Koh, E., et al. (Eds.) (2026). Proceedings of the 2nd International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education.

Trustworthy Secondary Use of Educational Big Data with ReLEAF

Hibiki ITO[a*], Chia-Yu HSU[b] & Hiroaki OGATA[b]
[a]*School of Informatics, Kyoto University, Japan*
[b]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*hibiki.itoo@gmail.com

Abstract: Digital infrastructures such as Learning and Evidence Analytics Framework (LEAF) involve educational big data, whose secondary use offers unprecedented opportunities for advancing research and impact on education. However, prior efforts have struggled to balance privacy protection with societal benefits. To address this issue, we present a novel data sharing system ReLEAF, developed on the LEAF infrastructure. It implements the two-stage data sharing approach: (1) differentially private synthetic data is shared and (2) real-data validation is performed on demand. Since the latter incurs additional privacy loss and operational costs associated with output checking, we conduct a formative user study to explore to what extent synthetic data alone support secondary use and when real-data validation becomes necessary. Results suggest that, while synthetic data alone could support exploratory analysis, validation is perceived as necessary to publish findings or apply to practice, unless the quality of synthetic data is guaranteed. Implications for system improvement and future directions are discussed.

Keywords: Data infrastructure, secondary use, data sharing, ReLEAF

1. Introduction

Increasing adoption of digital technologies in education has led to the development of large-scale infrastructures, such as Learning and Evidence Analytics Framework (LEAF; Ogata et al., 2022), the National Digital Education Architecture of India (DSEL, 2021) and the Taiwan Adaptive Learning Platform (MOE, 2025). These infrastructures collect and store educational big data, whose secondary use offers substantial research opportunities for understanding learning processes and identifying potential issues that need support. Despite this potential, access to such data is often limited to trusted researchers or even not provided to third parties due to privacy constraints (Fischer et al., 2020). This hinders research opportunities and open science practices in the education domain.

Prior efforts have attempted to address this problem yet have struggled to balance privacy protection with collective benefits. On one hand, the lack of proper privacy protection fails to obtain societal acceptance. The inBloom initiative aimed for exchanging educational data and best practices across schools, yet it was discontinued just one year after its launch following widespread public opposition about sharing sensitive personal information (Bulger et al., 2017). On the other hand, overly strong privacy protection loses usability and potential societal benefits. The MOOC Replication Framework (MORF) allows for remotely performing analyses on large-scale MOOC data, yet maximised privacy protection resulted in limited utility, hampering its wider adoption in the research community (Baker & Hutt, 2025).

To address this tension between privacy protection and data usability, we present a novel data sharing system ReLEAF, developed on the LEAF infrastructure. ReLEAF adopts the two-stage data sharing method proposed by Ito et al. (2026b): (1) differentially private synthetic data is shared, and (2) real-data validation—researchers submit code to a validation server—is performed on demand. In this paper, we discuss the system design and conduct a formative user study. Implications for improving the system and future directions are discussed.

2. System design

Figure 1 shows the overview of the ReLEAF system, consisting of two web UIs, Lab and Review, and *datasites* hosted at each institution (i.e. validation servers). For a datasite, we employ PySyft[1] to store both pseudonymised and synthetic data, and the web UIs function as a hub of datasites. Each datasite contains multiple datasets, each of which contains one or more assets, a pair of CSV files, each containing pseudonymous real data and corresponding synthetic data generated from that real data.

