ASCILITE 2025

Future-Focused:

Educating in an Era of Continuous Change

Enhancing self-regulated learning with large language models: A pilot study on the feasibility of local deployment

Xiaoyu Zhuang, Aneesha Bakharia
The University of Queensland

Self-regulated learning (SRL) is essential for academic success, yet many learners struggle to plan, monitor, and reflect their learning processes without support. Large Language Models (LLMs) offer opportunities for real-time, personalised learning guidance, but cloud-based deployments raise privacy and trust concerns. This pilot study investigates the feasibility of delivering SRL support through a locally deployed, privacy-preserving chatbot. Using a design-based research approach, we co-designed a chatbot platform with sixty-one university students and conducted a two-week field study with seven participants using both local (offline) and cloud-based (online) modes. Mixed-method findings indicate that the chatbot successfully prompted higher-order SRL activities such as goal setting and reflective monitoring in authentic study sessions. Participants reported greater trust when using the fully local LLMs due to data remaining on-device. However, the local LLMs demonstrated much slower response times and occasional inaccuracies, highlighting privacy-performance trade-offs. This research demonstrates the potential of locally deployed, privacy-preserving, human-centred AI to support SRL and offers empirical insights into the benefits and limitations of deploying LLMs on small-scale local devices in educational contexts.

Keywords: Self-regulated learning, generative AI, large language models, co-design, usability evaluation, pilot study

Introduction and motivation

Self-regulated learning (SRL) refers to learners' active management of their own learning process through cyclical phases of planning, monitoring, and reflecting (Zimmerman, 2002). High SRL capacity is associated with better academic outcomes and lifelong learning skills (Broadbent & Poon, 2015). Conversely, many students fail to effectively self-regulate (Broadbent et al., 2023). Researchers and educators have long sought interventions to measure and foster SRL skills. Traditional approaches include self-reports and think-aloud protocols, but these often provide only retrospective or surface-level feedback (Graaf et al., 2021). However, real-time, personalised scaffolding of SRL remains challenging with these conventional tools.

Recent advances in Generative AI (GenAI), especially Large Language Models (LLMs), open new possibilities for supporting SRL. LLM-based conversational agents can engage learners in dialogues to clarify goals, suggest study strategies, prompt self-reflection, and provide feedback on demand (Molenaar et al., 2023). Early studies conducted by Molenaar et al. (2023) indicate that GenAI can facilitate SRL processes like strategic planning and monitoring of understanding. However, most such GenAI support relies on cloud-hosted services, such as ChatGPT, Claude, and Gemini, which send students' data to external servers. This raises legitimate privacy, security, and ethical concerns for educational deployments (Das et al., 2025; Nguyen, 2025). Learners may be hesitant to share personal learning struggles or private content with a cloud AI, which can undermine trust and willingness to engage deeply. Furthermore, institutional policies like data sovereignty laws also complicate the use of cloud AI at scale in education (Polyportis & Pahos, 2025).

To address these challenges, Kumar and Ahmed (2024) suggested a privacy-preserving approach: running LLMs locally so that no personal data leaves learners' own devices. By using on-device inference, we aim to preserve data confidentiality and potentially improve user trust. The central research question is whether a locally deployed LLM-based application can effectively support SRL comparable to a cloud-based AI, while alleviating privacy concerns. We presented a prototype SRL chatbot called *LearnSphere*, which is capable of operating in dual modes: offline mode (local LLMs) and online mode (cloud-based LLMs). The system was co-designed with university students to ensure relevance and usability, then evaluated in a user test to compare SRL support, user experience, and outcomes across the two different deployment modes.

Methodology

We followed a design-based research (DBR) methodology (Brown, 1992) to iteratively design and evaluate the SRL support tool in a real-world context. Our project comprised two main phases corresponding to DBR cycles (Wang & Hannafin, 2005): (1) co-design and development: working with students to conceptualise and build the LLM-powered SRL application; and (2) testing and evaluation: deploying the application in students' authentic study routines and assessing its impact on SRL and user experience. We adopted a mixed-methods explanatory sequential design (Creswell & Clark, 2017). Quantitative data such as survey ratings were collected and analysed first via a Technology Acceptance Model (TAM) questionnaire (Davis, 1989) and a System Usability Scale (SUS) survey (Brooke, 1996) to capture usage patterns and outcomes, followed by qualitative data like conversation logs, interviews and think-aloud observations to explain the quantitative results and gain deeper insight into user experiences. Ethical approval for this study was obtained from the University's ethics committee prior to participant recruitment and data collection.

