SimuLe-Lux: A Neuro-Symbolic Teaching System for Diagnosing L2 Reading Comprehension

Christophe Friezas Gonçalves and **Christoph Schommer** and **Salima Lamsiyah**
Department of Computer Science, Faculty of Science, Technology and Medicine
University of Luxembourg, Esch-sur-Alzette, Luxembourg
{christophe.friezas,christoph.schommer,salima.lamsiyah}@uni.lu

Abstract

Understanding why second-language learners make reading comprehension errors is essential for effective instruction and teaching. However, providing fine-grained diagnostic feedback requires teacher expertise and time. While recent work has explored large language models as student simulators, existing approaches lack grounding in linguistic and educational evidence. We introduce *SimuLe-Lux*, a neuro-symbolic framework for simulating learner profiles and supporting teacher-facing diagnostic training in Luxembourgish reading comprehension. The system is built on *LuxDiagRC* and utilizes a large language model to create and simulate learner responses across six CEFR proficiency levels, including weaker-state variants. These responses are aggregated into interpretable diagnostic profiles. A retrieval-augmented teacher-facing agent then uses this structured knowledge base to generate grounded explanations of learner behaviour and instructional guidance. Evaluation results show that the simulated learners exhibit CEFR-consistent error patterns and that the teacher-facing agent produces relevant and informative diagnostic feedback for structured pedagogical queries.

1 Introduction

Learning to read in a second language inevitably involves making mistakes. Such errors are not merely failures of performance but indicators of underlying linguistic and cognitive difficulties. When teachers can identify and interpret these errors, they can better target instruction and support language development (Mullis and Martin, 2019; Suzuki et al., 2025). However, diagnosing the sources of reading comprehension errors requires substantial expertise and individual attention, which is difficult to provide at scale in classroom settings and for low-resource languages.

At the same time, large language models (LLMs) are increasingly used by students as learning aids (Ghimire and Edwards, 2024; Liu et al., 2025), and their widespread interaction with learner-generated text has opened new possibilities for modeling student behaviour. Recent studies have shown that LLMs can be prompted to simulate student responses in test and exam scenarios (Martynova et al., 2025; Benedetto et al., 2024; Xu et al., 2024), suggesting their potential as tools for generating synthetic learner data. Teachers also report positive attitudes toward AI-based educational support systems (Pokrivcakova, 2019). However, prior work has identified important limitations, including unrealistic response patterns in open-ended settings (Martynova et al., 2025) and poor calibration of learner proficiency levels (Benedetto et al., 2024; Xu et al., 2024). These shortcomings make it difficult to use existing LLM-based simulations for fine-grained diagnostic training.

In this paper, we propose *SimuLe-Lux*, a neuro-symbolic system designed to support teacher-facing diagnostic reasoning in reading comprehension. The system combines a diagnostically annotated multiple-choice reading comprehension dataset in Luxembourgishœ with large language model–based learner simulation across six CEFR proficiency levels. Rather than producing arbitrary errors, *SimuLe-Lux* generates learner responses that are grounded in annotated linguistic features and cognitively challenging question types, enabling structured analysis of why learners commit mistakes.

SimuLe-Lux is intended as a training partner for educators. It allows teachers to explore simulated learner profiles, examine patterns of misunderstanding, and receive explanations that link learner behaviour to specific linguistic and cognitive triggers. This design shifts the role of AI in education from task automation toward support for teacher diagnos-

Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages, pages 191–198
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_021

tic expertise and reflective practice (Pruski et al., 2025).

Through a series of targeted evaluations, we show that a carefully constructed diagnostically rich dataset can serve as the symbolic core of a neuro-symbolic system, enabling both realistic learner simulation and grounded teacher-facing explanations. Our work demonstrates how dataset design, learner modeling, and retrieval-augmented reasoning can be combined to support human-centered educational AI in a low-resource language setting.

Contributions. This work makes two main contributions. First, we present *SimuLe-Lux*, a teacher-facing system that simulates second-language learner behaviour using linguistically and cognitively grounded annotations. Second, we propose a neuro-symbolic diagnostic control mechanism that conditions simulated learner errors on annotated linguistic features and question types, enabling interpretable and controllable learner profiles.

2 Related Work

This section places our work in relation to three research areas: reading comprehension datasets, diagnostic modeling of learner weaknesses, and teacher-facing educational systems.

