An AI-based Chat Agent for Measuring Students' Self-Regulated Learning Skills

Slaviša Radović
Center of Advanced Technology for Assisted Learning and Predictive Analytics (CATALPA), FernUniversität in Hagen
Hagen, Germany
slavisa.radovic@fernuni-hagen.de

Elisabeth Wetchy
FernUniversität in Hagen
Hagen, Germany
elisabeth.wetchy@gmail.com

Niels Seidel
Center of Advanced Technology for Assisted Learning and Predictive Analytics (CATALPA), FernUniversität in Hagen
Hagen, Germany
niels.seidel@fernuni-hagen.de

2025 International Conference on Education Technology and Computers (ICETC) | 979-8-3315-9791-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICETC66579.2025.11387673

***Abstract*—While the recent potential of Large Language Models (LLMs) has been studied across various domains in education, their application in measuring students' Self-Regulated Learning (SRL) skills remains underexplored. Current SRL measurement initiatives (surveys and digital trace data) face several challenges, directly impeding the development of effective SRL interventions. To address this complex educational challenge, this study examines the implementation and evaluation of a generative artificial intelligence agent, *AI-SRLSI*, designed to conduct interviews based on Zimmerman and Martinez-Pons's Self-Regulated Learning Structured Interview (SRLSI). The system was tested with a total of 13 participants to explore efficiency, effectiveness and satisfaction. The results of the study indicate that the agent can successfully conduct the SRLSI interview, as well as demonstrate efficient automation of SRL assessments. Learners found the tool user-friendly and appreciated the conversational accuracy and quality. However, feedback on the utility and relevance of the recommendations was mixed, underscoring areas for improvement in future iterations and the potential of *AI-SRLSI* to enhance personalized learning support. These results offering direct insights for future advancements in both, SRL measurement and SRL interventions.**

Keywords—AI, Self-Regulated Learning, LLM.

I. INTRODUCTION

Self-Regulated Learning (SRL) has emerged as a critical competency for students in higher education to effectively manage the complex demands of academic success [1, 2, 20]. SRL entails learners taking an active role in directing their learning processes, including setting goals, planning strategies, monitoring progress, and reflecting on outcomes. Prominent SRL frameworks, proposed by Winne and Hadwin, Pintrich, and Zimmerman, offer detailed insights into these processes [3]. This article draws specifically on Zimmerman's influential model [4], which outlines three phases: forethought (goal setting and planning), performance (strategy implementation and monitoring), and self-reflection (evaluating and adjusting). Zimmerman's model integrates cognitive, emotional, and behavioral regulation, providing a comprehensive lens for understanding and supporting SRL [3, 6, 7].

Effective students' SRL is strongly associated with academic success, resilience, motivation, and the development of transferable skills such as time management and critical thinking—competencies essential for lifelong learning and professional growth [2]. Significant efforts have been made in fostering SRL through advances in research, the implementation of educational frameworks, and the integration of innovative technologies [7, 8].

Accurately measuring SRL remains a significant challenge and an essential prerequisite for advancing research and practice in this area [7, 9, 10, 20]. Commonly used methods include self-report questionnaires, such as the MSLQ, SRLIS, and LASSI [9], as well as log data from digital learning platforms [8, 10, 12]. Each of these approaches has its limitations. Self-report questionnaires are cost-effective and easy to administer but are often susceptible to biases, such as social desirability, self-selection, and overestimation of one's self-regulation abilities. In contrast, log data can provide objective insights into learners' behaviors but often lack fail to capture the metacognitive and motivational dimensions of SRL [6, 8, 11].

A less commonly used but valuable approach involves think-aloud protocols and interview methods, which provide deeper insights into learners' cognitive and regulatory processes. These methods require students to verbalize their actions and thoughts during specific learning tasks without justifying or explaining their reasoning [3, 12]. While such techniques reveal nuanced aspects of SRL, they are resource-intensive and time-consuming. For instance, structured interview protocols like Zimmerman and Martinez-Pons's Self-Regulated Learning Structured Interview (SRLSI) [5] have seen limited use in recent years due to their labor-intensive nature. Practical challenges—such as the need for additional interviewer training, manually conducting interviews, and analyzing qualitative data—have led researchers to favor less demanding alternatives, such as mentioned self-report questionnaires and digital trace data [10].