Participant recruitment

Design activities involved a broad sample of 61 university students from undergraduate to doctoral level and represented various disciplines, including IT, education, engineering, etc. They were recruited via voluntary response to gather diverse perspectives in the design phase, ensuring the tool's features would be broadly relevant and inclusive. For the field evaluation, a smaller cohort of 7 volunteer students (5 male, 2 female; 3 undergraduates, 3 postgraduates, 1 doctoral) participated in a two-week trial of the prototype. All had some prior exposure to LLMs, but none had used GenAI specifically for SRL support before.

Co-design features

The co-design phase began with an initial survey, adapted from the SRL-O questionnaire (Broadbent et al., 2023). It assessed students' self-reported SRL skills and captured their expectations for an AI support tool. Subsequently, we conducted interactive workshops with five diverse student representatives to address identified SRL challenges. Participants prioritised key chatbot features: (a) a goal-setting assistant for defining SMART goals (Doran, 1981); (b) strategic planning prompts; (c) real-time feedback and encouragement; (d) text summarisation; and (e) affective support. Notably, privacy emerged as a major topic during the workshops. Some participants had a strong preference for an offline AI to protect personal learning data, reinforcing our focus on local versus cloud-based LLM deployment. These findings informed the final design principles of the *LearnSphere* prototype.

User evaluation procedure

Seven university students participated in a two-week pilot study using the *LearnSphere* chatbot during their normal study routines. Participants experienced both offline and online modes, each for approximately one week for their daily academic requirements. The system logged human-AI interactions automatically and stored them in the database. Participants completed TAM questionnaires after trying each mode and a SUS survey at the end, alongside periodic think-aloud reflections and semi-structured post-study interviews. The LLMs used in the offline and online modes are summarised in Table 1.

Table 1
Large language models used during user test

LLM deployment mode (platform)	Model name	Model size
Offline (Ollama)	Llama 3.2	3.2B
	Gemma 3	4.3B
	Phi-4	14.7B
Online (Groq)	Llama 3	8B
	Gemma 2	9B
	Llama 3	70B

Results and analysis

Quantitative data from TAM and SUS were analysed descriptively, given the small sample. We computed mean ratings for Perceived Usefulness (PU) and Perceived Ease of Use (PEOU) for each mode, and the SUS score for each participant. Due to the limited sample size, we did not perform inferential statistical tests, but we examined the magnitude of differences between conditions. We also calculated simple usage metrics from the logs: number of chat conversations and messages in each mode, frequency of using each feature, and average session lengths. These were plotted to observe trends. Qualitative data, like interview transcripts and open-ended feedback, were analysed using empirical analysis.

SRL activation and usage patterns

The *LearnSphere* chatbot engaged students in SRL activities during their study sessions, with all seven participants using it multiple times across the two-week pilot. Average daily active use ranged from 15 to 30 minutes and resulted in 177 conversation threads. As shown in Figure 1, problem-solving and summarisation were the most frequently used features, indicating learners often sought immediate academic support. Higher-order SRL features, including goal setting and reflective monitoring, were also used, with all participants trying each module at least once. Interview feedback confirmed that using the chatbot encouraged SRL behaviours, with one participant noting it 'made me pause and plan my study better.' Both local and cloud modes supported these interactions, with participants often unaware of mode differences aside from response speed, suggesting that SRL facilitation was consistent across deployment modes.

