Automatic Evaluation of the Pedagogical Effectiveness of Open-Domain Chatbots in a Language-Learning Game

Bianca Ciobanica[1,2][0009−0000−9028−0912], Cosupervisor: Serge Bibauw[3][0000−0002−1264−6090], and Supervisor: Anaïs Tack[1,2][0000−0003−3086−8188]

[1] KU Leuven, Faculty of Arts, Research Unit Linguistics, Leuven, Belgium
[2] KU Leuven, imec research group itec, Kortrijk, Belgium
{bianca.ciobanica, anais.tack}@kuleuven.be
[3] Université catholique de Louvain, Louvain-la-Neuve, Belgium
serge.bibauw@uclouvain.be

Abstract. This master's thesis introduces SEED (Scoring Educational Effectiveness of a Dialogue), a novel metric and approach to evaluating the pedagogical value of responses generated by large language models (LLMs) in a language-learning conversational game. As LLM-powered chatbots are being deployed more widely as educational tools (Ji, Han, & Ko, 2022; Kochmar et al., 2025), their design requires thorough evaluation to ensure not only their effectiveness, but also their reliability, given their social impact and role as interactive agents (Xu, Chen, & Huang, 2022). This setting makes it difficult to apply traditional rule-based evaluation metrics effectively, and while recent advancements in neural-based dialogue evaluation have improved assessment methods (Maurya et al., 2025), there remains no clear consensus on automatic evaluation of dialogue data (Yeh, Eskenazi, & Mehri, 2021).

Therefore, assessing the quality of open-ended systems in educational settings remains a non-trivial task. It spans linguistic proficiency, conversational, social, and pedagogical skills, along with broader goals such as effectiveness and user satisfaction (Ji, Han, & Ko, 2022). In dialogue designed to support second language development with open dialogue flows, overall dialogue metrics capture only part of the picture as they do not take into account the learning experience. We need evaluation methods that not only reflect how language is performed, but also how it is acquired by accounting for pedagogical factors necessary for language acquisition, such as input exposure, output production, and negotiation of meaning (Loewen & Sato, 2018). Without a robust evaluation framework, it remains unclear whether current systems successfully fulfill these didactic functions. Moreover, in the context of evaluating a conversational language-learning game, chatbot assessment does not occur in a traditional mentor–learner classroom setting, but rather within a character–player learning experience.

To address this challenge, this thesis investigates the following research question: how do we evaluate the pedagogical effectiveness of a dialogue where the learning goals are not explicitly stated but rather embedded. Evaluating such a system adds an entirely new layer; it requires an approach that is sensitive to function, effect and takes into account the open-ended and multi-dimensional nature of those conversations (Kochmar et al., 2025; Maurya et al., 2025; Tack & Piech, 2022).

The empirical foundation of this study is provided by *Language Hero* (LH; Linguineo, 2024), a task-based conversational language learning game that diverges from traditional the teacher-student. Within the game, the chatbot functions as a narrative character, while the learner assumes the role of a player. Language acquisition is thus facilitated through open-ended dialogue embedded in the game's storyline.

We proceeded in three stages: annotation framework design, metric development, and model implementation. First, we introduced a turn-level annotation framework grounded in an interactionist second language acquisition theory (Mackey, 2020). It focuses on three core dimensions: communicative intent, learner output, and interactional support. These dimensions were operationalized through a manually annotated corpus of in-game dialogues from LH.

Subsequently, the annotations guide SEED, a metric aggregating the learning potential at a macro-level. It offers a lens through which to evaluate open-ended dialogue.

In the last stage, the evaluation was framed as a supervised classification task, using fine-tuned BERT models (Devlin et al., 2019) and their variants (Liu et al., 2019; Pires, Schlinger, & Garrette, 2019; Sanh et al., 2020) to develop the learnable metric. We implemented task-specific and unified architectures using a hybrid input encoding strategy, either including or excluding conversational context. The results show that the inclusion of context affects each pedagogical dimension differently. DistilBERT demonstrated the highest performance (F1 = 0.84) in predicting communicative intent when conversational context was included in the input. In contrast, output elicitation was most accurately predicted by RoBERTa when context was excluded (F1 = 0.98). Overall, predicting interactional support was most effective with BERT in a unified, context-aware architecture (F1 = 0.81), suggesting that shared representations across pedagogical dimensions enhance model performance.

Future work will validate the SEED metric with end users to assess effectiveness and reliability. SEED-informed scores could enable cross-dialogue evaluation, dynamic prompt adaptation, and serve as a baseline for semi-supervised learning. We also plan to compare fine-tuned BERT with few-shot LLM prompting in a hybrid model that combines their strengths (Wang & Chen, 2025).

References

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 4171–4186). Association for Computational Linguistics.

Ji, H., Han, I., & Ko, Y. (2022). A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education*, *55*, 1–16.

Kochmar, E., Alhafni, B., Bexte, M., Burstein, J., Horbach, A., Laarmann-Quante, R., Tack, A., Yaneva, V., & Yuan, Z. (2025). The 2025 bea shared task on pedagogical ability assessment of ai-powered tutors [Held July 31–August 1, 2025]. *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*. https://sig-edu.org/sharedtask/2025

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach [arXiv:1907.11692 [cs]]. *arXiv*.

Loewen, S., & Sato, M. (2018). Interaction and instructed second language acquisition. *Language Teaching*, *51*(3), 285–329.

Mackey, A. (2020). Theory and approaches in research into interaction, corrective feedback, and tasks in l2 learning. In *Interaction, feedback and task research in second language learning: Methods and design* (pp. 1–26). Cambridge University Press.

Maurya, K. K., Srivatsa, K. A., Petukhova, K., & Kochmar, E. (2025). Unifying AI Tutor Evaluation: An Evaluation Taxonomy for Pedagogical Ability Assessment of LLM-Powered AI Tutors [arXiv:2412.09416 [cs]]. *arXiv* Comment: 9 pages.

Pires, T., Schlinger, E., & Garrette, D. (2019, July). How multilingual is multilingual BERT? In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4996–5001). Association for Computational Linguistics.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter [arXiv:1910.01108 [cs]]. *arXiv*
Comment: February 2020 - Revision: fix bug in evaluation metrics, updated metrics, argumentation unchanged. 5 pages, 1 figure, 4 tables. Accepted at the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019.

Tack, A., & Piech, C. (2022, July). The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues. In A. Mitrovic & N. Bosch (Eds.), *Proceedings of the 15th international conference on educational data mining* (pp. 522–529). International Educational Data Mining Society.

Wang, D., & Chen, G. (2025). Evaluating the use of BERT and Llama to analyse classroom dialogue for teachers' learning of dialogic pedagogy [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13604]. *British Journal of Educational Technology*, *n/a*(n/a).

Xu, K., Chen, X., & Huang, L. (2022). Deep mind in social responses to technologies: A new approach to explaining the Computers are Social Actors phenomena. *Computers in Human Behavior*, *134*, 107321.

Yeh, Y.-T., Eskenazi, M., & Mehri, S. (2021, November). A comprehensive assessment of dialog evaluation metrics. In W. Wei, B. Dai, T. Zhao, L. Li, D. Yang, Y.-N. Chen, Y.-L. Boureau, A. Celikyilmaz, A. Geramifard, A. Ahuja, & H. Jiang (Eds.), *The first workshop on evaluations and assessments of neural conversation systems* (pp. 15–33). Association for Computational Linguistics.