This study revisits the potential of the structured interview protocol as a viable SRL measurement tool by exploring its automation through a generative AI agent (AI-SRLSI).

Recently, generative AI systems, such as ChatGPT, have demonstrated significant potential for addressing specific educational challenges, prompting growing interest from researchers and educators [13, 14]. These systems excel in powering interactive question-answering platforms and educational chatbots, offering meaningful communication with students [14, 15]. Moreover, their ability to adapt and customize learning experiences enables them to align with varied teaching strategies and accommodate individual learners' characteristics

and preferences. This was also recognized in Yan et al. [14] systematic scoping review of 118 peer-reviewed papers unveiling benefits such as detection of learning mistakes, grading students' knowledge, providing teaching support, tailoring knowledge representation, providing feedback and recommendation.

While LLMs hold significant potential for educational applications, their use is not without limitations [16]. Challenges such as hallucinations, overstatements, and the generation of false or misleading information are well-documented [13]. As a result, educators and researchers have become increasingly vigilant about the biases and inaccuracies in AI-generated content. These concerns have driven the development of more carefully designed, thoughtfully implemented, and contextually adapted AI solutions to minimize risks. For instance, a recent study by Kwartler et al. [16] highlights the use of multi-agent workflows, where AI models interact collaboratively to identify, capture, and correct hallucinations in generated content. Such nonlinear workflows are gaining traction for their ability to improve the robustness and accuracy of AI outputs, paving the way for safer and more reliable educational applications. Others recommend complex agents design, which includes pedagogical knowledge, technological knowledge, as well as models of conversation and agent behavior [16].

II. Research Question

This research is grounded in the pearls and perils of generative artificial intelligence used to design an AI agent capable of conducting the SRLSI. Building on Zimmerman's [5] foundational structured interview protocol, the study integrates AI technologies with learning sciences to evaluate the efficiency, effectiveness, and student satisfaction of this innovative approach to assessing SRL skills. Specifically, the research aims to address the following questions:

RQ1. How efficient is AI-SRLSI in conducting the Self-Regulated Learning Structures Interview (SRLSI)?

RQ2. How effective is AI-SRLSI in implementing the Self-Regulated Learning Structured Interview (SRLSI)?

RQ3. How satisfied are students with the conversational quality of AI-SRLSI?

III. Research Method

A. Participants and context of the study

The system was tested with a total of 13 participants who did interviews conducted via an integration with the Discord platform. Participants were between 19 and 34 years old. Out of the 13 interviews, 11 were considered successful, representing 84.62% of all interviews. These were used in following data analysis.

The research procedure begins by inviting students to join a Discord (https://discord.com) channel and initiating the AI-SRLSI (Fig. 1). This would not differ from other chat-like interfaces; except one characteristic: instead of the learners leading the interaction, the AI-SRLSI agent would guide the conversation. The AI-SRLSI agent starts the conversation with a greeting and guides students through six learning contexts relevant to the cohort: in-class situations, at home (completing assignments or preparing for class), when working on writing assignments outside of class, when doing assignments, preparing for and taking tests, and when poorly motivated (in line with SRLSI). For each context, the agent describes a scenario and prompts students to reflect on and write down the learning strategies they typically use. For each strategy mentioned, the Agent asks the student to rate its frequency of use. If no strategies are provided (or if strategy is unclear, incomplete, or ambiguous), the agent prompts the student again before moving to the next context. After completing all six contexts, the Agent summarizes the session, highlighting the most frequently used strategies [19] and any overlooked ones, encouraging personal reflection. This is in line with Zimmerman and Martinez-Pons's Self-Regulated Learning Structured Interview [5]. Following the interview finish, students were asked to complete a post-questionnaire about s*atisfaction* as explained in sections on measurement.