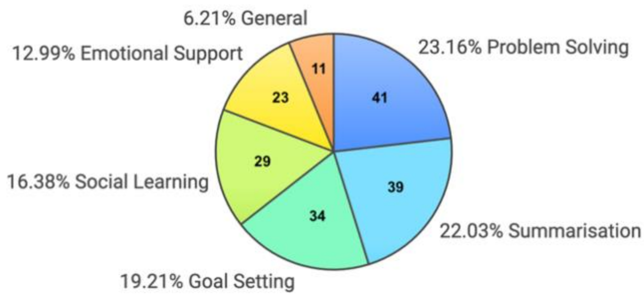

Figure 1. Distribution of feature usage

User acceptance and usability

To quantitatively assess user acceptance of the system, we examined the TAM survey results. Figure 2 shows the average ratings for Perceived Usefulness (PU) and Perceived Ease of Use (PEOU) in each deployment mode. Participants generally agreed that the chatbot was useful for their learning and fairly easy to use in both modes, but there was a consistent advantage for the cloud-based LLM on these metrics. Interviews confirmed that the slower response time and occasional hiccups with the local model made it feel 'less smooth' and 'less intelligent and helpful' at times, which likely affected these ratings. However, it's important to note that the local mode's TAM scores were still above the neutral midpoint, indicating moderate acceptance. None of the participants found the local mode completely unusable. Rather, they 'would prefer it if it could be as fast and accurate as the cloud version.' One participant summarised, 'The concept is great and I felt safer with it offline, but it was a bit slow so it tested my patience when I was in a hurry.' Another said, 'If the local AI were as powerful as the cloud AI, I'd use it exclusively.' These comments highlight that performance differences, like the speed and quality of responses, drove the slight drop in perceived usefulness of the local LLM.

In terms of system usability, the overall SUS score combining both modes averaged 68. In SUS interpretation, a score of 68 is right around the 50th percentile of usability. It is an acceptable usability level for a new system prototype.

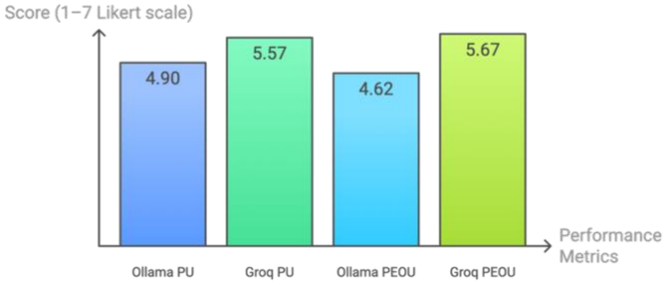

Figure 2. Perceived usefulness (PU) and perceived ease of use (PEOU) ratings for local and cloud-based modes

Trust and privacy perceptions

Beyond usability and acceptance, a critical focus was on privacy concerns. A key question was whether users indeed felt more comfortable and trusting of the local AI given that it ran offline. Interview responses were mixed but revealing. Some participants said they appreciated the privacy of the local mode. Three participants noted that knowing the data stayed on their device made them 'feel safer asking dumb questions'. This supports the idea that privacy assurances can enhance user trust and openness. Nevertheless, privacy was not the top priority for everyone. A couple of participants admitted they did not really mind using the cloud model either, since they assumed reputable services would handle data responsibly. For them, the improved performance of the cloud AI was more noticeable than the difference in privacy. In fact, a majority of participants ultimately valued accuracy and speed over absolute privacy for academic tasks. This pragmatic stance is telling that while local AI deployment can remove a barrier to trust and minimise data misuse concerns, users still need the AI to meet their performance expectations.

Output quality and cognitive load

Indeed, the quality of the AI's outputs diverged between modes due to model size differences. The cloud-based LLMs with a larger parameter size generally produced more accurate and elaborate responses, whereas the smaller local models sometimes struggled with advanced questions. We noticed that the local LLM would occasionally ask clarifying questions repetitively without resolving the issue, or give answers that were off-base, which eroded trust in its reliability. In one documented instance, the local LLM hallucinated a detailed but incorrect summary of a fictitious concept in a reading passage, whereas the cloud LLM provided a more accurate summary. Although participants often could tell when the local AI was wrong as they would double-check and catch obvious mistakes, it made them cautious. As a result, learners usually spent more time repeating or correcting their input prompts, and checking the outputs of the chatbot, thus increasing their cognitive load and affecting the effectiveness of SRL (Zhai et al., 2024).

Limitations and future directions

Overall, this pilot study makes two main contributions. First, it demonstrates the feasibility of a privacy-preserving, fully local LLM chatbot that can activate key SRL behaviours during authentic study tasks. Second, it provides initial empirical insights into the trade-offs between local and cloud-based LLMs for SRL support: local deployment can boost trust and data privacy, but currently lags in speed and output quality. This early evidence sets the stage for scaling up with larger samples and longer trials to test whether local LLMs can deliver comparable educational benefits at scale.

However, the small sample size and short two-week duration limit generalisability and preclude robust statistical analysis. Hardware variability across participants' devices introduced uncontrolled factors affecting performance perceptions. Additionally, cognitive load and learning outcomes were assessed only through self-report, and the chatbot covered a limited scope of SRL strategies with manually designed prompts.

