Claude Opus 4.7 performs best on benchmark tasks (85.5%), but is marginally less performant on some competencies than other frontier models (managing activities, presenting linguistic points, student assessment). On “hard” tasks, top tier model performance ranges from 69.9% to 73.4%, suggesting both room for improvement and reason to develop more complex evaluation datasets.

We propose this benchmark as the first offering of our wider program to develop context-specific AIED evaluation methodologies. As such, L2-Bench prioritizes essential desiderata that second language learning design benchmarking should seek first to capture. We also provide insights into new directions for AIED evaluations methods in general.

Repositories. The evaluation codebase and benchmark dataset (tasks, rubrics, and reference answers) is available at <https://huggingface.co/datasets/OUP/l2-bench>.

Licensing. The dataset, rubrics, and reference answers are released under the Creative Commons Attribution–ShareAlike 4.0 International licence (CC-BY-SA-4.0); the evaluation code is released under the MIT licence.

^{*}These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Acknowledgements

L2-Bench was a collaborative effort and there are a great many people for their assistance and advice on this project. With thanks to Professor Elizabeth Wonnacott, Department of Education, University of Oxford, who provided advice on statistical tests and experimental design to support the validation analysis, and to Beatrice Segura Harvey, ELT specialist, for contributing to L2-Bench dataset creation. Thank you to our wider team and colleagues at Oxford University Press, in particular to Megan Gericke, for her support in facilitating the validation study and helping project delivery, and Dorian McCree for his continued support and advice since project inception. We would like to thank all 221 education practitioners who contributed their time and expertise to the 2026 “OUP Global Practitioner Challenge”, in particular to those who ultimately helped to guide the development of L2-Bench and future iterations: Anna Król, Warsaw University of Technology; Yolanda Xavier, NOVA University Lisbon; Ozlem Terzioğlu, British Council; Rachel Toncelli, Northeastern University; Wiktorja Allan, Berlin School of Economics and Law; Gülbahar Vidinel, Dartaşşafaka Educational Institutions; Richard Delme Phillips; Deise Amaral, UFRGS; Serhii Andrusienko, Swiss NeuroLanguage Academy ANDRUSENKO.PRO; Hsin Yun Ho, University of Birmingham; Taiki Shimosakai, University of Birmingham; Rosangela Misciagna, Ca’ Foscari University Venice; Alan S. Mackenzie, Mackenzie Education; David E. Ponce, Instituto de Lenguas Extranjeras; Adriana Maria Butnariu, Institut Camps Blancs; Marli Silva Pereira; Agnieszka Tyszkiewicz-Zora, University of Łódź; Carla Marmorale; Sabrina Boem; Monica Lunardon; Silvia Barlassina, IIS “Altiero Spinelli” Sesto San Giovanni; Masooma Amjad, University of Azad Jammu and Kashmir; Anna Maria Rucolo, Milan ILC; John Bletsas, Hamlet EFL school; Tory S. Thorkelson, Sejong University; Magdalena Ecaterina Tolea, Ienachita Vacarescu National College; Valentina Tur-rini, Lower secondary school “F.Cipriani” Nogara; Liliia Okhotina Okhotina; Magdalena Muszyńska; Jemma Hillyer, Oxford University Press; Ruyang Chloe Ye, Oxford University Press; Jasmin Anderson, Oxford University Press; Megan Hurley, Oxford University Press; Phil Davis, Oxford University Press. Finally, we would like to thank all 39 postgraduate participants and organisers of the 2025 University of Birmingham “PGT SHAPE AI Challenge” for their hard work and valuable contributions in helping us iterate on early versions of L2-Bench, in particular the winners and runner-ups of the challenge: Venkata Vyjayanthi Pedapati (Vy), Yernur Niyetkaliyev, Aparajitha Magnesh, Manh Nguyen (Leo), Niamh Evans, Hsin-Yun Ho (Sydney), Sofia Muñoz, Saniya Saheer, Taiki Shimosakai, Yang Yu, and Dr. Liam Knight for his help facilitating this pilot.

All substantive research decisions, interpretations, and conclusions remain solely the responsibility of the authors.

References

- Alcaras, G.; and Ricci, D. 2025. Configuration Work: Four Consequences of LLMs-in-use. arXiv:2512.19189.
- Bastani, H.; Bastani, O.; Sungu, A.; Ge, H.; Kabakci, O.; and Mariman, R. 2024. Generative AI Can Harm Learning. SSRN Working Paper, DOI: 10.2139/ssrn.4895486.
- Cuban, L. 2001. *Oversold and Underused: Computers in the Classroom*. Cambridge, MA: Harvard University Press.
- Eriksson, M.; Purificato, E.; Noroozian, A.; Vinagre, J.; Chaslot, G.; Gomez, E.; and Fernandez-Llorca, D. 2025. Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation. arXiv:2502.06559.
- Holmes, W. 2024. AIED—Coming of Age? *International Journal of Artificial Intelligence in Education*, 34(1): 1–11.
- Jurenka, I.; Kunesch, M.; McKee, K. R.; Gillick, D.; et al. 2024. Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach. ArXiv:2407.12687.
- Kennedy, W. M.; and Campos, D. V. 2025. Vernacularizing Taxonomies of Harm is Essential for Operationalizing Holistic AI Safety. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '24, 698–710. AAAI Press.
- Reuel, A.; Ghosh, A.; Chim, J.; Tran, A.; Long, Y.; Mickel, J.; Gohar, U.; Yadav, S.; Ammanamanchi, P. S.; Allaham, M.; Rahmani, H. A.; Akhtar, M.; Friedrich, F.; Scholz, R.; Riegler, M. A.; Batzner, J.; Habba, E.; Saxena, A.; Kornilova, A.; Wei, K.; Soni, P.; Mathew, Y.; Klyman, K.; Sania, J.; Sahoo, S.; Bruvik, O. B.; Sadeghi, P.; Goswami, S.; Wang, A.; Jernite, Y.; Talat, Z.; Biderman, S.; Kochenderfer, M.; Koyejo, S.; and Solaiman, I. 2025. Who Evaluates AI’s Social Impacts? Mapping Coverage and Gaps in First and Third Party Evaluations. arXiv:2511.05613.
- Schwartz, R.; Chowdhury, R.; Kundu, A.; Frase, H.; Fadaee, M.; David, T.; Waters, G.; Taik, A.; Briggs, M.; Hall, P.; Jain, S.; Yee, K.; Thomas, S.; Bhandari, S.; Duncan, P.; Thompson, A.; Carlyle, M.; Lu, Q.; Holmes, M.; and Skeadas, T. 2025. Reality Check: A New Evaluation Ecosystem Is Necessary to Understand AI’s Real World Effects. arXiv:2505.18893.
- Selwyn, N. 2012. *Education and Technology: Key Issues and Debates*. United Kingdom: Continuum International Publishing Group, 1st edition. ISBN 9781441150363.
- Selwyn, N. 2019. *Should Robots Replace Teachers?: AI and the Future of Education*. London: Polity.
- Weidinger, L.; Raji, I. D.; Wallach, H.; Mitchell, M.; Wang, A.; Salaudeen, O.; Bommasani, R.; Ganguli, D.; Koyejo, S.; and Isaac, W. 2025. Toward an Evaluation Science for Generative AI Systems. arXiv:2503.05336.
- Zawacki-Richter, O.; Marín, V. I.; Bond, M.; et al. 2019. Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(39).