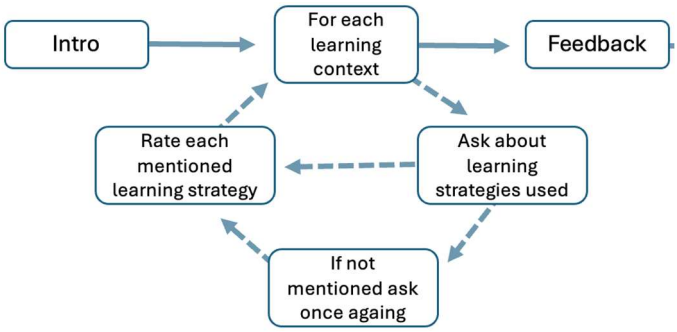

Fig 1. The research procedure

B. Implementation Summary

System Architecture. The system follows a three-tier architecture with a frontend (Discord client) and a backend API to access the database and LLM (Fig. 2, [19]). The app leverages the Discord platform, widely used by FernUniversität Hagen students for learning and discussions. Through an API the backend supports various frontend clients. The database stores essential information, including supported languages, learning contexts, strategies, and user data (e.g., conversation state, identified strategies). User and LLM messages are tracked with timestamps for full conversation reconstruction. Ollama is used for provision of open-source LLMs.

Dialogue Loop. The SRLSI is a step-by-step process guiding the user through an interview to gather information about their learning strategies (Fig. 1). Each step must be completed before moving on to the next, and there are defined next steps at the end of each. This structure aligns with Task-Oriented Dialogue (TOD) systems, which aim to help users achieve specific goals [15]. The system in this development uses a simple agent to reason whether a dialogue step is complete and generates structured JSON output for processing.

Prompt Development. The LLM used by the system is Llama 3 70b, which outperforms other open-source models on multiple benchmarks. Each dialogue step was validated to determine if the user's response answered the original question or deviated, such as with a clarifying question or unrelated comment. To assess if the step was completed, a chain-of-thought prompt was used, following [15] by asking the LLM to reason whether the step had been completed.

Categorization of user answers. The AI-SRLSI relies on recognizing learning strategies in user descriptions of study habits. To replicate this in the language agent, the system was provided with a list of strategies for analysis [5, 10]. Retrieval Augmented Generation (RAG) was implemented in the AI-SRLSI to enhance task-specific fine-tuning and minimize hallucinations. This was achieved by creating a vector database to store strategy descriptions and processing user statements through the same system. Development showed that providing the full list of strategies yielded better results than using a filtered top 5 list, especially with limited data.

Structured Output Generation. Storing user responses required a structured output, which LLMs often struggle to generate. To address this, regular expressions were used to extract and format the data as valid JSON. If no valid JSON was generated, the process was retried up to five times.

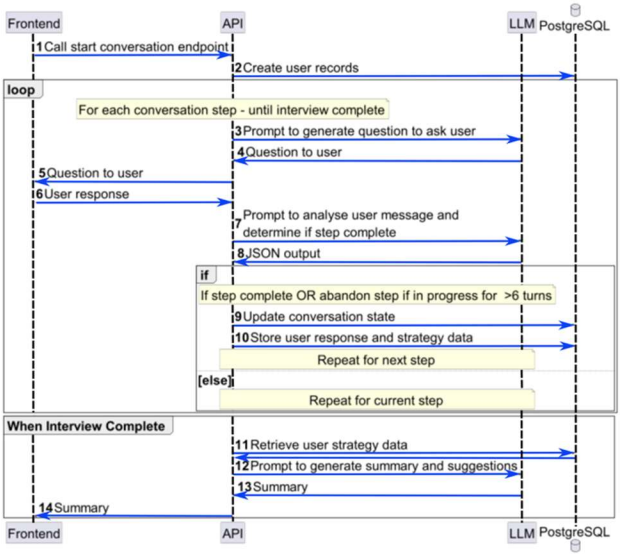

Fig 2. The design of the Agent's three-tier architecture [as introduced in 19]

Designing prompts for specific conversation steps. To achieve a natural conversation flow, the conversation history was stored in the database and used when asking users to rate strategy frequency, or when moving between learning contexts in interview stages [16].