Therefore, future work should involve larger, more diverse samples and longer-term studies to examine sustained impacts on SRL skills and academic performance. The outcome is encouraging for future innovation in decentralised learning technologies. It opens the door to follow-up research in several directions. One is exploring fine-tuned local models for educational dialogue to see if they can close the quality gap with general large models. Another is scaling up the evaluation: would these findings hold with a larger group of learners with longer learning periods? Would we see measurable improvements in SRL skills or grades for those using the AI regularly versus a control group? In addition, there are implications for future learning analytics research design: pre-post or longitudinal designs could help assess whether consistent use of such tools improves goal quality, study habits, and retention. Standardising hardware or recording device specifications would clarify performance impacts, while benchmarking evolving models will help maintain relevance. Incorporating objective measures of cognitive load and learning outcomes like NASA-Task Load Index (NASA-TLX) (Hart & Staveland, 1988) would strengthen claims of educational benefit.

References

Broadbent, J., Panadero, E., Lodge, J. M., & Fuller-Tyszkiewicz, M. (2023). The self-regulation for learning online (SRL-O) questionnaire. *Metacognition and Learning*, *18*(1), 135-163. https://doi.org/10.1007/s11409-022-09319-6

Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education*, *27*, 1-13. https://doi.org/10.1016/j.iheduc.2015.04.007

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, *189*(194), 4-7. https://doi.org/10.1201/9781498710411-35

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences*, *2*(2), 141-178. https://doi.org/10.1207/s15327809jls0202_2

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research*. SAGE Publications.

Das, B. C., Amini, M. H., & Wu, Y. (2025). Security and privacy challenges of large language models: A survey. *ACM Comput. Surv.*, *57*(6), Article 152. https://doi.org/10.1145/3712001

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319-340. https://doi.org/10.2307/249008

Doran, G. T. (1981). There's a S.M.A.R.T. way to write management's goals and objectives. *Management Review*, *70*(11), 35.

Graaf, J. v. d., Lim, L., Fan, Y., Kilgour, J., Moore, J., Bannert, M., Gasevic, D., & Molenaar, I. (2021). *Do instrumentation tools capture self-regulated learning?* LAK21: 11th International Learning Analytics and Knowledge Conference, Irvine, CA, USA. https://doi.org/10.1145/3448139.3448181

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in Psychology* (Vol. 52, pp. 139-183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Kumar, B. V. P., & Ahmed, M. D. S. (2024). Beyond clouds: Locally runnable LLMs as a secure solution for AI applications. *Digital Society*, *3*(3), 49. https://doi.org/10.1007/s44206-024-00141-y

Molenaar, I., Mooij, S. d., Azevedo, R., Bannert, M., Järvelä, S., & Gašević, D. (2023). Measuring self-regulated learning and the role of AI: Five years of research using multimodal multichannel data. *Computers in Human Behavior*, *139*, 107540. https://doi.org/10.1016/j.chb.2022.107540

Nguyen, K. V. (2025). The use of generative AI tools in higher education: Ethical and pedagogical principles. *Journal of Academic Ethics*. https://doi.org/10.1007/s10805-025-09607-1

Polyportis, A., & Pahos, N. (2025). Understanding students' adoption of the ChatGPT chatbot in higher education: The role of anthropomorphism, trust, design novelty and institutional policy. *Behaviour & Information Technology*, *44*(2), 315-336. https://doi.org/10.1080/0144929X.2024.2317364

Wang, F., & Hannafin, M. J. (2005). Design-based research and technology-enhanced learning environments. *Educational Technology Research and Development*, *53*(4), 5-23. https://doi.org/10.1007/BF02504682

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, *11*(1), 28. https://doi.org/10.1186/s40561-024-00316-7

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, *41*(2), 64-70. https://doi.org/10.1207/s15430421tip4102_2

ASCILITE 2025

Future-Focused:

Educating in an Era of Continuous Change

Zhuang, X., & Bakharia, A. (2025). Enhancing self-regulated learning with large language models: A pilot study on the feasibility of local deployment. In Barker, S., Kelly, S., McInnes, R., & Dinmore, S. (Eds.), *Future Focussed. Educating in an era of continuous change*. Proceedings ASCILITE 2025. Adelaide (pp. 562-567). https://doi.org/10.65106/apubs.2025.2654