Reading Comprehension Datasets. Datasets are a cornerstone for the development and evaluation of NLP systems. Large-scale benchmarks such as SQuAD (Rajpurkar et al., 2016) and RACE (Lai et al., 2017) have demonstrated strong performance in multiple-choice question answering and generation (Dutulescu et al., 2025). Domain-specific datasets such as MedMCQA (Pal et al., 2022) and SciDQA (Singh et al., 2024) extend this paradigm to medical and scientific texts. However, despite their scale and diversity, most of these datasets frame reading comprehension as a binary task, scoring responses simply as correct or incorrect. As a result, they provide little insight into the underlying causes of learner errors, which is a critical limitation in educational settings where diagnosing misconceptions is central to language acquisition and instruction (Berzak et al., 2020).

Student Weaknesses Detection. The detection of weaknesses in reading comprehension is a multifaceted problem that traditionally requires substantial teacher expertise and extensive one-on-one instructional time (Mullis and Martin, 2019; Suzuki et al., 2025). Diagnostic models analyse patterns in student responses to identify learner profiles and categories of misunderstanding (Hemati and Baghaei, 2020; Suzuki et al., 2025). While these approaches demonstrate the feasibility of AI-based weakness detection, they typically rely on specialized models, handcrafted diagnostic rules, and fixed category sets, which limit their adaptability across learners and domains. In contrast, our framework introduces a data-driven diagnostic controller that models emerging learner patterns and profiles from question responses, enabling flexible and dynamically evolving representations using state-of-the-art large language models.

Teacher-Facing Tools. LLM-based educational tools span a wide range of applications, from simple translation systems to intelligent tutoring systems (ITS) (Pokrivcakova, 2019). Recent research on teacher-facing AI has largely focused on automating instructional tasks such as curriculum design, content generation, and grading (Zhang et al., 2025; Hu et al., 2025; Gonnermann-Müller et al., 2025). While this increases efficiency, it risks encouraging forms of work substitution similar to those observed among students who over-rely on generative AI (Liu et al., 2025). In contrast, emerging work advocates for AI systems that support teacher skill development and reflective practice, enabling lifelong learning and pedagogical growth (Pruski et al., 2025).

Positioning of Our Work. *SimuLe-Lux* combines a diagnostically annotated reading comprehension dataset, learner simulation, and a teacher-facing analysis agent within a single neuro-symbolic framework. Unlike rule-based diagnostic systems, it derives learner profiles from simulated responses grounded in linguistic annotations. Compared to standard reading comprehension datasets that support only correct–incorrect scoring, the use of *LuxDiagRC* (Gonçalves et al., 2026) enables analysis of the sources of learner errors. The system is designed to support teacher diagnostic reasoning rather than automate instructional tasks.

3 Methodology

This section presents the methodological design of *SimuLe-Lux* , a neuro-symbolic framework for simulating second-language learners and supporting teacher-facing diagnostic training in reading comprehension. As illustrated in Figure 1, the

Table 1: Example *Retrieve* question from LuxDiagRC showing how answer options are grounded in annotated text spans and linguistic feature tags.

Question	Wéi schwätzt den Marc wann en rosen ass? *How does Marc talk when he is angry?*
Span Type	**Span**
Critical span	. . . seet de <ORTHO-PHONO-DIVERGE> <MORPH-N-RULE> Marc zu mir, wann hien <ORTHO-PHONO-DIVERGE> <MORPH-N-RULE> besonnesch rosen <LEX-FALSE-FRIEND> ass <LEX-FALSE-FRIEND>, schwätzt hien <ORTHO-PHONO-DIVERGE> <MORPH-N-RULE> franséisch, fir seng Roserei besser an <LEX-FALSE-FRIEND> der <LEX-FALSE-FRIEND> Gidd ze halen, seet en <DISC-COREF-AMBIG>.
Distractor span	wéi <ORTHO-PHONO-DIVERGE> wier et, wa <ORTHO-PHONO-DIVERGE> <MORPH-N-RULE> mir géifen erwuesse <ORTHO-PHONO-DIVERGE> ginn <SYN-VERB-FINAL>?!
Answer Type	**Answer**
Correct	En schwätzt Franséisch *He speaks in French*
Misunderstanding	En schwätzt séier fir sech an der Gidd ze halen *He speaks fast to keep his temper in check*
Distractor	En schwätzt wéi en Erwuessenen *He speaks like an adult*
No support	En schwätzt Däitsch *He speaks in German*

framework integrates three tightly coupled components: (i) *LuxDiagRC*, a diagnostically annotated reading comprehension dataset (Gonçalves et al., 2026), (ii) a learner simulation module with neuro-symbolic diagnostic control, and (iii) a retrieval-augmented teacher-facing rationalization agent. Together, these components form a pipeline that transforms annotated texts into simulated learner behaviours and grounded pedagogical explanations.