C. Measurement

Based on the posed research questions the conversations between users and AI agent were evaluated using categories derived from the quality assessment methods described by Boukes et al. [17] and Borsci et al. [18]. The conversations were assessed in several categories of efficiency and effectiveness.

Efficiency was measured by the number of completed conversations, time taken to complete conversations (i.e. time between first and last message from the agent), number of interactions that resulted in an error (successfully continue to conduct interview even if components fail, and to handle interview comment that that do not fall within the planned interview path), and ability of the system when faced with unexpected input. Responses were considered as unexpected when participants' comment that did not contain an answer to the question asked (for example a request for clarification or a comment on the conversation itself), or when answer that did not contain the expected data for this conversation step)

Effectiveness was measured by success rate for using appropriate degrees of formality, success rate in executing steps correctly, and success rate in classifying any learning strategies mentioned by the user correctly.

Satisfaction was measured by success rate for responding to the user mood and sentiment, keeping the participant engaged and enabling them to enjoy the interaction, as measured by the tone of user responses ranked on a 5 point scale from very positive to very negative. We have also measured the user perception of the agent quality and accuracy using a survey with questions taken from the Chatbot Usability Scale [18].

IV. Results

The evaluation was conducted by retrieving the conversation data from the database, with metadata for each message containing the conversation turn, learning context being discussed, strategy recognized by the agent/being discussed in the message (see section 2.1 for an explanation of the steps), and message time. The full conversation context was then considered for the analysis. Agents' conversations, ratings and classifications were evaluated by researchers using theoretical categories described in [19] and classifications process established by Zimmerman and Martinez-Pons [5].

To assess reliability, the researcher examined all 510 messages (sent by participants and agent) in 11 successfully completed interviews. In two interviews more than 60% of the steps were executed incorrectly due to hallucinations and disregard of the prompts, leading to a deviation from the intended interview flow. This was the reason for excluding these particular interviews from the analysis below

Reliability was examined by exploring the agent coding and the researchers coding, evaluated using the interrater reliability procedure of the kappa statistic test. The results showed almost perfect agreement (of 0.97) and high percentages of accurate categorization, as reported is subsection B Effectiveness.

A. Efficiency

All conversations were completed, with 84.62% deemed successful. The time to completion, measured from the first to the last agent message had a median completion time of 41 minutes. The agent was able to pick the conversation where left or interrupted and to successfully continue dialog, with no errors in this respect. When excluding inactive periods (gaps of 15 minutes or more), the longest active completion time was 53 minutes, with an average of 29 minutes and a median of 27 minutes. When analyzing the response times, based on API log timestamps (with an average discrepancy of less than 3 ms compared to Discord front-end timestamps), most responses (72.5%) were delivered within 10–30 seconds of receiving the user's message, with average of 23 seconds.

The system successfully handled unexpected user input—responses not directly answering the question or completing the current step—in 68.69% of cases, advancing the dialogue appropriately. Unexpected inputs that deviated from the process outlined in Figure 1 and Zimmerman and Martinez-Pons's Self-Regulated Learning Structured Interview (SRLSI) included

instances where students did not answer the question directly. Instead, they sought clarification, asked for confirmation, or requested additional information. The agent managed these conversational requests, guiding the interaction back to the planned interview process.

B. *Effectiveness*

The system used an appropriate tone and formality in 74.09% of cases. Most deviations involved overly positive responses to neutral user messages, contrary to instructions to avoid evaluative commentary (e.g., "That's a great approach").

The system correctly executed 73.6% of steps. A common issue occurred in the final step, where the summary of user responses often omitted the requested suggestions. However, these were typically provided if the user requested them afterward.