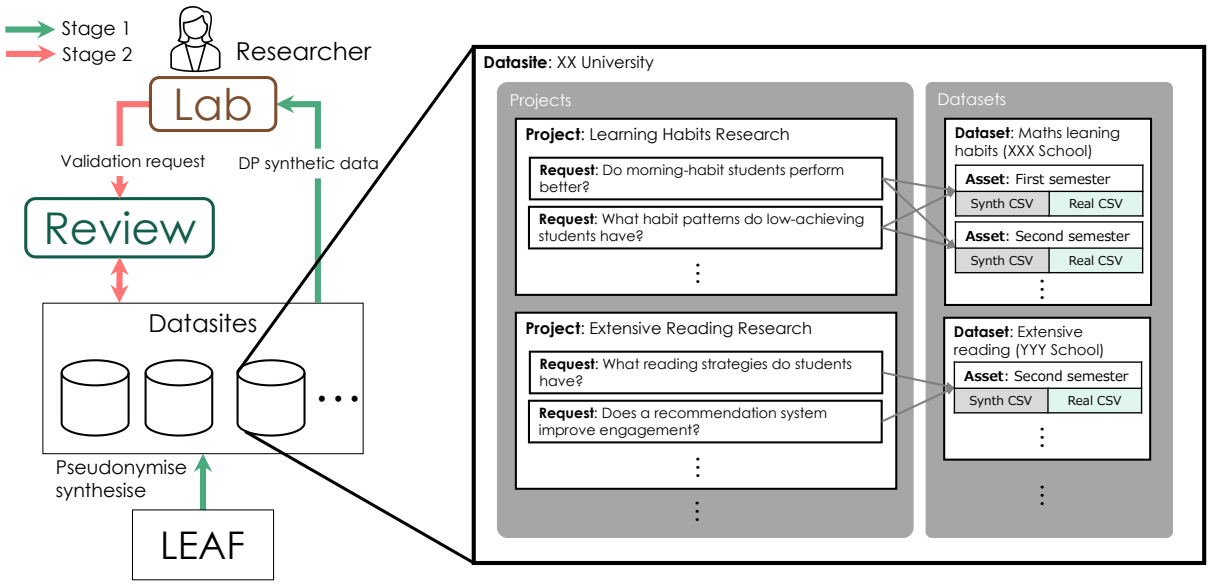

Figure 1. Overview of the ReLEAF system.

The two-stage approach combining differentially private synthetic data generation (DP-SDG) and real-data validation was recently proposed as a practical method for sharing educational data while protecting privacy (Ito et al., 2026b). DP-SDG offers robust privacy protection, unlike traditional anonymisation that is unpredictably vulnerable to privacy attacks. Ideally, if synthetic data achieve sufficient fidelity, researchers can rely exclusively on them. However, as synthetic data can easily produce misleading or false discoveries (Montoya Perez et al., 2024), real-data validation allows for non-DP output, intending to support researchers in confirming findings. While Ito et al. (2026b) conceptually proposed the two-stage method and experimentally assessed its utility and risk, our work designs a system that implements the method. Specifically, ReLEAF implements the method as follows:

Stage 1. A datasite is deployed at each institution and a data custodian uploads DP synthetic data and corresponding pseudonymous data through Review. Third-party researchers download synthetic data through Lab and analyse the data locally.

Stage 2. Researchers may perform real-data validation on demand, following the workflow shown in Figure 2. A researcher first creates a project in the datasite by describing the research objective and public interest rationale and then develops a validation request within a project with the help of LLM review. After a human reviewer approves the request, the researcher can run the code on the datasite and obtain output.

The Python code must be a self-contained function to comply with PySyft requirements. No internet access or file writes are allowed, and common data science packages are preinstalled on the datasites. Since real-data validation could unintentionally disclose personal information, output checking is needed for statistical disclosure control (SDC; Griffiths et al.,

[1] https://github.com/OpenMined/PySyft

2024). Outputs are reviewed for their necessity to answer the research question associated with a validation request and their collective benefits justified by the project description. We explicitly prohibit outputting individual-level information, only permitting aggregated output.