3.1 LuxDiagRC Dataset

The *LuxDiagRC* dataset is constructed using a two-layer human annotation paradigm designed to link reading comprehension errors to underlying linguistic and cognitive difficulties. Each answer option in the dataset is explicitly grounded in annotated spans of the source text, enabling diagnostic interpretation of learner choices (Gonçalves et al., 2026). The dataset applies their proposed annotation pipeline to 16 texts used in primary and secondary language teaching (Gonçalves et al., 2026).

The dataset contains 640 multiple-choice questions across 16 texts (40 per text, 10 per annotator), with each text comprising 8 *Retrieve*, 12 *Interpretation*, 12 *Inferential*, and 8 *Evaluative* items. Each answer option is explicitly grounded in annotated text spans, enabling direct mapping from learner choices to underlying linguistic and cognitive difficulties, as presented in Table 1.

3.2 Simulated Students and Neuro-symbolic Diagnostic Control

The neuro-symbolic diagnostic controller requires stable and interpretable probability vectors to represent learner profiles. To obtain these, we employ a large language model (GPT-5, August 2025) to simulate student responses across the six proficiency levels defined by the Common European Framework of Reference for Languages (CEFR): A1 to C2.

The model is prompted in a zero-shot setting to answer all questions in the dataset. It is instructed to respond as a student taking a language test and to produce realistic, level-appropriate errors. For each proficiency level, the model answers the questions twice: once at the nominal level and once in a deliberately weaker state. This weaker-state variant enables the diagnostic controller to model smooth transitions between adjacent proficiency levels, avoiding hard boundaries and allowing both weaker and stronger learner behaviours to be represented within each level.

Questions are grouped by text, with 40 questions per prompt, ensuring that the model's errors arise from reading comprehension difficulty within a single text rather than from cross-text interference or memory effects. This design emphasizes the emergence of systematic error patterns linked to linguistic and cognitive challenges rather than to recall alone.

The resulting simulated responses are aggregated to compute probability distributions over linguistic features, question types, and answer options for each proficiency level, yielding a diagnostic probability vector for each learner profile. In addition to these level-specific vectors, level-text-specific distributions are also computed, providing additional layers of diagnostic and contextual information in the knowledge base.

3.3 Teacher-Facing Diagnostic Agent

The teacher-facing component is implemented as a retrieval-augmented generation (RAG) system (Lewis et al., 2020) built on GPT-5 (August 2025), which integrates the *LuxDiagRC* dataset with the neuro-symbolic diagnostic controller to produce grounded explanations of learner behaviour. Through retrieval over the full knowledge base, the agent can link student responses to underlying linguistic features, cognitive question types, and proficiency-level vulnerability profiles. The model operates in a zero-shot setting and is instructed to assist teachers in analysing student errors, identifying the linguistic and cognitive sources of misunderstanding, and reflecting on instructional strategies for Luxembourgish second-language learners. By combining symbolic diagnostic representations with generative reasoning,