The system achieved 80.37% success rate in classifying learning strategies to match human coding. Challenges arose due to ambiguities in strategy definitions, overlaps in categories, and brief user statements requiring interpretation. Improved definitions and clearer guidelines are recommended for future iterations. The agent's classifications were reviewed by the second author of this article, following the SRLSI guidelines, achieving near-perfect agreement of raters of 0.97 calculated with Cohen's Kappa. This indicates high reliability in the reviewed classifications.

C. *Satisfaction*

The system demonstrated a high success rate of 93.65% in appropriately responding to user mood and sentiment, with most unsuccessful cases involving overly positive replies to neutral messages. Regarding user engagement, 95.34% of user messages were neutral in tone, with only 2.54% rated as positive and 2.12% as negative. No messages were classified as very positive or very negative.

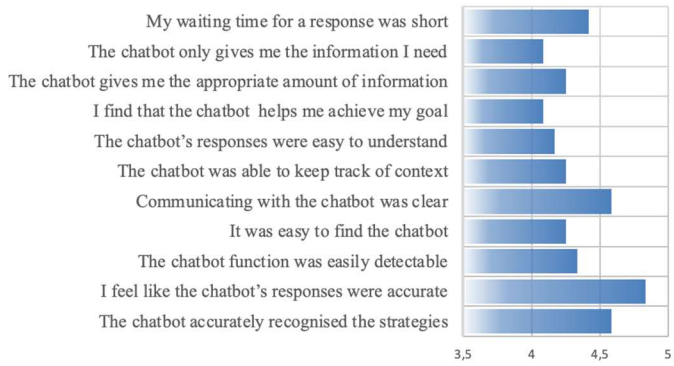

Fig 3. Participant responses to Chatbot usability questionnaire (rated on a scale from 1 to 5)

User perceptions, gathered via a survey based on the Chatbot Usability Scale [18], were highly positive (Fig. 3). Out of 11 participants, the average ratings of all items was high than 4, on a scale from 1 to 5, indicating students high agreement with statement. Participants expressed positive feedback about the chatbot's usability, with most agreeing that it was easy to communicate with, tracked context well, and provided understandable and accurate responses. Sixty percent of users found the response times quick, indicating good overall usability. See Figure 3 for additional analysis in respect to all items.

User perceptions regarding the final feedback were evaluated based on the responses. While many users (50%) reported that the chatbot prompted reflection on their learning strategies and provided useful suggestions (e.g., "It made me recognize the strategies I use"), others (25%) felt the recommendations were too general or lacked personal relevance (e.g., "The suggestions were so general and not really related to my needs"). The summary step's failure to provide suggestions unless explicitly requested contributed to some (25%) feeling that no additional strategies were offered.

V. Discussion

Building on our prior research [19, 20] on supporting students' SRL, this research provides empirical evidence on the effectiveness and efficiency of a novel approach to measuring SRL skills. Additionally, previous studies utilizing the SRLSI demonstrated that a structured interview procedure is a valid tool for assessing students' use of SRL strategies [5], but its time-intensive and laborious nature has limited its application in both research and practice. Therefore, this empirical study has examined the potential of generative artificial intelligence to support learning and self-regulation specifically by conducting an interview based on the SRLSI by Zimmerman and Martinez-Pons [5]. We designed and evaluated AI-SRLSI, a tool that was intended to establish a baseline estimate of the learner's level of self-regulation based on their reported use of learning strategies in different contexts. Several key points emerge for discussion.

Regarding the first research question, The system successfully enabled all participants to complete the interview, with some finishing in a single session and others returning over multiple days. The app has been able to work with very low percentage of errors and high percentages of handling unexpected messages, suggests that the app's design, which allowed for flexibility in communication, helped participants finish the interview successfully. The agent successfully resumed the conversation after students' breaks and interruptions. Despite varying session lengths, the active time for agent to respond for all participants remained within a acceptable range and were perceived as not too long.