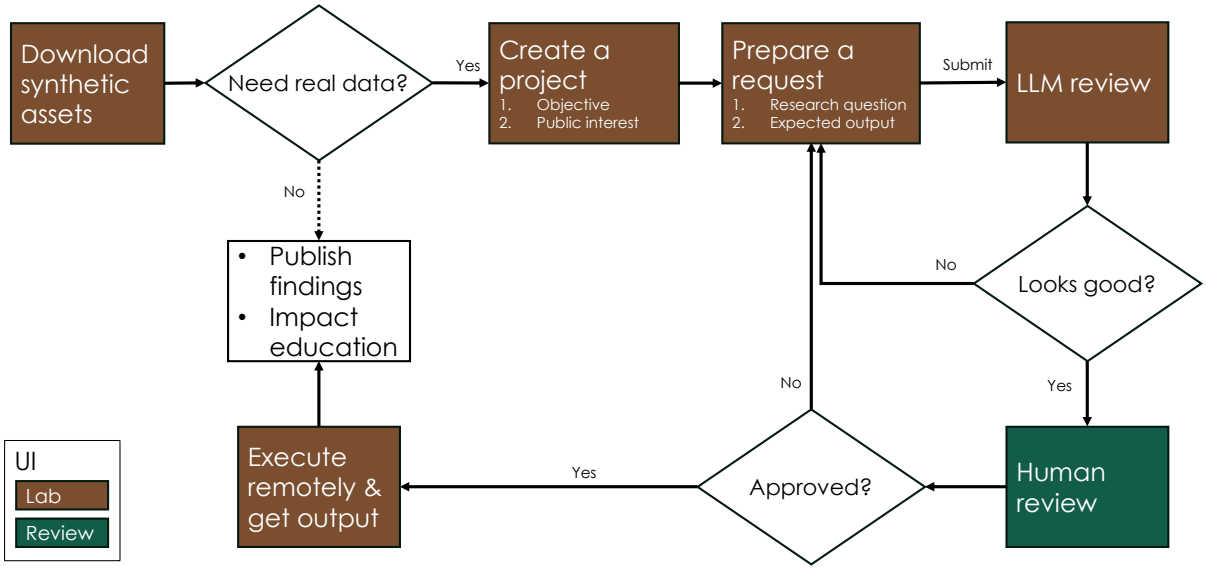

Figure 2. Workflow of real-data validation (stage 2) through ReLEAF.

3. Formative user study

While ReLEAF proposes an architecture designed to balance privacy protection with research accessibility, real-data validation introduces additional information-theoretic privacy loss and operational costs associated with output checking. Consequently, it is important to understand to what extent synthetic data alone (stage 1) can support secondary use and when real-data validation (stage 2) becomes necessary. Therefore, we conducted a formative user study exploring how researchers perform secondary analyses using synthetic data alone and with the two-stage approach implemented in ReLEAF. To this end, we set the following RQ:

- **RQ**: To what extent can researchers perform secondary use with synthetic data alone and when does real-data validation become necessary?

3.1 Experiment settings

A pilot experiment was conducted with four graduate students over five weeks. Three participants majored in educational technology and one in another field of computer science. In the first week, participants went through a tutorial of ReLEAF and practiced output checking. Then, we created the common project "Learning Habits and Outcomes" on which everyone worked independently. The project objective was set to the following:

- **Project objective**: Understanding the relationship between learning habits and outcomes and identifying potential issues with learners that would need support.

Three LEAF datasets from years 2022 to 2024 were prepared. Each dataset contains seventh graders' ($N = 120$ per year) learning outcomes (low, average and high) and time spent on BookRoll over 17 weeks, an e-book system within LEAF (Ogata et al., 2017). The BookRoll session data was aggregated by week and time windows—morning, afternoon, evening and overnight—following the method of Hsu et al. (2023). We designed a multi-shot synthesis setting, where similar data of different learner cohorts arrive regularly (Ito et al., 2026a):

1. Dataset 2022 was uploaded to ReLEAF, and students analysed it during the second week
2. Dataset 2023 was uploaded at the beginning of the third week, and students analysed both datasets during the week.

3. In the fourth week, students shared findings during the class.
4. Finally, dataset 2024 was uploaded at the beginning of the fifth week, and students analysed it as well as previous datasets during the final week.

This way, we aimed to emulate a three-year secondary-use project. We generated DP synthetic data by following the LLM-based multi-synthesis method of Ito et al. (2026b) with privacy parameters $\varepsilon = 1$ and $\delta = 10^{-3}$. Specifically, summary statistics were calculated from real data with DP, and then LLMs generated Python code to produce synthetic data that aligns the given context, schema and the DP statistics. Generation of the second and third datasets were informed of real-data validation results of previous years so that synthetic data quality is expected to cyclically improve (Ito et al., 2026b, 2026a). Informed consents were obtained from both LEAF users and the pilot participants for using their data for research.

3.2 Focus group

A semi-structured focus group was conducted at the end of the pilot because it generates rich *language in vivo* which helps designing subsequent studies, thereby suitable for a formative study (Tracy, 2012). A moderator prompted the below questions, followed by free discussions:

- To what extent did you achieve the project objective with synthetic data?
- To what extent did you achieve the project objective with the addition of real-data validation?