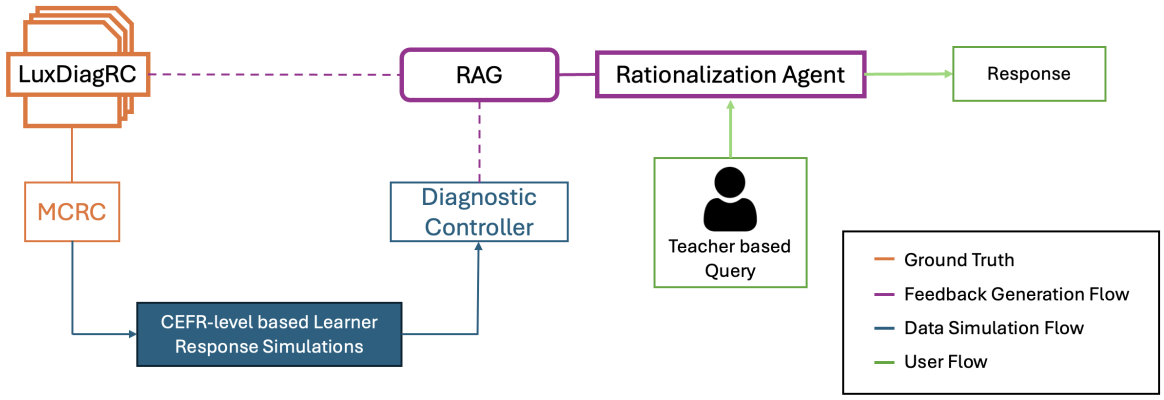

Figure 1: Overview of the *SimuLe-Lux* framework. *LuxDiagRC*) texts with the diagnostically rich multiple-choice reading comprehension questions (MCRC) are used to simulate CEFR-level learner responses. These responses are aggregated into neuro-symbolic diagnostic profiles and accessed by a retrieval-augmented teacher-facing agent to provide grounded explanations and pedagogical feedback.

the agent produces explanations that are both pedagogically meaningful and directly traceable to evidence in the dataset.

4 SimuLe-Lux Framework Evaluation

This section evaluates *SimuLe-Lux* as a multi-agent, teacher-facing framework for diagnostic reading comprehension training. In line with the methodology described in Section 3, the evaluation targets the two core computational components of the system: (i) the learner simulation and neuro-symbolic diagnostic controller, and (ii) the retrieval-augmented teacher-facing agent. Here we focus on whether the simulated learners exhibit CEFR-consistent behaviour and whether the teacher-facing agent produces grounded, pedagogically useful responses.

4.1 Evaluation of Learner Simulation

The learner simulation is evaluated with respect to its alignment with CEFR proficiency levels and the smoothness of transitions between adjacent learner profiles. After generating responses for all texts and proficiency levels, error patterns are analysed per question type and level to assess whether the neuro-symbolic diagnostic controller captures meaningful differences in reading comprehension ability.

4.1.1 Prompt Engineering and Simulation Protocol

The learner simulation underwent a systematic prompt engineering process. Initial prompts produced nearly identical response patterns across proficiency levels. This was resolved by including all CEFR levels (A1–C2) within a single prompt, enabling the model to differentiate between learner profiles within a shared context window. A minimum-correct-answer constraint was further introduced for the weaker-state simulations to avoid degenerate cases with unrealistically poor performance.

Table 2: Diagnostic controller probabilities derived from simulated learner responses. $_W$ denotes the weaker-state profiles. Each value represents the probability of committing an error for the corresponding question type.

Proficiency	Retrieve	Interpret	Infer	Evaluate
$A1_W$	0.77	0.81	0.79	0.76
A1	0.63	0.70	0.67	0.64
$A2_W$	0.68	0.79	0.73	0.78
A2	0.55	0.65	0.57	0.60
$B1_W$	0.66	0.77	0.66	0.72
B1	0.54	0.57	0.49	0.50
$B2_W$	0.64	0.64	0.62	0.65
B2	0.49	0.47	0.45	0.43
$C1_W$	0.59	0.61	0.58	0.55
C1	0.44	0.44	0.42	0.40
$C2_W$	0.52	0.51	0.51	0.50
C2	0.42	0.41	0.42	0.38

The final prompt was held fixed for all texts and question batches. It instructs the model to answer each question twice per proficiency level, once at the nominal level and once in a deliberately weaker state, with responses restricted to multiple-choice labels (A–D), ensuring consistency across runs:

- Beäntwer des Froen als géifs du als Schüler am Sproochentest setzen. Lëtzebuergesch

ass net deng Mammesprooch. Du mess realistesch an typesch Examensfeeler fir dei spezifesch Sproochen Niveauen (A1,A2,B1,B2,C1,C2). Äntwer simpel just als A, B, C oder D fir all Fro an fir all Sprochen Niveau. Pro Sproochen Niveau Beäntwer déi 40 Froën eng keier normal an nach eng kéier mä méi schwaach mat minimum 8 richtechen Äntwerten.

- Answer these questions as if you were a student in a language test. Luxembourgish is not your native language. You make realistic and typical exam mistakes for your specified language level (A1,A2,B1,B2,C1,C2). Answer only with A, B, C, or D for each question and for all language levels. For each language level, answer the 40 questions once normally and a second time in a weaker state with a minimum of 8 correct answers.