Regarding the second research question, the system's overall success rate for executing conversation steps correctly was 73.6%, with significant variation across participants, ranging from over 80% to below 50%. Failures in steps typically were caused students did not answer the question directly (but sought clarification, asked for confirmation, or requested additional information). The conversation deviated from its intended flow until a correct user response allowed progression or until turn limit was reached (up to 5 retries were attempted, see Structural output generation subsection). Despite these failures, the Llama3 model, optimized for dialogue, enabled continued conversation, even if slightly off-track. Some failures were due to "hallucinations," where the agent generated information not part of the dataset, such as suggesting the "5-second rule" for procrastination. Techniques like embedding factual knowledge in prompts or using an additional "reviewer agent" may mitigate these issues but they remain persistent. The agent generally succeeded in responding to user tone and sentiment, despite frequently using a positive

tone for neutral user messages. The positive ratings on the user questionnaire indicated that the agent effectively engaged participants, helping most reflect on their learning strategies.

Regarding the third research question, participants expressed positive feedback about the chatbot's usability, with most agreeing that it was easy to communicate with, tracked context well, and provided understandable and accurate responses. Sixty percent of users found the response times quick, indicating good overall usability. However, some users desired more detailed feedback beyond familiar strategies. Feedback on the usefulness of suggestions was mixed; some found them helpful, others perceived them as generic and sought more creative or in-depth advice. Additionally, some users felt they received no further suggestions, as the summary and suggestion step failed to provide them.

Limitation. Several limitations were noted. First, this application was tested with a small sample of participants, offering valuable insights into the feasibility of the approach and identifying potential improvements. Second, the participants came from different courses rather than a single cohort, making it challenging to formulate meaningful and cohesive final feedback (cannot be related to a certain context, e.g. a course or subject.). Third, the findings were not triangulated with other measurement tools, such as self-report questionnaires or digital trace data, which could provide a more comprehensive understanding of students' SRL. Lastly, the study highlighted challenges in implementing a tool based on Task-Oriented Dialogue principles.

Further work. In future versions, deliberately controlling response times could enhance performance. Additional issues included the agent's failure to adhere to tone guidelines, often providing unwarranted positive feedback. Enhancing the system's ability to contextualize responses by extracting user behaviors could reduce misunderstandings and improve the conversational flow. Moreover, requesting additional information from learners could help build a clearer picture of their challenges and support needs, enabling the development of a more effective feedback system. Another avenue of research could focus on triangulating students' SRL skills using trace data and questionnaires, as commonly practiced. With such an agent, future studies could also explore implementing interview protocols at multiple points during the learning process to measure how students' strategy use evolves over time. Additionally, with this implementation technical foundations are placed to implement other interview guidelines using the AI tool, which is available as open-source: https://discord.gg/waRfuS7y9f.

ACKNOWLEDGMENT

This work was funded by the Center of Advanced Technology for Assisted Learning and Predictive Analytics (CATALPA) of the FernUniversität in Hagen.

REFERENCES

[1] N. Edisherashvili, K. Saks, M. Pedaste, and Ä. Leijen, "Supporting Self-Regulated Learning in Distance Learning Contexts at Higher Education Level: Systematic Literature Review," *Frontiers in Psychology*, vol. 12, Jan. 2022.

[2] R. Guan, M. Raković, G. Chen, and D. Gašević, "How educational chatbots support self-regulated learning? A systematic review of the literature," *Education and Information Technologies*, Aug. 2024.

[3] E. Panadero, "A Review of Self-regulated Learning: Six Models and Four Directions for Research," *Frontiers in Psychology*, vol. 8, Apr. 2017.

[4] B. J. Zimmerman, "Attaining Self-Regulation," in *Elsevier eBooks*, 2000, pp. 13–39.

[5] B. J. Zimmerman and M. M. Pons, "Development of a Structured Interview for Assessing Student Use of Self-Regulated Learning Strategies," *American Educational Research Journal*, vol. 23, no. 4, pp. 614–628, Jan. 1986, doi: 10.3102/00028312023004614.