For analysing the focus group data, we used the Steps for Coding And Theorising (SCAT), as it is suitable for one-time, small-scale data (Otani, 2008). SCAT codes utterances into themes and constructs and then contextualises them back into a single storyline. Given evolving evidence that LLMs can effectively support qualitative analysis (Hayes, 2025), an LLM-based SCAT pipeline was developed.[2] Specifically, a human researcher, GPT-5.4, Gemini 3.1 Pro and Claude Opus 4.6 generated themes and constructs independently. The researcher filters LLMs' outputs that are important but missing in the human output, and a storyline is then manually created from the updated human output.

3.3 Results

For descriptive purposes, we report the number of requests over the pilot period in Figure 3. Period 1 started when dataset 2022 was uploaded (week 2), period 2 started when dataset 2023 was uploaded (week 3 and 4), and period 3 started when dataset 2024 was uploaded (week 5). 38 requests were created in total, 4 requests were rejected, and 8 requests were withdrawn after LLM reviews. The number of requests increased over time, suggesting validation requests typically occurred after an initial period of exploratory synthetic-data analysis.

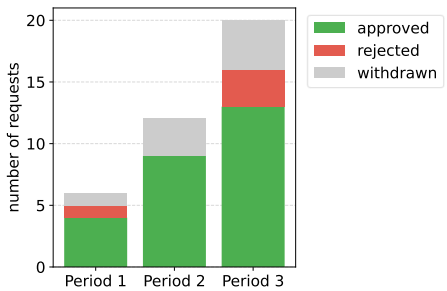

Figure 3. Number of requests over time during the pilot.

Figure 4 shows the storyline derived by SCAT on the focus group data. The result suggests that synthetic findings are perceived as infeasible to publish or apply to educational practice. The participants reported that real-data validation becomes necessary for such

[2] https://github.com/hibiki-i/scat-llm-pipeline-v2

downstream integration, unless the fidelity of synthetic data is guaranteed. Nonetheless, we also find challenges in the two-stage workflow. Discrepancies between synthetic and real data could damage credibility in synthetic data, and transparency in the SDG process shapes the analysis plan. As a trade-off with privacy protection, the review process can constrain the speed of research processes.

The participants observed that LLM-generated synthetic data tended to produce hypothesis-confirming results that are not validated on real data, pointing out the infeasibility to achieve synthetic-data-alone scientific discoveries and practice applications. This included particular concerns about longitudinal risk of false-discovery accumulation. To mitigate false discoveries, participants demanded SDG transparency, which **shapes the analysis plan**. For conditions when real-data validation becomes necessary, a participant optimistically viewed future synthetic-data-alone analysis given sufficient fidelity, while still showing **apprehension towards practical and academic applications without real-data validation**. On the other hand, another participant posited **delegation of validity assurance to data providers**, demanding either validity guarantee or real-data validation. Although real-data validation enables scientific publications and practice applications, a participant experienced damaged credibility of synthetic data with real-data invalidation and the constraints on the speed of the research process due to manual reviews.

Figure 4. Storyline of the focus group. Underlines indicate themes and constructs derived from participants' utterances through SCAT, where bolded ones are extracted by LLMs.

4. Discussion and conclusion

We presented a novel data sharing system ReLEAF. By implementing the two-stage data sharing method, it aims to offer strong privacy protection while maintaining data accessibility for researchers. Our formative user study suggests that, while synthetic data can support exploratory analyses, findings would be difficult for researchers to publish or apply to practice, requiring validation. This indicates ReLEAF's key trade-off between privacy/operational costs and societal benefits. Further work would be needed on what concrete benefits ReLEAF may deliver by offering research opportunities and promoting open science practices and what risks are accepted by stakeholders as a trade-off. This would inform the design of ReLEAF, especially output checking of real-data validation.

Additionally, challenges are also identified. As the review process may block speedy research processes, it would be beneficial to automate part of the process and make the results and process time more predictable. Improving the quality of synthetic data is also prioritised, since the lack of fidelity can lead to unnecessary validation requests and damage credibility of synthetic data. Moreover, future work should also consider longitudinal sustainability of ReLEAF such as reviewer training.