4.1.2 Neuro-Symbolic Diagnostic Controller

The aggregated error probabilities produced by the simulated learners are shown in Table 2. Each value represents the probability of committing an error for a given question type and proficiency level, with $_W$ denoting the weaker-state profiles. The results exhibit a CEFR-consistent trend: error rates decrease monotonically as proficiency increases across all question types.

Importantly, weaker-state profiles (e.g., $B2_W$) consistently fall between their nominal level and the adjacent lower level (e.g., B1), producing smooth transitions between learner profiles. This reflects a well-established property of second-language learning, where the strongest learners at a lower level may outperform the weakest learners at a higher level. Without weaker-state modeling, such transitional behaviour would be lost under hard level boundaries.

4.2 Evaluation of the Teacher-Facing Diagnostic Agent

The teacher-facing agent is evaluated on its ability to generate grounded, informative, and pedagogically appropriate responses to teacher queries. A total of 40 prompts were used, grouped into four categories: dataset-specific, proficiency-level–specific, teaching-skill–oriented, and vague. Representative prompts for each category are shown in Table 3.

Table 3: Sample prompt categories used to evaluate the teacher-facing agent. Prompts were originally in Luxembourgish and are translated here for clarity.

Prompt category	Example prompt
Dataset	A student performed very well on the "Blues" questions but struggled with the "D'Schnurreli" questions. Why might this be?
Proficiency level	Which difficulties does a B2 student encounter while reading?
Teaching skill	How can I improve the inferential skills of my weaker students?
Vague	How can I help strong students?

Table 4: Mean human evaluation scores for the teacher-facing agent across prompt categories. Scores are averaged over three judges on three categories with a five-point scale. Relevance (Rele.), Informativeness (Info.), Appropriateness (App.).

Prompt type	Rele.	Info.	Appr.
Dataset	4.43	3.92	3.97
Proficiency level	4.87	4.58	4.73
Teaching skill	4.27	4.37	4.77
Vague	3.10	2.60	3.83

4.2.1 Prompt Categories and Human Scoring

Dataset-specific prompts test whether the agent can retrieve and reason over LuxDiagRC content. Proficiency-level prompts assess its ability to connect learner profiles to expected weaknesses. Teaching-skill prompts evaluate whether the agent can support instructional reflection and diagnostic reasoning. Vague prompts test robustness to underspecified input.

All responses were scored by three independent human judges on a five-point scale along three rubrics: *Relevance* (use of dataset and diagnostic evidence), *Informativeness* (usefulness for teachers), and *Appropriateness* (pedagogical coherence). Mean scores for each prompt category are reported in Table 4.

4.2.2 Quantitative Results

As shown in Table 4, the agent achieves strong performance on dataset-specific, proficiency-level, and teaching-skill prompts, indicating that it effectively integrates the diagnostic controller with *LuxDiagRC* to generate grounded pedagogical explanations. Performance drops for vague prompts, particularly in relevance and informativeness, reflecting reduced access to dataset evidence and increased reliance on the model's internal knowledge.

4.2.3 Qualitative Analysis

Qualitative inspection of responses supports these findings. For teaching-skill prompts (Table 3), the agent explicitly references simulated learner performance and linguistic features when proposing instructional strategies, for example:

- *Start with texts where your learners already did better on interpret/retrieve and add one inferential item with strong evidence cues. Your own results show retrieve > inferential on Den Ersatzschoulmeeschter at lower levels; build up from there and explicitly neutralize false-friend traps in distractors during feedback.*

In contrast, for vague prompts the agent relies largely on generic pedagogical knowledge rather than diagnostic evidence, as illustrated by:

- *Compare across texts. Set quick synthesis tasks (e.g., compare how two texts set up a problem/solution or portray relationships) to push transfer of skills.*

- *Feedback that grows expertise. Use concise rubrics that reward evidence citation, metalinguistic precision, and quality of reasoning (not just speed or quantity). Ask students to annotate where their inference "leaps" occur using the text's key span labels.*

This contrast highlights the importance of *LuxDiagRC* and the neuro-symbolic diagnostic controller in enabling evidence-based teacher support.

5 Concrete Example: End-to-End SimuLe-Lux Workflow

To support understanding of *SimuLe-Lux's* operation, we present an end-to-end example following a LuxDiagRC input through to the diagnostic agent's final response, illustrating how the system's components interact throughout the process.