[6] A. J. Sebesta and E. B. Speth, "How Should I Study for the Exam? Self-Regulated Learning Strategies and Achievement in Introductory Biology," *CBE—Life Sciences Education*, vol. 16, no. 2, p. ar30, May 2017, doi: 10.1187/cbe.16-09-0269.

[7] S. Radovic and N. Seidel, "Self-regulated learning support in technology enhanced learning environments: A reliability analysis of the SRL-S rubric," *International Journal of Assessment Tools in Education*, vol. 11, no. 4, pp. 675–698, Sep. 2024, doi: 10.21449/ijate.1502786.

[8] S. Radović, N. Seidel, D. Menze, and R. Kasakowskij, "Investigating the effects of different levels of students' regulation support on learning process and outcome: In search of the optimal level of support for self-regulated learning," *Computers & Education*, vol. 215, p. 105041, Mar. 2024, doi: 10.1016/j.compedu.2024.105041.

[9] J. Van Der Graaf *et al.*, "Do Instrumentation Tools Capture Self-Regulated Learning?," *Conference: LAK21: 11th International Learning Analytics and Knowledge Conference*, pp. 438–448, Apr. 2021, doi: 10.1145/3448139.3448181.

[10] S. F. E. Rovers, G. Clarebout, H. H. C. M. Savelberg, A. B. H. De Bruin, and J. J. G. Van Merriënboer, "Granularity matters: comparing different ways of measuring self-regulated learning," *Metacognition and Learning*, vol. 14, no. 1, pp. 1–19, Feb. 2019, doi: 10.1007/s11409-019-09188-6.

[11] Y. Fan *et al.*, "Towards investigating the validity of measurement of self-regulated learning based on trace data," *Metacognition and Learning*, vol. 17, no. 3, pp. 949–987, May 2022, doi: 10.1007/s11409-022-09291-1.

[12] T. J. Cleary and M. R. Russo, "A multilevel framework for assessing self-regulated learning in school contexts: Innovations, challenges, and future directions," *Psychology in the Schools*, vol. 61, no. 1, pp. 80–102, Aug. 2023, doi: 10.1002/pits.23035.

[13] W. X. Zhao *et al.*, "A Survey of Large Language Models," *arXiv*, Jan. 2023, doi: 10.48550/arxiv.2303.18223.

[14] D. Yan, "Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation," *Education and Information Technologies*, vol. 28, no. 11, pp. 13943–13967, Apr. 2023, doi: 10.1007/s10639-023-11742-4.

[15] W. Wang, Z. Zhang, J. Guo, Y. Dai, B. Chen, and W. Luo, "Task-Oriented Dialogue System as Natural Language Generation," *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2022, doi: 10.1145/3477495.3531920.

[16] T. Kwartler, M. Berman, and A. Aqrawi, "Good Parenting is all you need - Multi-agentic LLM Hallucination Mitigation," *arXiv*, Oct. 2024, doi: 10.48550/arxiv.2410.14262.

[17] M. Boukes, B. Van De Velde, T. Araujo, and R. Vliegenthart, "What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools," *Communication Methods and Measures*, vol. 14, no. 2, pp. 83–104, Oct. 2019, doi: 10.1080/19312458.2019.1671966.

[18] S. Borsci, M. Schmettow, A. Malizia, A. Chamberlain, and F. Van Der Velde, "A confirmatory factorial analysis of the Chatbot Usability Scale: a multilanguage validation," *Personal and Ubiquitous Computing*, vol. 27, no. 2, pp. 317–330, Aug. 2022, doi: 10.1007/s00779-022-01690-0.

[19] E. Wetchy, An AI-Based Chat Agent to Support Students' Self-Regulated Learning Skills. *Master thesis* Fern Universitait in Hagen, Jan. 2025.

[20] S. Radović, N. Seidel, L.M. Haake, and R. Kasakowskij, R. "Analyzing students' self-assessment practice in a distance education environment: Student behavior, accuracy, and task characteristics" Journal of Computer Assisted Learning, 40(2), 654–666. 2024 DOI:10.1111/jcal.12907