As limitations of the formative user study, the pilot employed a small sample size, and participants may not represent the researcher population who would conduct secondary use. Thus, the results lack generalisability, and more comprehensive, larger-scale evaluation is needed. Nonetheless, the goal of the study is to generate design insights to inform future system development rather than to provide reproducible evidence about researcher behaviour.

In conclusion, to address data access issues with digital infrastructures in education, ReLEAF is designed for trustworthy secondary use balancing privacy protection with collective benefits. Our formative study highlights the trade-off between the benefits of reliably publishing findings and applying to practice and privacy/operational costs, while also identifying challenges that necessitate further system development. Future work needs to evaluate this trade-off and address the identified challenges.

Acknowledgements

This work was supported by CSTI SIP Grant Number JPJ012347 and JSPS KAKENHI Grant Number 23H00505, 25KJ1515.

References

Baker, R., & Hutt, S. (2025). MORF: A Post-Mortem. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 797–802.

Bulger, M., McCormick, P., & Pitcan, M. (2017). *The Legacy of inBloom*. https://datasociety.net/wp-content/uploads/2017/02/InBloom_feb_2017.pdf

DSEL. (2021). *National Digital Education Architecture*. Department of School Education and Literacy (DSEL), Government of India. https://www.ndear.gov.in/images/pdf/NDEAR%20Main%20Report_July%2026_210728_194926.pdf

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education*, *44*(1), 130–160.

Griffiths, E., Greci, C., Kotrotsios, Y., Parker, S., Scott, J., Welpton, R., Wolters, A., & Woods, C. (2024). *Handbook on Statistical Disclosure Control for Outputs*. Safe Data Access Professionals Working Group.

Hayes, A. S. (2025). "conversing" with qualitative data: Enhancing qualitative research through large Language Models (LLMs). *International Journal of Qualitative Methods*, *24*(16094069251322346). https://doi.org/10.1177/16094069251322346

Hsu, C.-Y., Otgonbaatar, M., Horikoshi, I., Li, H., Majumdar, R., & Ogata, H. (2023). Chronotypes of Learning Habits in Weekly Math Learning of Junior High School. *Proceedings of the 31st International Conference on Computers in Education*, *1*, 566–568.

Ito, H., Hsu, C.-Y., & Ogata, H. (2026a). Cyclic adaptive private synthesis for sharing real-world data in education. *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference*, 32–41.

Ito, H., Hsu, C.-Y., & Ogata, H. (2026b). Training-free private synthesis with validation: A new frontier for practical educational data sharing. In *arXiv [cs.CY]*. arXiv. https://doi.org/10.48550/arXiv.2604.01821

MOE. (2025). *Education in Taiwan (2025-2026)*. Ministry of Education, Republic of China (Taiwan). https://stats.moe.gov.tw/files/ebook/Education_in_Taiwan/2025-2026_Education_in_Taiwan.pdf

Montoya Perez, I., Movahedi, P., Nieminen, V., Airola, A., & Pahikkala, T. (2024). Does differentially private synthetic data lead to synthetic discoveries? *Methods of Information in Medicine*, *63*(1–02), 35–51.

Ogata, H., Majumdar, R., Yang, S. J. H., & Warriem, J. M. (2022). Learning and evidence analytics framework (LEAF): Research and practice in international collaboration. *Information and Technology in Education and Learning*, *2*(1), Inv-p001-Inv-p001.

Ogata, H., Oi, M., Mohri, K., Okubo, F., Shimada, A., Yamada, M., Wang, J., & Hirokawa, S. (2017). Learning analytics for E-book-based educational big data in higher education. In *Smart Sensors at the IoT Frontier* (pp. 327–350). Springer International Publishing.

Otani T. (2008). SCAT A Qualitative Data Analysis Method by Four-Step Coding: Easy Startable and Small Scale Data-Applicable Process of Theorization. *Bulletin of the Graduate School of Education and Human Development: Psychology and Human Developmental Sciences*, *52*(2), 27–44.

Tracy, S. J. (2012). *Qualitative research methods: Collecting evidence, crafting analysis, communicating impact*. Wiley-Blackwell.