Input (LuxDiagRC) A Luxembourgish text annotated with linguistic features is used to generate a multiple-choice question for LuxDiagRC:

Question:

> *'Wisou hun d'Kanner forcéiert gelaacht?'* (*Why did the kids force a laughter?*)

Critical span (annotated):

> *Vun eis* `<LEX-FALSE-FRIEND>` *krut hien no all Witz* `<LEX-LOAN-DE>` *e* `<ORTHO-PHONO-DIVERGE>` `<MORPH-N-RULE>` *forcéiert* `<LEX-LOAN-FR>` *Laachen ze héieren* `<SYN-VERB-FINAL>`

Distractor span (annotated):

> *Seng Pedanterie* `<LEX-RARE-ZLS>` `<LEX-LOAN-FR>` *ass* `<LEX-FALSE-FRIEND>` *wait gang* `<SYN-VERB-FINAL>`

Answer options:

- A) *Vir dass se op dem Proff senger gudder seit stinn.* (Correct)
- B) *Well se wollten laachen.* (Misunderstanding)
- C) *Well den Prof pedant war.* (Distractor span)
- D) *Well d'Kanner den Witz gutt fannen.* (No textual support)

Simulated Learner Output (LLM Student Simulation). SimuLe-Lux generates CEFR-level learner responses:

Level	Answer	
	Normal	Weaker
A1	B	D
A2	B	D
B1	C	D
B2	A	B
C1	A	C
C2	A	A

The weaker-state simulation (e.g., $B2_W$) may still select either option B or C, allowing overlap between A2, B1 and B2. Thus, producing realistic transitional learner profiles.

Teacher Query.

> *"Why do my B2 learners often misunderstand this question?"*

Teacher-Facing Output (Rationalization Agent).

> *"B2 Learners often misread the LEX-LOAN-FR item forcéiert and map it to a volitional "wanted to laugh," so they pick the agentive answer "Well se wollten laachen" instead of the Inferential,*

social motive, which the critical span signals via "forcéiert Laachen". The Wisou prompt demands Inferential reasoning beyond surface form, but attention is drawn to clause-final structure (SYN-VERB-FINAL), so form-focused parsing overrides the causal reading and yields the misunderstanding."

This example demonstrates how *LuxDiagRC*, learner simulation, and the diagnostic controller are combined through retrieval-augmented generation to produce grounded, pedagogically meaningful teacher feedback.

6 Conclusion

This paper presents a novel paradigm for teacher-facing educational tools that addresses key gaps in the literature on reading comprehension datasets, learner simulation, and diagnostic AI systems. The proposed framework combines a dual-layer annotation pipeline with large language model–based learner simulation to construct a diagnostically grounded dataset and a retrieval-augmented teacher-facing agent capable of supporting pedagogical reasoning.

The results support three main conclusions. First, the evaluation of the teacher-facing agent shows that combining simulated learner data with a diagnostically rich dataset enables the generation of grounded and pedagogically meaningful feedback. Second, the learner simulation results demonstrate that large language models can be used to generate realistic and proficiency-sensitive student behaviour, strengthening their role as tools for learner modeling. Third, the inclusion of intermediate (weaker-state) proficiency profiles enhances the neuro-symbolic diagnostic controller by capturing smooth transitions and realistic overlaps between CEFR levels, allowing the system to model transitional learner weaknesses more faithfully.

Limitations and Future Work

This work stands as a proof of concept for the two main contributions of the paper, the *SimuLe-Lux* framework and the implementation of a neuro-symbolic diagnostic controller. However, several limitations remain. First, although *LuxDiagRC* is carefully designed to be diagnostically rich, it is relatively small in scale compared to large reading comprehension benchmarks. Its primary purpose is not broad coverage but the construction of a high-quality gold standard for diagnostic analysis. While this enables fine-grained grounding of learner errors, it limits the diversity of genres and topics.

Second, learner behaviour is generated using a single large language model. Although the CEFR-consistent patterns observed in Table 2 indicate that the simulation produces meaningful proficiency distinctions, the simulated learners may still reflect model-specific biases and stylistic regularities. In particular, some error patterns may be influenced by the internal representations of the underlying language model rather than by genuine second-language acquisition processes.

Third, the teacher-facing rationalization agent is evaluated using human judgments on synthetic prompts rather than through deployment with real teachers. While the results in Table 4 demonstrate strong grounding and pedagogical coherence for structured queries, the system's effectiveness in authentic classroom workflows and teacher training contexts remains to be validated.

Future work will address these limitations along three directions. First, *LuxDiagRC* will be expanded through controlled data synthesis, using the existing annotation framework to generate new diagnostically grounded texts and questions, thereby increasing coverage while preserving interpretability. Second, simulated learner profiles will be compared against real student response data to calibrate and validate the neuro-symbolic diagnostic controller, enabling more accurate modeling of second-language error distributions. Third, the teacher-facing agent will be evaluated in longitudinal user studies with educators, allowing the impact of *SimuLe-Lux* on diagnostic skill development and instructional decision-making to be assessed in realistic educational settings.

References

Luca Benedetto, Giovanni Aradelli, Antonia Donvito, Alberto Lucchetti, Andrea Cappelli, and Paula Buttery. 2024. Using LLMs to simulate students' responses to exam questions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11351–11368, Miami, Florida, USA. Association for Computational Linguistics.

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Lin-*

guistics, pages 5726–5735, Online. Association for Computational Linguistics.

Andreea Dutulescu, Stefan Ruseti, Denis Iorga, Mihai Dascalu, and Danielle S. McNamara. 2025. Ymcq: Reasoning-enhanced mcq generation. In *Artificial Intelligence in Education*, pages 308–315, Cham. Springer Nature Switzerland.

Aashish Ghimire and John Edwards. 2024. *Coding with AI: How Are Tools Like ChatGPT Being Used by Students in Foundational Programming Courses*, pages 259–267.

Christophe Friezas Gonçalves, Salima Lamsiyah, and Christoph Schommer. 2026. LuxDiagRC: A diagnostic reading comprehension corpus for Luxembourgish with linguistic and cognitive annotation layers. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 532–541, Rabat, Morocco. Association for Computational Linguistics.

Jana Gonnermann-Müller, Jennifer Haase, Konstantin Fackeldey, and Sebastian Pokutta. 2025. Facet: Teacher-centred llm-based multi-agent systems-towards personalized educational worksheets.

Seyed Jamal Hemati and Purya Baghaei. 2020. A cognitive diagnostic modeling analysis of the english reading comprehension section of the iranian national university entrance examination. *International Journal of Language Testing*, 10(1):11–32.

Bihao Hu, Jiayi Zhu, Yiying Pei, and Xiaoqing Gu. 2025. Exploring the potential of llm to enhance teaching plans through teaching simulation. *npj Science of Learning*, 10(1):7.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chang Liu, Loc Hoang, Andrew Stolman, Rene F. Kizilcec, and Bo Wu. 2025. Understanding student engagement with large language model-powered course assistants. In *Artificial Intelligence in Education*, pages 3–10, Cham. Springer Nature Switzerland.

Daria Martynova, Jakub Macina, Nico Daheim, Nilay Yalcin, Xiaoyu Zhang, and Mrinmaya Sachan. 2025. Can llms effectively simulate human learners? teachers' insights from tutoring llm students. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 100–117.

Ina VS Mullis and Michael O Martin. 2019. *PIRLS 2021 Assessment Frameworks.* ERIC.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Silvia Pokrivcakova. 2019. Preparing teachers for the application of ai-powered technologies in foreign language education. *Journal of language and cultural education*, 7(3):135–153.

Cédric Pruski, Célia da Costa Pereira, Marcos Da Silveira, Gabriele Marconi, Marie Gallais, and Andrea Tettamanzi. 2025. Bridging skill gaps: Combining generative and symbolic ai for personalized lifelong learning pathways. In *Artificial Intelligence in Education*, pages 19–27, Cham. Springer Nature Switzerland.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Shruti Singh, Nandan Sarkar, and Arman Cohan. 2024. SciDQA: A deep reading comprehension dataset over scientific papers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20908–20923, Miami, Florida, USA. Association for Computational Linguistics.

Shungo Suzuki, Hiroaki Takatsu, Ryuki Matsuura, Miina Koyama, Mao Saeki, and Yoichi Matsuyama. 2025. Feedforwarding diagnostic language assessment: Artificial intelligence- (ai-) driven weakness identification and contextualised feedback for second language speaking. *Language Testing*, 42(4):476–507.

Songlin Xu, Xinyu Zhang, and Lianhui Qin. 2024. Eduagent: Generative student agents in learning.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2025. Simulating classroom education with LLM-empowered agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10364–10379, Albuquerque, New Mexico. Association for Computational Linguistics.